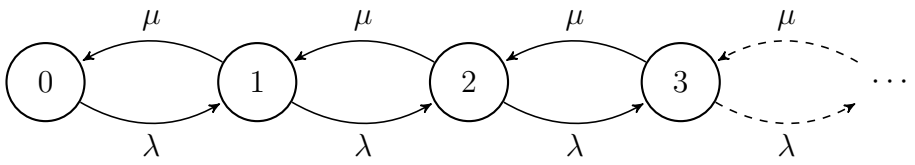
TAMS15: SS2
Köteori

Johan Thim (jothi@mai.liu.se)

21 november 2018

12.1 M/M/1 kö

Ankomsttider och väntetider är exponentialfördelade och vi kan beskriva systemet som en födelse-döds-process där vi har konstant födelseintensitet λ (ny kund anländer) och konstant dödsintensitet μ (kund har fått betjäning och går sin väg). Kedjan är oändlig (finns inget max-antal kunder i kön). Figur:



Om

$$1+\frac{\lambda}{\mu}+\frac{\lambda^2}{\mu^2}+\cdots=\sum_{k=0}^{\infty}\rho^k=\frac{1}{1-\rho}=\frac{\mu}{\mu-\lambda}<\infty,$$

det vill säga om $\rho=\lambda/\mu<1$ (dödsintensiteten större än födelseintensiteten), så uppfyller jämnviktsfördelningen $\underline{\pi}$ att

$$\pi_0=1-\rho \qquad \text{och} \qquad \pi_k=\pi_0\rho^k=(1-\rho)\,\rho^k, \quad k\geq 1.$$

Detta känner vi igen som den geometriska fördelning. Vi skriver att $X\sim \mathrm{Ge}(1-\rho)$ om

$$p_X(k)=(1-\rho)\rho^k, \quad k=0,1,2,3,\ldots.$$

Vi sammanfattar lite information om den geometriska fördelningen.

Geometrisk fördelning

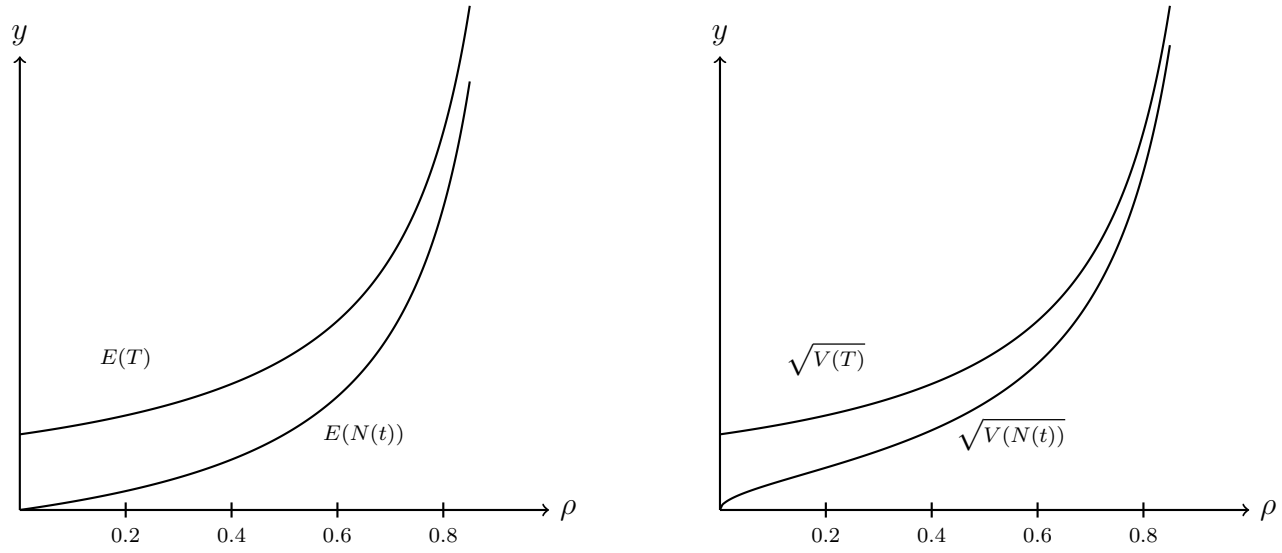
Sats. Om $X\sim \mathrm{Ge}(p)$ med $0<p<1$ så är $E(X)=\dfrac{1-p}{p}$ och $V(X)=\dfrac{1-p}{\rho^2}$.

