9. Fluid forgalmak, forgalom prioritizálás

Kommunikációs hálózatok teljesítményének elemzése

Horváth Illés

2024/11/06

1. Bevezetés
2. Fluid modell
3. Markov-modulált fluid sor
4. Forgalmak prioritizálása
5. Statikus WFQ hálózat elemzése

Bevezetés

Tekintsünk egy M/M/1 sort, amiben a kiszolgálási ráta μ, az érkezési ráta λ, $\lambda < \mu$.

Modellezzük azt a helyzetet, hogy az egyes igények mérete pici, de sűrűn érkeznek!

Ha a szerver egy egységnyi méretű igényt μ rátával szolgál ki, akkor ha egy igény mérete $1/n$, akkor azt μn rátával szolgálja ki (n értéke nagy).

Legyen az érkezési ráta arányosan λn.

Hogyan viselkedik ekkor a sor?

Bevezetés

A sor terhelése ugyanúgy

$$\rho = \frac{\lambda}{\mu} = \frac{\lambda n}{\mu n}.$$

A sorhossz stacionárius eloszlása is PGEO$(1-\rho)$, viszont az igények mérete $1/n$, tehát ha a buffer telítettségét nem a sorhosszal, hanem az igények összméretével mérjük, akkor a bufferben lévő adatmennyiség $1/n$-nel arányos.

A rendszerben töltött idő eloszlása EXP$((1-\rho)\mu n)$, aminek a várható értéke $\frac{1}{(1-\rho)\mu n}$.

Fluid modell

Összességében, amint $n \to \infty$, a sor a következőképpen viselkedik:

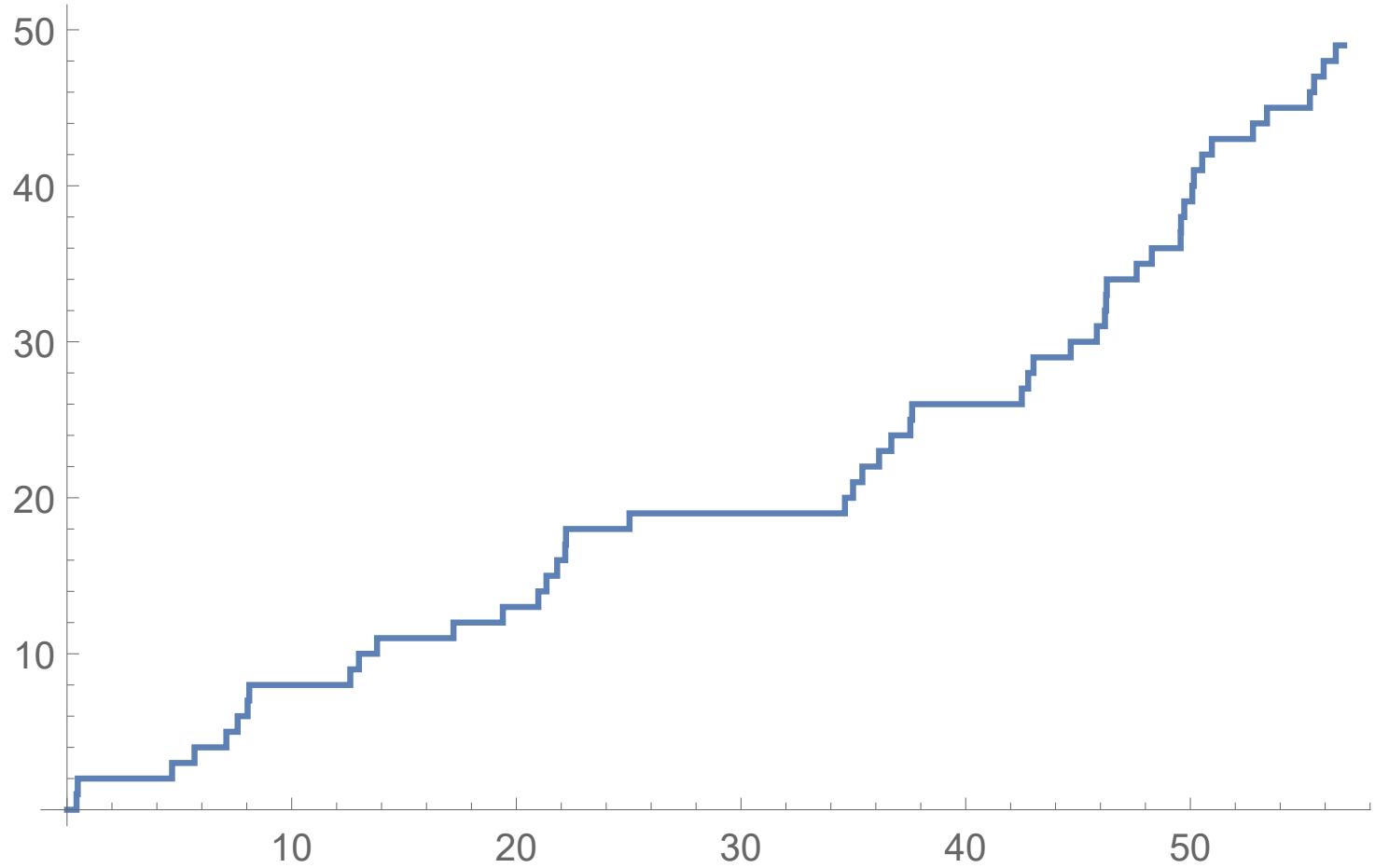
- az egyes igények rendszerben töltött ideje 0, és
- a bufferben lévő adatmennyiség is 0.

