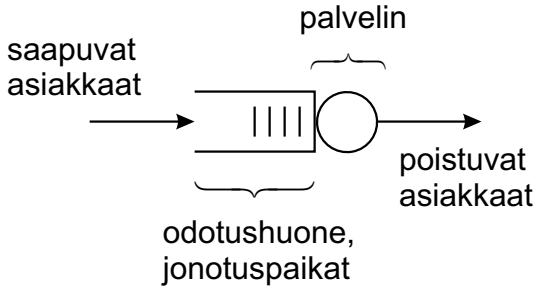
Odotusjärjestelmät

Siirrytään tarkastelemaan odotusjärjestelmiä. Nämä ovat "aitoja" jonojärjestelmiä siinä mielessä, että niissä on odotuspaikkoja ja asiakkat voivat todella joutua jonottamaan palveluun pääsyä.

Aluksi esitellään allaolevan kuvan mukaisen yhden palvelimen jonoon liittyvät perussuureet.



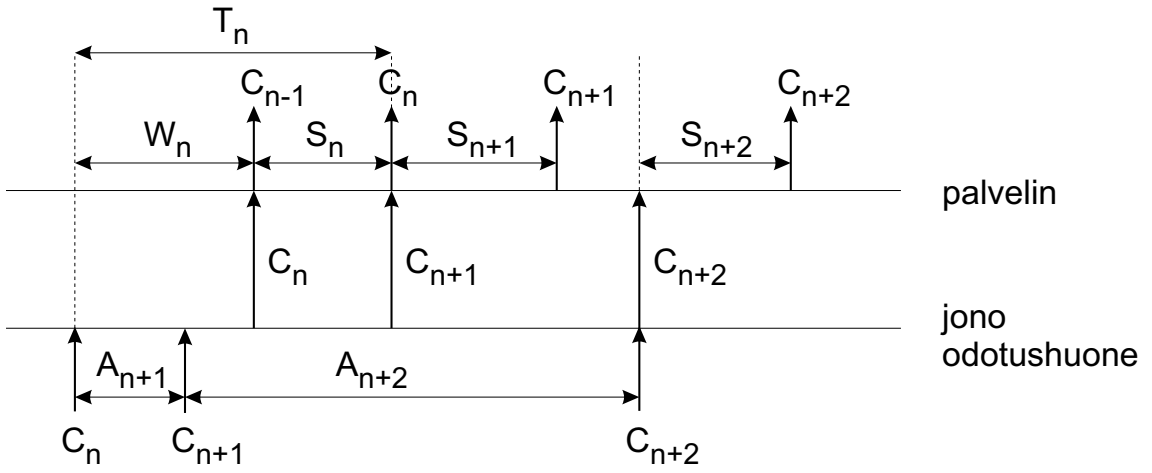
Kaksoisaika-akseli (yhden palvelimen järjestelmässä)

C_n	asiakas n
S_n	asiakkaan n palveluaika (työn purkamiseen kuluva aika)
X_n	asiakkaan n palveluvaade (palvelutyö)
W_n	asiakkaan n odotusaika
T_n	$W_n + S_n$ asiakkaan n järjestelmässä viettämä kokonaisaika
	viipymisaika, viipymä (time in system, sojourn time)
A_n	(tai t_n) asiakkaiden $n-1$ ja n saapumisten välinen aika (interarrival time)
C	palvelimen palvelunopeus eli kapasiteetti (merkitään myös c)

Palveluaika riippuu palveluvaateesta ja palvelimen nopeudesta: $S_n = X_n/C$.

Tietoliikennesovelluksissa palvelu on tyypillisesti esim. paketin lähettämistä linjalle.