Faktum är att vi redan bevisat uttrycket för $E(X)$ och $V(X)$ i föreläsning fem.
Vi har alltså $N(t) \sim \text{Ge}(1-\rho)$, så $E(N(t)) = \rho/(1-\rho)$. Littles sats implicerar att den förväntade totala tiden i systemet för en kund ges av

$$E(T) = \frac{E(N(t))}{\lambda} = \frac{1}{\mu - \lambda}.$$

Även varianserna kan beräknas:

$$V(N(t)) = \frac{\rho}{(1-\rho)^2} \qquad \text{och} \qquad V(T) = \frac{1}{(\mu-\lambda)^2}.$$



Vi ser i figurerna ovan att systemet blir instabilt då nyttjandegraden ρ närmar sig ett. Väntevärderna växer kraftigt och även variationen i väntetid och antal personer i systemet ökar kraftigt, vilket gör förutsägelser opålitliga. Detta är ett fenomen i de flesta kösystem.
Vi kan även räkna ut förväntad kötid och förväntat antal kunder i kön:

$$E(W) = E(T) - E(S) = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)},$$
$$E(N_q(t)) = \lambda E(W) = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{1-\rho}.$$

12.1.1 Väntetider

Antag att det är n personer i kön när nästa person anländer. Den totala tiden T i systemet för denna person är summan av $n+1$ stycken oberoende variabler $T_i \sim \text{Exp}(\mu)$. Enligt föregående föreläsning ger detta upphov till Gamma-fördelningen, så

$$F_T(t) = \int_0^t f_T(x)\,dx = 1 - e^{-\mu t}\sum_{k=0}^{n}\frac{(\mu t)^k}{k!}, \quad t \geq 0.$$

Men det är inte alltid precis n kunder före oss när vi kommer, så hur finner vi fördelningen för T utan detta antagande? Svaret kommer i form av lagen om total sannolikhet:

$$\begin{aligned}
F_T(t) = P(T \le t) &= \sum_{n=0}^{\infty} P(N(t)=n) P\big(T \le t \mid N(t)=n\big) \\
&= \sum_{n=0}^{\infty} (1-\rho)\rho^n P\big(T \le t \mid N(t)=n\big) = \sum_{n=0}^{\infty} (1-\rho)\rho^n \left(1 - e^{-\mu t} \sum_{k=0}^{n} \frac{(\mu t)^k}{k!}\right) \\
&= 1 - e^{-\mu t} \sum_{n=0}^{\infty} (1-\rho)\rho^n \sum_{k=0}^{n} \frac{(\mu t)^k}{k!} = 1 - e^{-\mu t} \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \sum_{n=k}^{\infty} (1-\rho)\rho^n \\
&= 1 - e^{-\mu t} \sum_{k=0}^{\infty} \frac{(\rho\mu t)^k}{k!} = 1 - e^{-\mu t(1-\rho)} = 1 - e^{-t(\mu-\lambda)}.
\end{aligned}$$

Vi har alltså visat att $T \sim \text{Exp}(\mu - \lambda)$, så speciellt är $E(T) = 1/(\mu - \lambda)$ (vilket vi redan visste). På samma sätt kan man visa att $F_W(w) = P(W \le w) = 1 - \rho e^{-w(\mu-\lambda)}$, $w \ge 0$. Notera att $F_W(0) = P(W = 0) = 1 - \rho$ är sannolikheten att kön är tom när en kund anländer. Vi kan även finna detta genom sannolikheten att betjäningen är upptagen:

$$P(N(t) > 0) = P(N(t) \ge 1) = 1 - \pi_0 = \rho,$$

så $\rho = \dfrac{\lambda}{\mu}$ är verkligen nytjandegraden.