Gyakorlatilag az történik, hogy az érkezés sűrűn, apró egységekben történik és "átfolyik" a szerveren, mint egy folyadék.

Ez a fő gondolata a fluid modelleknek: a forgalmat nem csomagszintű egységekben modellezzük, hanem folytonos folyadékként.

Fluid modell

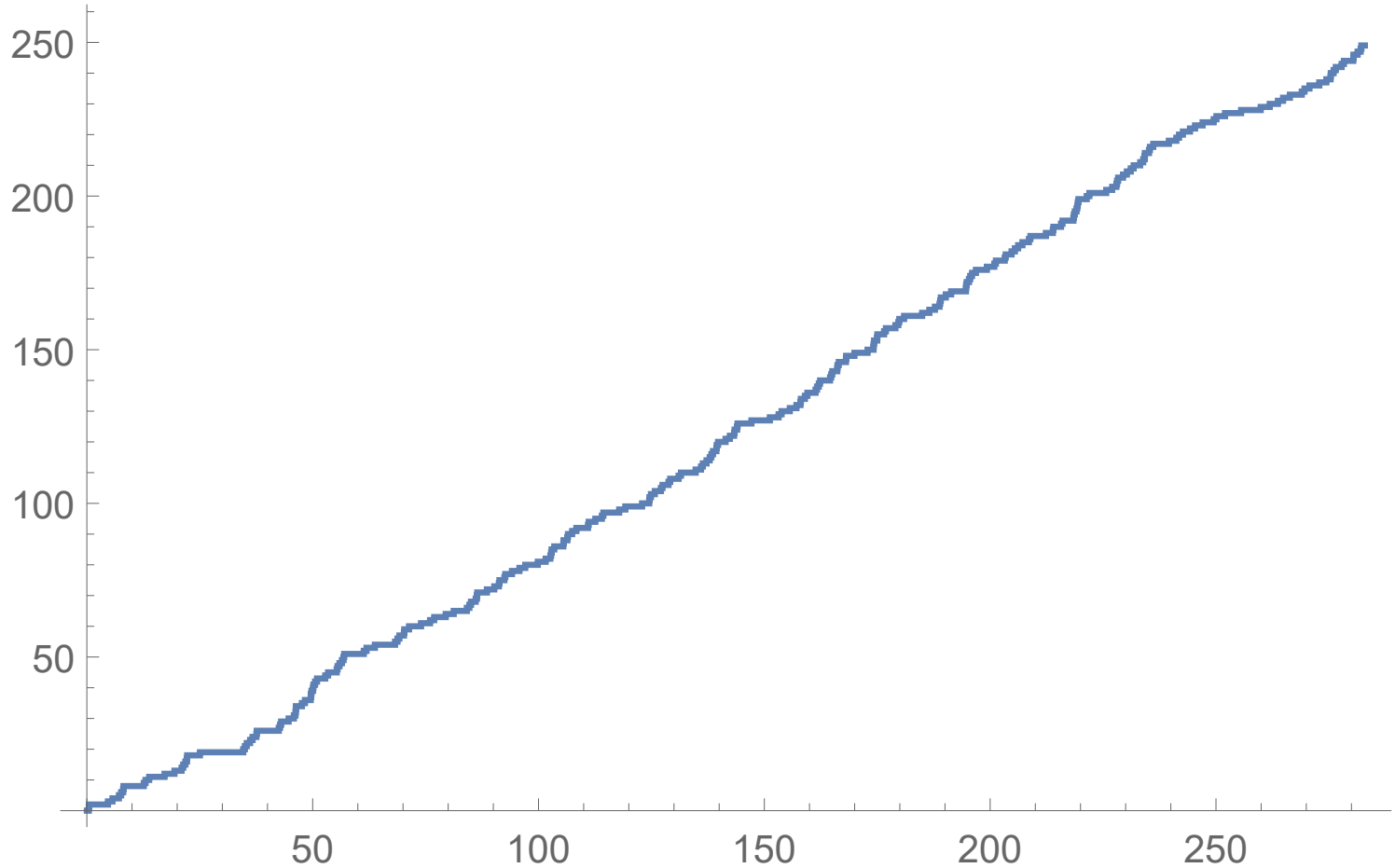
Az érkezési folyamat Poisson-pontfolyamat λn rátával; hogy néz ez ki, amint $n \to \infty$?