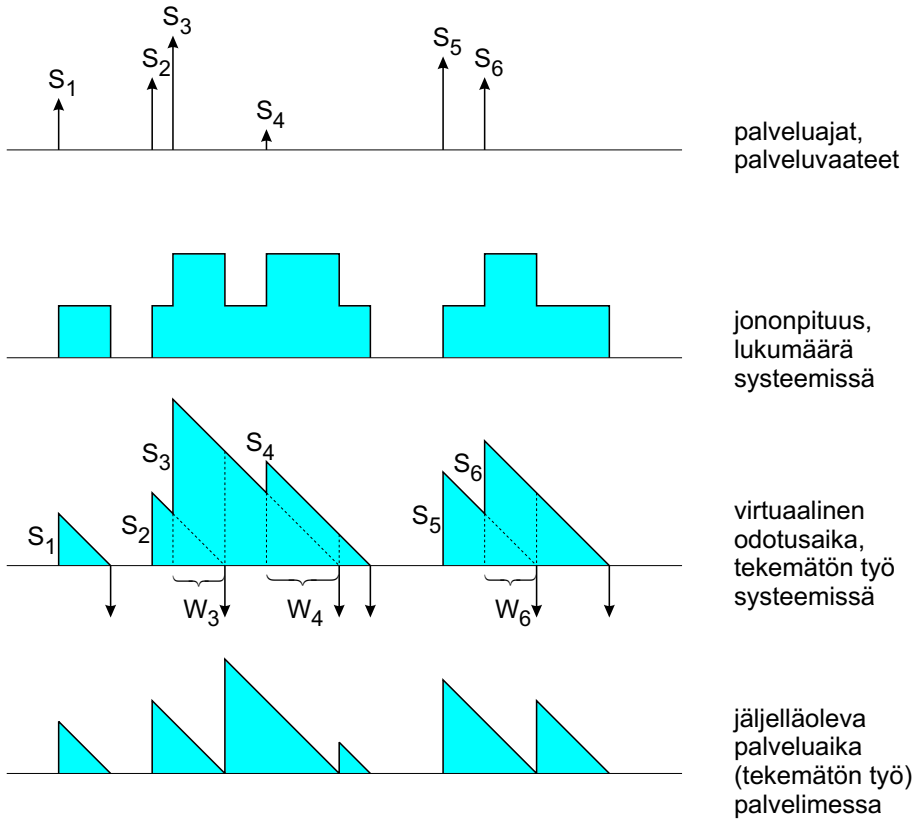
Tällöin työn yksikkö voi olla kbit ja palvelunopeuden yksikkö vastaavasti kbit/s.



Kuvasta nähdään, että FIFO:lle pätee yhtälö $\boxed{W_{n+1} = (W_n + S_n - A_{n+1})^+}$ missä $(x)^+ = \max(x, 0)$

Jononpituus, tekemätön työ ja virtuaalinen odotusaika

N_t	(tai Q_t tai L_t) asiakkaiden lukumäärä järjestelmässä ("jononpituus")
S_n	asiakkaan n palveluaika (työn purkamiseen kuluva aika)
X_t	jonossa oleva tekemätön työ hetkellä t
V_t	virtuaalinen odotusaika hetkellä t
W_n	asiakkaan n kokema odotusaika
C	palvelimen palvelunopeus eli kapasiteetti (merkitään myös c)



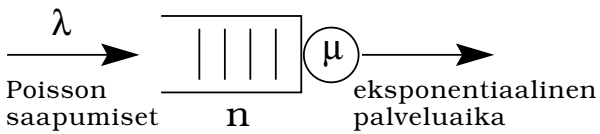
- Virtuaalinen odotusaika V_t tarkoittaa sitä aikaa, jonka hetkellä t saapuva asiakas joutuisi odottamaan (jos sattuisi juuri silloin tulemaan jonoon) FIFO-jonossa.

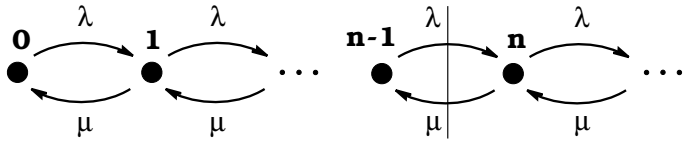
  V_t on siis jonossa sisällä olevan tekemättömän työn X_t purkamiseen kuluva aika, $V_t = X_t/C$.