Exempel

En router med stort arbetsminne (oändlig bufferkapacitet) och väldigt snabb processor har en utgående bandbredd på 100 Mbps (som routern kan fylla). Paket som kommer in till routern har en storlek som är exponentialfördelad med väntevärde 16 Kbit, och paket anländer enligt en Poissonprocess med $\lambda = 2000$ paket per sekund. Hitta en lämplig kö-modell, och svara på följande:

(a) Förväntat antal paket i routern och förväntad svarstid.
(b) Sannolikhet att ett paket tar mer än 0.5 ms att behandla.
(c) Sannolikhet att routern är ”idle” (inte har några paket i systemet).
(d) Sannolikhet att routern har mer än ett paket i systemet.

Lösning: Vi väljer en M/M/1 kö med $\lambda = 2000$ paket/s och

$$\mu = \frac{100 \text{ Mbps}}{16 \text{ Kbit}} = \frac{100000}{16} = 6250 \text{ paket/s}.$$

Låt $\rho = \lambda/\mu = 0.32$.

(a) $E(N(t)) = \dfrac{\rho}{1-\rho} \approx 0.47$. Ett halvt paket i kön alltså. Svarstiden är tiden för paketet att ta sig genom systemet, vilket blir

$$E(T) = \frac{1}{\mu - \lambda} = 0.2353 \text{ ms}.$$

(b) Sannolikheten att det tar mer än 0.5 ms för ett paket fås enligt

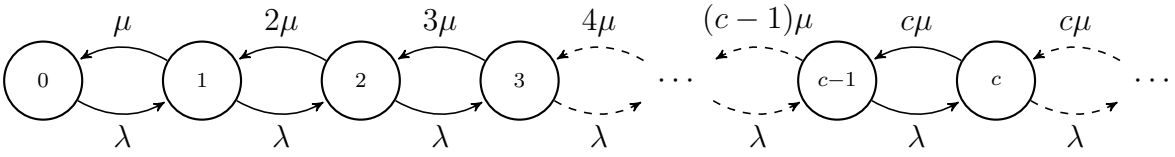
$$P(T > 0.5 \text{ ms}) = 1 - P(T \leq 0.5 \text{ ms}) = 1 - (1 - e^{-0.5(\mu-\lambda)/1000}) \approx 0.1194.$$

(c) Sannolikheten att kön är tom ges av $P(N(t) = 0) = \pi_0 = 1 - \rho = 0.68$.

(d) $P(N(t) > 1) = 1 - P(N(t) \leq 1) = 1 - \pi_0 - \pi_1 \approx 1 - 0.68 - 0.22 = 0.10$.

12.2 M/M/c kö

Ankomsttider och väntetider är exponentialfördelade och vi kan beskriva systemet som en födelse-döds-process där vi har konstant födelseintensitet λ (ny kund anländer) och en dödsintensitet som är μ vid en kund i systemet, 2μ om det är två, 3μ om det är tre och så vidare upp till c. För fler än c kunder är dödsintensiteten $c\mu$. Kedjan är oändlig (finns inget max-antal kunder i kön).



Låt $\rho = \dfrac{\lambda}{c\mu}$. För att kunna prata om en stationär fördelning är det tillräckligt att $\rho < 1$, för då är

$$\sum_{k=0}^{c-1} \frac{\lambda^k}{k!\mu^k} + \sum_{k=c}^{\infty} \frac{\lambda^k}{c!c^{k-c}\mu^k} = 1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2} + \frac{\lambda^3}{3!\mu^3} + \cdots + \frac{\lambda^c}{c!\mu^c} + \frac{\lambda^{c+1}}{c!c\mu^{c+1}} + \frac{\lambda^{c+2}}{c!c^2\mu^{c+2}} \cdots$$
$$= \sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!}\frac{1}{1-\rho} < \infty.$$