$\lambda = 1, n = 50$

Fluid modell

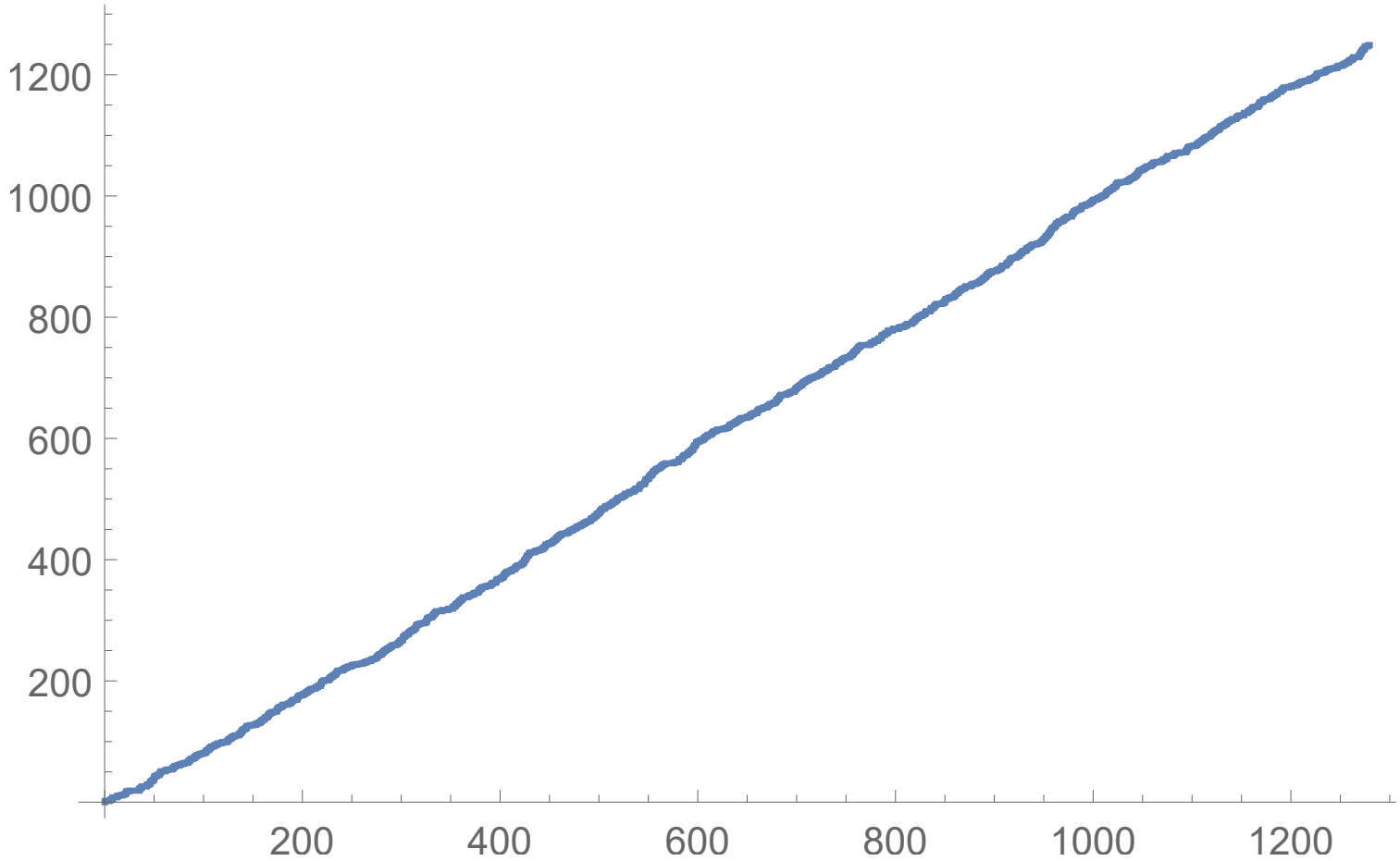
Az érkezési folyamat Poisson-pontfolyamat λn rátával; hogy néz ez ki, amint $n \to \infty$?



$\lambda = 1, n = 250$

Fluid modell

Az érkezési folyamat Poisson-pontfolyamat λn rátával; hogy néz ez ki, amint $n \to \infty$?



$\lambda = 1, n = 1250$

Fluid modell

Amint $n \to \infty$, az érkezési folyamatban lévő véletlen ingadozások egyre kisebbek; gyakorlatilag az adatcsomag szintű véletlen ingadozások olyan kis léptékűek, hogy a fluid modellben már nem látszanak, és a határértékben az érkezési folyamat determinisztikus, ahol a forgalom állandó λ sebességű folyadékként érkezik.

Fluid modellekben a szerver szerepe olyan, mint egy cső: egy μ kiszolgálási rátájú szerverből egy μ kapacitású kifolyó cső lesz.

Fluid modell

Ha $\lambda < \mu$, akkor a befolyó λ sebességű folyadék egyből ki is folyik szintén λ sebességgel.

Ha $\lambda > \mu$, akkor alapesetben adatvesztés történik. Ennek elkerülésére több lehetséges megoldás van:

- Késleltetés (bufferek alkalmazása), a bufferek méretétől függő időtávon segít.
- Visszaszabályozás (ha a szerver adatvesztést jelez, akkor a forrás lassabban ad): a forgalom típusától függően ennek egyéb következményei is lehetnek, pl. a kiszolgálás minősége lesz gyengébb (pl. zene vagy videó stream), vagy hosszabb ideig ad a forrás (pl. fix méretű le- és feltöltések). Ilyen szabályozást valósít meg pl. a TCP vagy más protokollok. Lásd még: variable bit rate (VBR) forgalom.
- Újraküldés: nem kezeli a probléma gyökerét, viszont a többivel kombinálva hasznos lehet.

Fluid modell

Mi fluid forgalom esetén feltesszük, hogy a visszaszabályozás instant és tökéletes, nincs adatvesztés.