- Poisson-saapumisten tapauksessa W_n:n jakauma on PASTA-ominaisuuden vuoksi sama kuin V_t:n jakauma.

$M/M/1$-jono

Asiakkaiden lukumäärä $M/M/1$-jonossa





Leikkausmenetelmää soveltamalla saadaan tasapainoehto

$\lambda\pi_{n-1} = \mu\pi_n$ eli $\pi_n = \rho\pi_{n-1}$ missä $\rho = \lambda/\mu$ (liikenneintensiteetti, kuorma),

josta rekursiivisesti

$\pi_n = \rho^n\pi_0$ (jotta jono olisi stabiili, vaaditaan $\rho < 1$)

Tyhjän jonon todennäköisyys π_0 saadaan normiehdosta $\pi_0 + \pi_1 + \pi_2 + \ldots = 1$

$$\pi_0 = 1/\sum_{n=0}^{\infty}\rho^n = 1-\rho$$

(todennäköisyys, että palvelin (jono) on tyhjä $= 1-\rho$ $\Rightarrow$ todennäköisyys, että palvelin on käytössä $= \rho$)

$M/M/1$-jonon pituusjakauma, $\pi_n = \mathrm{P}\{N = n\}$,

$$\boxed{\pi_n = (1-\rho)\,\rho^n} \quad n = 0, 1, \ldots$$

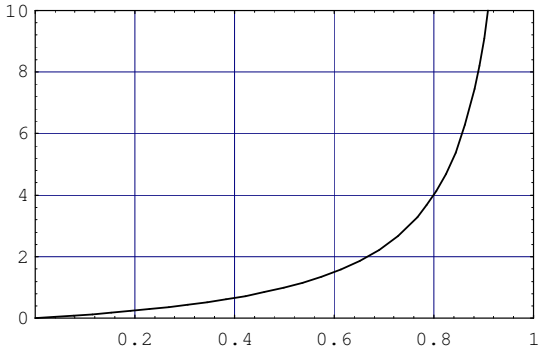
$\mathrm{Geom}_0(\rho)$-jakauma (alkaa nollasta)

Asiakkaiden keskimääräinen lukumäärä

$$\mathrm{E}[N] = \sum_{i=0}^{\infty} i\,\pi_i = (1-\rho)\sum_{i=0}^{\infty} i\rho^i = (1-\rho)\,\rho\,\frac{d}{d\rho}\sum_{i=0}^{\infty}\rho^i$$

$$= (1-\rho)\,\rho\,\frac{d}{d\rho}\Big(\frac{1}{1-\rho}\Big) = \frac{\rho}{1-\rho} \qquad \mathrm{Geom}_0(\rho)\text{-jakauman keskiarvo (alkaa 0:sta)}$$

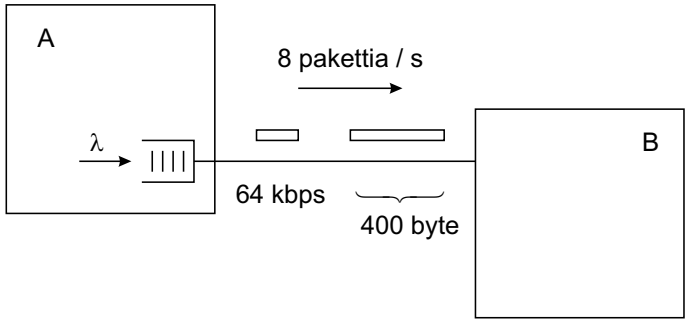
$$\mathrm{E}[N] = \frac{\rho}{1-\rho} = \underbrace{\rho}_{\substack{\text{asiakkaat}\\ \text{palvelimessa}}} + \underbrace{\frac{\rho^2}{1-\rho}}_{\substack{\text{odottavat}\\ \text{asiakkaat}}}$$



Jonon häntätodennäköisyys: todennäköisyys sille, että systeemissä on vähintään n asiakasta,

$$\mathrm{P}\{N \geq n\} = \sum_{i=n}^{\infty} \pi_i = (1-\rho)\sum_{i=n}^{\infty}\rho^i = \rho^n$$

Esimerkki.