Den stationära fördelningen ges nu av

$$\pi_0 = \left(\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!}\frac{1}{1-\rho} \right)^{-1} \tag{1}$$

och

$$\pi_k = \begin{cases} \pi_0 \dfrac{(c\rho)^k}{k!}, & 1 \leq k \leq c, \\ \pi_0 \dfrac{\rho^k c^c}{c!}, & k \geq c. \end{cases}$$

Sannolikheten att alla betjäningsställen är upptagna ges av uttrycket

$$C(c, \lambda/\mu) = \sum_{k=c}^{\infty} \pi_k = \frac{c^c \sum_{k=c}^{\infty} \rho^k}{c! \left(\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!} \cdot \frac{1}{1-\rho} \right)} = \frac{(c\rho)^c}{(c\rho)^c + c!(1-\rho)\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!}}.$$

Denna formel är känd som **Erlang C-formel**. Ofta beräknas den enklast genom

$$C(c, \lambda/\mu) = \pi_0 \frac{(c\rho)^c}{c!}\frac{1}{1-\rho} = \frac{\pi_c}{1-\rho}.$$

Förväntat antal kunder ges av

$$E(N(t)) = \sum_{k=0}^{\infty} k\pi_k = \pi_0 \left(\sum_{k=0}^{c} k\frac{(c\rho)^k}{k!} + \sum_{k=c+1}^{\infty} \frac{k\rho^k c^c}{c!} \right).$$

Vi betraktar en del av högerledet i taget. Först,

$$\begin{aligned}\pi_0 \sum_{k=0}^{c} k\frac{(c\rho)^k}{k!} &= \pi_0 c\rho \sum_{k=1}^{c} \frac{(c\rho)^{k-1}}{(k-1)!} = \pi_0 c\rho \sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} = \pi_0 c\rho \left(\frac{1}{\pi_0} - \frac{(c\rho)^c}{c!}\frac{1}{1-\rho} \right) \\ &= c\rho - c\rho C(c, \lambda/\mu),\end{aligned}$$

om vi utnyttjar uttrycket för π_0 givet i (1) ovan. För den andra delen,

$$\begin{aligned}\pi_0 \sum_{k=c+1}^{\infty} \frac{k\rho^k c^c}{c!} &= \frac{\pi_c}{\rho^{c-1}} \sum_{k=c+1}^{\infty} k\rho^{k-1} = \frac{\pi_c}{\rho^{c-1}} \frac{d}{d\rho} \sum_{k=c+1}^{\infty} \rho^k \\ &= \frac{\pi_c}{\rho^{c-1}} \frac{d}{d\rho} \left(\rho^{c+1} \frac{1}{1-\rho} \right) = \frac{\pi_c}{1-\rho} \left((c+1)\rho + \frac{\rho^2}{1-\rho} \right) \\ &= \frac{\pi_c}{1-\rho} \left(c\rho + \frac{\rho}{1-\rho} \right).\end{aligned}$$

Vi summerar dessa två och finner att

$$E(N(t)) = c\rho - c\rho C(c, \lambda/\mu) + \frac{\pi_c}{1-\rho} \left(c\rho + \frac{\rho}{1-\rho} \right) = c\rho + \frac{\rho}{1-\rho} C(c, \lambda/\mu).$$

Littles sats kan nu användas för att ta fram följande uttryck för relevanta förväntade värden:

$$\begin{aligned}E(T) &= \frac{E(N(t))}{\lambda} = \frac{1}{c\mu - \lambda} C(c, \lambda/\mu) + \frac{1}{\mu}, \\ E(W) &= E(T) - E(S) = \frac{1}{c\mu - \lambda} C(c, \lambda/\mu), \\ E(N_q(t)) &= \lambda E(W) = \frac{\lambda/c}{1-\rho} C(c, \lambda/\mu).\end{aligned}$$