A fő vizsgált teljesítményjellemző az átviteli sebesség hosszú távú átlaga (throughput).

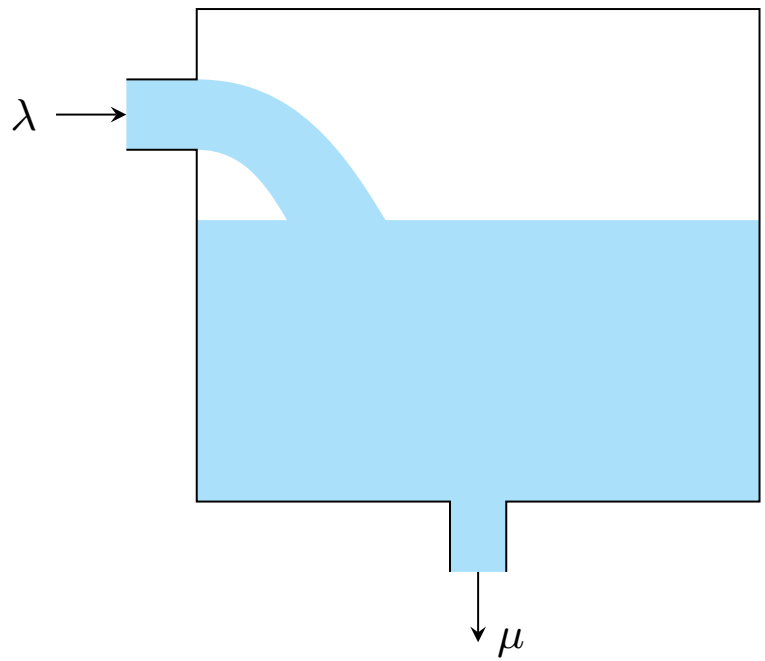
Buffereket megengedünk; ha vannak bufferek egy rendszerben, akkor a késleltetés is releváns teljesítményjellemző.

Kitekintés: attól függően, hogy egy adott forgalomra nézva a visszaszabályozás milyen következményekkel jár (pl. gyengébb minőségű szolgáltatás, vagy hosszabb ideig aktív forgalom stb.), más teljesítményjellemző is releváns lehet.

Fluid buffer

Ha van a szervernek buffere, akkor az lehet véges vagy végtelen kapacitású. $\lambda > \mu$ esetén elkezd megtelni $\lambda - \mu$ sebességgel, és a forrás akkor szabályoz vissza, amikor a buffer megtelik.

A bufferben egy „csepp” késleltetése attól függ, hogy a beérkezésekor mekkora volt a buffer szint.



Markov-modulált fluid modell

Fluid modellekben a λ érkezési ráta lehet időben változó is.

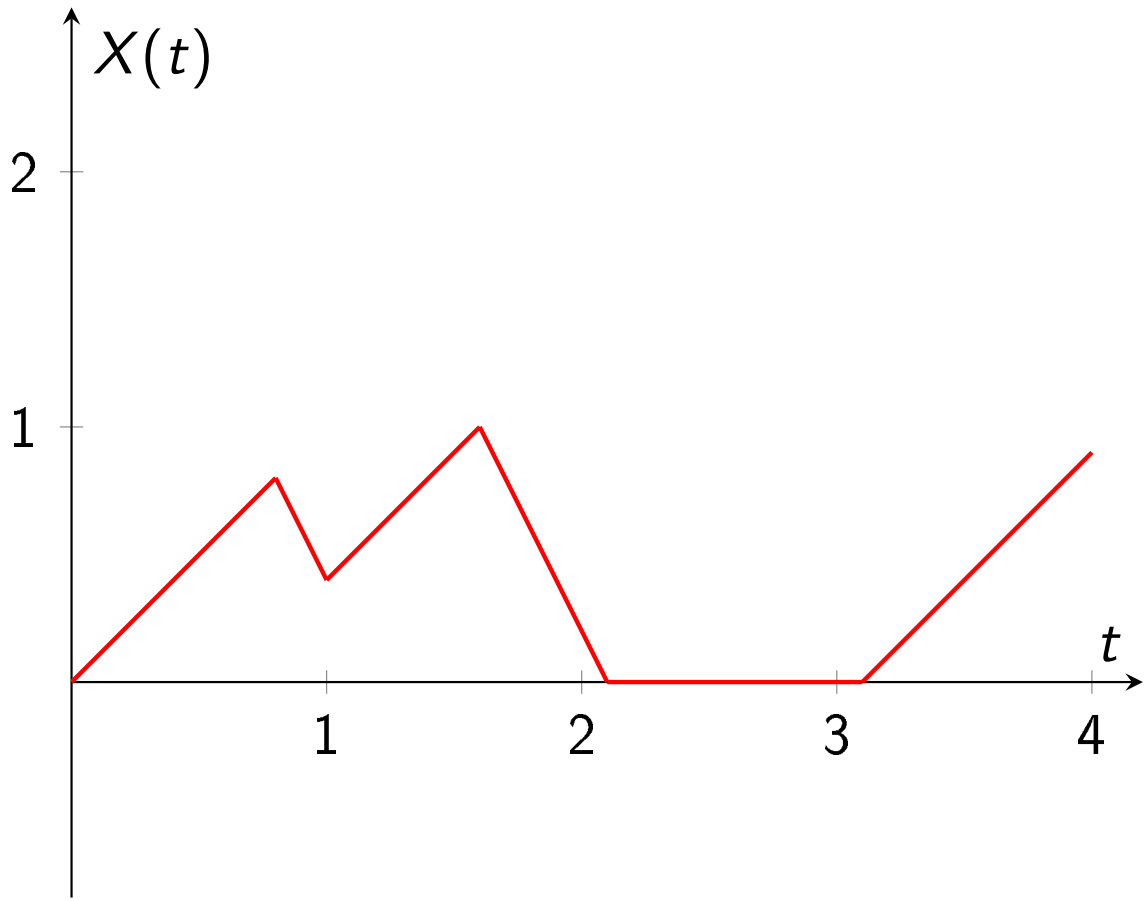
Egy gyakori feltevés az, hogy az érkezési ráta *Markov-modulált*: λ értéke egy $S(t)$ háttér folytonos idejű Markov-folyamat szerint változik ($S(t)$ véges állapotterű, irreducibilis, generátora Q). Ilyenkor a háttér Markov-lánc minden i állapotához tartozik egy λ_i érkezési sebesség.

Egy olyan fluid sort fogunk elemezni, amiben a μ kiszolgálási sebesség konstans, és a szervernek végtelen buffere van.

Jelölje a buffer szintjét a t időpontban $X(t)$. Ekkor az $(S(t), X(t))$ párra teljesül a Markov-tulajdonság.

Markov-modulált fluid sor

Példa. Egy kétállapotú háttér-Markov lánc által vezérelt fluid sor buffer-folyamata.



Markov-modulált fluid sor

Jelölje az $S(t)$ folyamat stacionárius eloszlását π, azaz

$$\pi Q = 0, \quad \pi_1 + \cdots + \pi_k = 1$$

és

$$\lim_{t\to\infty} P(S(t) = i) = \pi_i.$$

Lemma

Egy Markov-modulált fluid sor pontosan akkor stabil, ha

$$\sum_{i=1}^{k} \pi_i \lambda_i < \mu.$$

Biz. Ergodtétel alapján a hosszú távú átlagos érkezési ráta $\sum_{i=1}^{k} \pi_i \lambda_i$, μ-nek ennél nagyobbnak kell lennie.

Markov-modulált fluid sor

$(S(t), X(t))$ stacionárius eloszlását szeretnénk kiszámítani.

Ehhez előkészületnek a következő mennyiségekre lesz szükségünk:

$$F_i(t,x) = P(X(t) < x, S(t) = i) \qquad (i = 1, \ldots, k).$$

F_i tulajdonságai: $\forall t > 0$ rögzített értékre

- $\lim_{x\to\infty} F_i(t,x) = P(S(t) = i)$;
- $\lim_{x\to 0+} F_i(t,x) = P(S(t) = i, X(t) = 0)$.