Man kan även härleda fördelningar för T och W likt förra fallet, men vi nöjer oss med de förväntade värderna.
En annan fördelning som dock kan vara intressant är avgångstiden X om alla betjäningar är upptagna. Vi har c stycken betjäningar och var och en har exponentialfördelad betjäningstid S_i med intensitet μ. Variabeln X kan skrivas som minimum av $S_1, S_2, \ldots, S_c$ och får då fördelningsfunktionen

$$F_X(x) = 1 - e^{-c\mu x}, \quad x \geq 0.$$

Detta är fördelningsfunktionen för en exponentialfördelad variabel med intensitet $c\mu$. Vi har alltså precis visat att $X \sim \text{Exp}(c\mu)$. Om bara k betjäningar är upptagna, $1 \leq k \leq c$, så är $X \sim \text{Exp}(k\mu)$.

Exempel

En M/M/3 kö har $\lambda = 3$ och $\mu = 2$. Vad är den förväntade kö-tiden? Vad är sannolikheten att en ny kund får stå i kö? Om alla betjäningar är fulla, vad är den förväntade tiden till en betjäning blir ledig?

Lösning: Vi har $c = 3$ och en Markov/Markov modell, så tider blir exponentialfördelade. Sannolikheten att vi får stå i kö ges av, med $c = 3$ och $\rho = 1/2$,

$$P(N(t) \geq c) = \sum_{k=c}^{\infty} \pi_k = C(c, \lambda/\mu) = 0.2368$$

eftersom

$$\pi_0^{-1} = \sum_{k=0}^{2} \frac{(3\rho)^k}{k!} + \frac{(3\rho)^3}{3!} \cdot \frac{1}{1-\rho} = 4.75$$

och

$$C(c, \lambda/\mu) = \frac{\pi_c}{1-\rho} = \pi_0 \frac{(c\rho)^c}{c!} \frac{1}{1-\rho}.$$

Förväntad kötid: $E(W) = 0.0789$. Om alla betjäningar är fulla vet vi att fördelningen för tiden till dess att någon lämnar systemet är $\text{Exp}(c\mu)$, så den förväntade tiden blir $1/c\mu = 1/6$.

12.2.1 M/M/2

Om $c = 2$ har vi med $\rho = \lambda/2\mu$ specialfallet

$$1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2} + \frac{\lambda^3}{2^2\mu^3} \cdots = 1 + 2\sum_{k=1}^{\infty} \left(\frac{\lambda}{2\mu}\right)^k = 1 + 2\frac{\rho}{1-\rho} = \frac{1+\rho}{1-\rho},$$

vilket är ändligt om $\lambda < 2\mu$. Jämnviktsfördelningen $\underline{\pi}$ uppfyller att

$$\pi_0 = \frac{1-\rho}{1+\rho} \qquad \text{och} \qquad \pi_k = 2\rho^k \cdot \frac{1+\rho}{1-\rho}, \quad k = 1, 2, 3 \ldots.$$

Formlerna för förväntade värden blir nu lite enklare:

$$E(N(t)) = \frac{2\rho}{1-\rho^2}, \qquad E(T) = \frac{1/\mu}{1-\rho^2}, \qquad E(W) = \frac{\rho^2}{\mu(1-\rho^2)}, \qquad E(N_q(t)) = \frac{2\rho^3}{1-\rho^2}.$$

Exempel

Den nuvarande databasservern har en betjäningsintensitet μ och en ankomstintensitet λ. Prestandan börjar bli för dålig och man ska uppgradera. Jämför följande val:

1. Köper en ny server med betjäningsintensitet 4μ.
2. Köper två stycken nya servrar med betjäningsintensitet 2μ och kör dessa med en gemensam kö.

Vilket alternativ ger bäst prestanda?

Lösning: Alternativ ett är en M/M/1 situation och alternativ två en M/M/2 situation. Vi jämför. Låt $\rho_1 = \dfrac{\lambda}{4\mu}$ och $\rho_2 = \dfrac{\lambda}{2 \cdot 2\mu}$. Numeriskt är alltså $\rho_1 = \rho_2$.

	$E(N(t))$	$E(T)$	$E(W)$
Alt. 1	$\dfrac{\rho_1}{1-\rho_1}$	$\dfrac{1}{4\mu(1-\rho_1)}$	$\dfrac{\rho_1}{4\mu(1-\rho_1)}$
Alt. 2	$\dfrac{2\rho_2}{1-\rho_2^2}$	$\dfrac{1}{2\mu(1-\rho_2^2)}$	$\dfrac{\rho_2^2}{2\mu(1-\rho_2^2)}$

Om vi jämför kvoten av de olika $E(N(t))$ för alternativen ser vi att

$$\frac{\text{Alt. 1}}{\text{Alt. 2}} = \frac{\frac{\rho}{1-\rho}}{\frac{2\rho}{1-\rho^2}} = \frac{1+\rho}{2} < 1,$$

eftersom $\rho < 1$. Kölängden blir alltså i snitt längre för alternativ två, vilket borde innebära sämre prestanda. Samma sak gäller övriga kvoter, båda ger samma uttryck $(1+\rho)/2 < 1$, så alternativ ett har kortare väntetid och kötid. Speciellt om belastningen är ganska liten får vi en stor skillnad.
Detta är en allmän princip. Om vi inte har ett problem som har en struktur som gör att parallell hantering snabbar upp så går det snabbare med ett dubbelt så snabbt enkelsystem som två långsammare.

12.2.2 M/M/∞

Om det finns oändligt många betjäningsställen konvergerar π_0 till

$$\pi_0 = \left(\sum_{k=0}^{\infty} \frac{(\lambda/\mu)^k}{k!} \right)^{-1} = e^{-a},$$

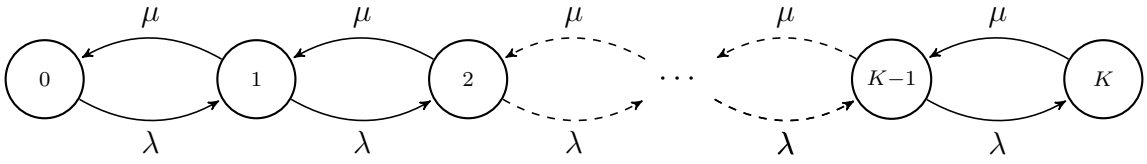
där $a = \lambda/\mu$, och därmed blir

$$\pi_k = \frac{a^k}{k!} e^{-a}, \quad k \geq 0.$$

Vi har alltså visat att $N(t) \sim \text{Po}(a)$. I detta fall kan vi alltså använda egenskaper för Poissonfördelningen (tabeller etc) för att räkna ut sannolikheter.
Så när har vi oändligt många betjäningsställen? Kanske inte så ofta i verkligheten, men om c är väldigt stort kan denna modell enkelt ge approximativa svar på parametrar av intresse.

12.3 M/M/1/K

I detta fall har vi ett betjäningsställe och en begränsad kölängd. Maximala antalet kunder i systemet är K. Typiskt exempel är dataswitchar av olika slag som har en buffert för inkommande paket men endast kan behandla ett paket i taget. Frågan blir ofta hur stor man ska dimensionera bufferten (K) för att inte tappa bort för många paket.



Låt $a = \lambda/\mu$. Den stationära fördelningen, som finns oberoende av λ och μ (varför?), ges av

$$\pi_0 = \left(\sum_{k=0}^{K} a^k \right)^{-1} = \frac{1-a}{1-a^{K+1}} \qquad \text{och} \qquad \pi_k = \pi_0 a^k, \quad 1 \le k \le K,$$

om $a \neq 1$. Om $a = 1$ blir $N(t)$ likformigt fördelad på tillstånden $\{\, 0, 1, 2, \ldots, K \,\}$.
Om $a \neq 1$ följer det att