Figyelem: $F_i(t,x)$-nek azon i-kre, melyekre $\lambda_i < \mu$, ugrása van $x = 0$-ban; az ugrás nagysága annak a valószínűsége, hogy a buffer t-kor üres.Jelölje ezt

$$\ell_i(t) = P(X(t) = 0, S(t) = i),$$

és legyen a buffer sűrűség t-kor

$$f_i(t,x) = \frac{\partial}{\partial x} F_i(t,x) \qquad (i = 1, \ldots, k).$$

Kolmogorov-egyenletek

Tétel (Kolmogorov-egyenletek)

$$\frac{\partial}{\partial t} f_i(t,x) + (\lambda_i - \mu)\frac{\partial}{\partial x} f_i(t,x) = \sum_{j=1}^{k} q_{ji} f_j(t,x) \qquad (i = 1, \ldots, k),$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \ell_i(t) = -(\lambda_i - \mu) f_i(t,0) + \sum_{j=1}^{k} q_{ji} \ell_j(t) \qquad (i = 1, \ldots, k).$$

Nem biz., de azt kell végiggondolni, hogy kis idő alatt hogyan változhat $S(t)$ és $X(t)$, és ez mit jelent $f_i(t,x)$-re és $\ell_i(t)$-re.

Stacionárius limesz

Az eddigi leírás tetszőleges t értékre érvényes; a stacionárius esetben $t \to \infty$, és a t-től való függés megszűnik:

$$F_i(x) = \lim_{t\to\infty} F_i(t,x) \qquad (i = 1,\dots,k),$$
$$f_i(x) = \lim_{t\to\infty} f_i(t,x) \qquad (i = 1,\dots,k),$$
$$\ell_i = \lim_{t\to\infty} \ell_i(t) \qquad (i = 1,\dots,k).$$

A Kolmogorov-egyenletek ilyenkor a következőre egyszerűsödnek:

$$(\lambda_i - \mu)\frac{\mathrm{d}}{\mathrm{d}x} f_i(x) = \sum_{j=1}^{k} q_{ji} f_j(x) \qquad (i = 1,\dots,k),$$

$$0 = -(\lambda_i - \mu) f_i(0) + \sum_{j=1}^{k} q_{ji}\ell_j \qquad \text{ha } \lambda_i > \mu,$$

$$\ell_i = 0 \qquad \text{ha } \lambda_i < \mu.$$

Stacionárius eloszlás

Az előbbi egyenletek megoldhatóak expliciten, és a megoldás PH-eloszlás lesz (Asmussen, 1995):

$$f_i(x) = \sum_{j=1}^{k} a_j e^{\lambda_j x}(\phi_j)_i \qquad (i = 1, \ldots, k),$$

$$\ell_i = \pi_i - \int_0^{\infty} f_i(x)\mathrm{d}x \qquad (i = 1, \ldots, k)$$

alakú, ahol a (λ_j, ϕ_j)-k az $R^{-1}Q$ mátrix sajátérték-sajátvektor párjai, ahol

$$R = \begin{bmatrix} \lambda_1 - \mu & 0 & \ldots & 0 \\ 0 & \lambda_2 - \mu & \ldots & 0 \\ & & \ddots & \\ 0 & \ldots & 0 & \lambda_k - \mu \end{bmatrix},$$

az a_{ij} konstansok pedig a megfelelő lineáris egyenletrendszerből számíthatóak ki.

ON/OFF fluid sor

Egy egyszerű, de gyakran releváns eset az ON/OFF fluid sor, ahol csak két állapot van, az 1-es állapotban az érkezési sebesség λ (ON), a 2-es állapotban 0 (OFF). Tegyük fel, hogy $\lambda > \mu$.

Erre a stacionárius eloszlás a következő:

$$f_1(x) = \frac{q_{21}}{q_{12}+q_{21}}\left(\frac{q_{12}}{\lambda-\mu}-\frac{q_{21}}{\mu}\right)e^{-\left(\frac{q_{12}}{\lambda-\mu}-\frac{q_{21}}{\mu}\right)x},$$

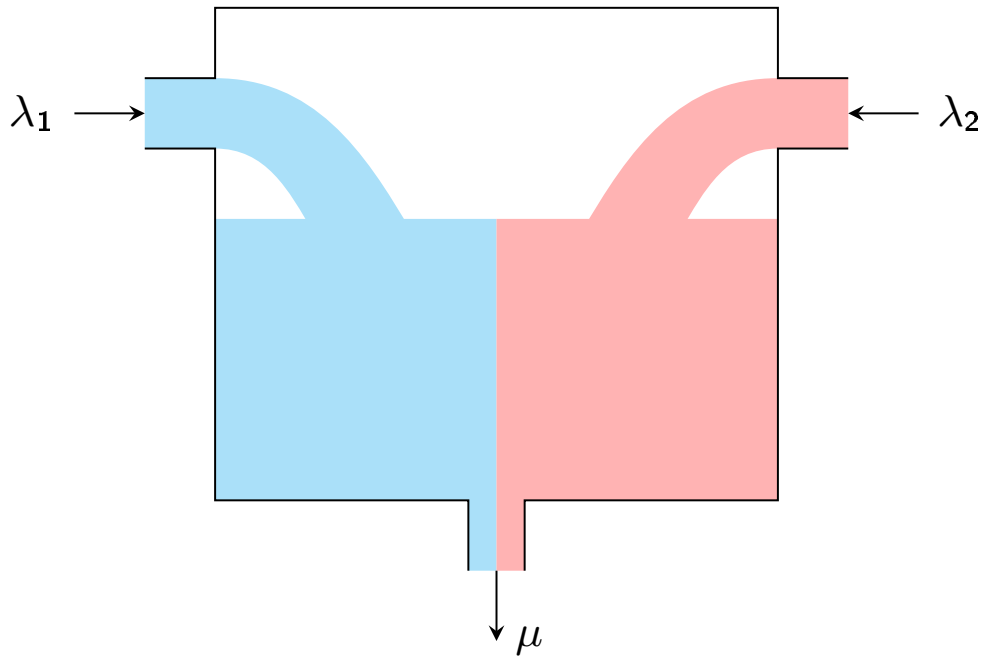
$$f_2(x) = \frac{q_{21}}{q_{12}+q_{21}}\frac{q_{12}\mu-q_{21}(\lambda-\mu)}{\mu^2}e^{-\left(\frac{q_{12}}{\lambda-\mu}-\frac{q_{21}}{\mu}\right)x},$$

$$\ell_1 = 0,$$

$$\ell_2 = \frac{q_{12}\mu-q_{21}(\lambda-\mu)}{(q_{12}+q_{21})\mu}.$$

Forgalomszabályozás

Mi történik, ha egy szerverhez több forrásból is érkezik forgalom?



Ha $\sum_i \lambda_i > \mu$, akkor forgalom prioritizálásra van szükség (congestion control).

Forgalom prioritizálás

Több forrásból érkező forgalmak prioritizálásra két fő típusú elvet szoktak használni:

- Szigorú prioritizálás. Ilyenkor az egyik forgalom teljes elsőbbséget élvez a másikkal szemben.
- WFQ (weighted fair queueing): Ilyenkor minden egyes forgalomnak van egy előre adott súlya, és a kiszolgálási sebességet a súlyok arányában kapják meg.

A gyakorlatban egy nagy hálózatban (több forgalom, több szerver) ezt úgy szokás megvalósítani, hogy minden forgalom néhány forgalmi osztály valamelyikébe tartozik, és az osztályokat prioritizáljuk, illetve azonos prioritású osztályokhoz WFQ súlyokat rendelünk (a WFQ súlyok szerverenként eltérhetnek).