$$\begin{aligned} E(N(t)) &= \sum_{k=0}^{K} k\pi_k = \pi_0 \sum_{k=0}^{K} k a^k = \pi_0 a \frac{d}{da} \left(\sum_{k=1}^{K} a^k \right) = \pi_0 a \cdot \frac{1 + a^K(K(a-1)-1)}{(1-a)^2} \\ &= \frac{a}{a-1} \cdot \frac{1 + a^K(K(a-1)-1)}{1-a^{K+1}} = \frac{a}{a-1} - \frac{(K+1)a^{K+1}}{1-a^{K+1}}, \end{aligned}$$

och om $a = 1$ blir $E(N(t)) = K/2$.
Den verkliga intensiteten λ_{in} in i systemet ges av

$$\lambda_{\text{in}} = \big(1 - P(N(t) = K)\big)\lambda = (1 - \pi_K)\lambda.$$

Intensiteten att kunder avvisas ges alltså av $\lambda\pi_K$.
Vi har även $E(T) = E(N(t))/\lambda_{\text{in}}$ och $E(W) = E(T) - E(S)$, samt att nyttjandegraden ρ ges av $\rho = \lambda_{\text{in}}/\mu = a(1 - \pi_K)$.

Exempel

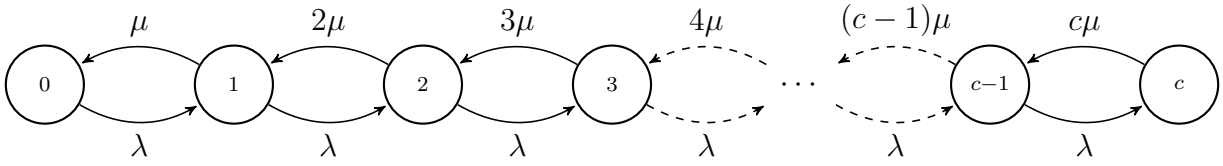
En dataswitch har förhållandet 0.5 mellan ankomstintensiteten λ och betjäningsintensiteten μ, dvs $a = \lambda/\mu = 0.5$. Hur stor buffert behövs för att sannolikheten för paketförlust blir $< 10^{-3}$?

Lösning: Sannolikheten för paketförlust är $P(N(t) = K) = \pi_K = \dfrac{(1-a)a^K}{1-a^{K+1}}$. Vi testar olika värden på K och finner att $K = 9$ stycken är den minsta längden på buffer vi kan acceptera. Sitter man vid en dator kan man använda MATLAB:

```
>> a = 0.5;
>> K = (1:1:10);
>> ((1-a).*a.^K) ./ (1 - a.^(K+1))
ans =
    0.3333    0.1429    0.0667    0.0323    0.0159    0.0079    0.0039    0.0020    0.0010    0.0005
>> ((1-a).*a.^K) ./ (1 - a.^(K+1)) < 10^(-3)
ans =
     0     0     0     0     0     0     0     0     1     1
```

12.4 M/M/c/c

Specialfallet när $K = c$ är intressant då det innebär att vi helt enkelt inte har någon kö, utan bara c betjäningsställen och kunder som anländer när det är fullt vid betjäningarna avvisas. Markovkedjan är ändlig, och kan ses nedan.



Eftersom kedjan är ändlig behövs inget extra villkor på λ och μ. Låt $a = \lambda/\mu$. Den stationära fördelningen ges av

$$\pi_0 = \left(\sum_{k=0}^{c} \frac{a^k}{k!} \right)^{-1} \qquad \text{och} \qquad \pi_k = \pi_0 \frac{a^k}{k!}, \quad 1 \leq k \leq c.$$

Det följer att

$$E(N(t)) = \sum_{k=0}^{c} k\pi_k = \pi_0 a \sum_{k=1}^{c} \frac{a^{k-1}}{(k-1)!} = \pi_0 a \sum_{k=0}^{c-1} \frac{a^k}{k!} = a \sum_{k=0}^{c-1} \pi_k = a(1 - \pi_c).$$

Det finns ingen kö, så $E(W) = 0$ och $E(T) = E(S) = \frac{1}{\mu}$.