- Reitittimeltä A lähtee keskimäärin 8 pakettia sekunnissa reitittimelle B.
- Paketin keskikoko on 400 byteä (jakautuma eksponentiaalinen).
- Linjan nopeus on 64 kbit/s.

Kuinka monta pakettia keskimäärin on reitittimessä A lähetysvaiheessa ja mikä on todennäköisyys, että pakettien lukumäärä on 10 tai enemmän?

Linjan eli palvelimen kuormitusaste on

$$\rho = 8\ \mathrm{s}^{-1} \times 400 \times 8\ \mathrm{bit}/64{\times}10^3\ \mathrm{bit\ s}^{-1} = 0.4.$$

Tämä voidaan laskea myös muodossa λ/μ, missä

$\lambda = 8$ pakettia/s, $\mu = 64$ kbit/s/(400×8 bit/paketti) $= 20$ pakettia/s $\Rightarrow$ $\lambda/\mu = 8/20 = 0.4$

Tällöin on $\underline{\mathrm{E}[N] = 0.4/(1-0.4) = 0.67}$.

Todennäköisyys, että paketteja on 10 tai enemmän on $\underline{0.4^{10} = 10^{-4}}$.

$M/M/1$-jonon viipymis- ja odotusajat

Littlen tulos:

Keskimääräinen viipymä systeemissä $\mathrm{E}[T] = \mathrm{E}[N]/\lambda$

Keskimääräinen odotusaika $\mathrm{E}[W] = (\mathrm{E}[N] - \rho)/\lambda$

$$\boxed{\mathrm{E}[T] = \frac{1}{1-\rho} \cdot \frac{1}{\mu} = \frac{1}{\mu - \lambda}}$$

$$\boxed{\mathrm{E}[W] = \frac{\rho}{1-\rho} \cdot \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}}$$

Riippumattomuus jonokurista

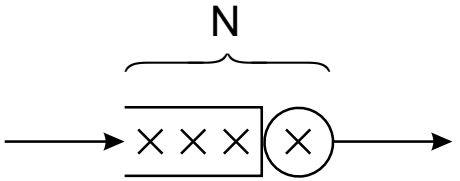
$M/M/1$-FIFO-jonolle on edellä johdettu jononpituusjakauma $\pi_n = (1-\rho)\rho^n$.

- Tämä jakauma on riippumaton jonokurista (FIFO, LIFO, PS),
  - kaikilla näillä jonokureilla tasapainoyhtälöt ovat täsmälleen samat (todistus harjoitustehtävänä)
- Siten myös keskimääräinen aika systeemissä, $\mathrm{E}[T] = 1/(\mu - \lambda)$, on riippumaton jonokurista (Littlen tuloksen perusteella keskimääräinen aika systeemissä on keskimääräinen jononpituus jaettuna λ:lla.)
- Sen sijaan mm. W:n ja T:n jakaumat riippuvat jonokurista.

Huom. Jononpituusjakauma itse ei ole insensitiivi palveluaikajakauman suhteen $M/M/1$-FIFO-jonossa. Sen sijaan kyseinen jakauma pätee LIFO- ja PS-jonoissa palveluaikajakaumasta riippumatta.

Viipymäajan jakauma

Oletetaan, että saapuva asiakas näkee edellään N asiakasta (mukaanlukien palvelimessa oleva asiakas).



Eksponenttijakauman muistittomuuden vuoksi myös palveltavana olevan asiakkaan (jos sellainen on) jäljelläoleva palveluaika $\sim \text{Exp}(\mu)$.

Aika T, jonka asiakas viettää systeemissä, muodostuu edelläolevien asiakkaiden ja sen omasta palveluajasta

$$T = \underbrace{S_1 + S_2 + \cdots + S_N}_{\text{edelläolevat}} + \underbrace{S_{N+1}}_{\text{oma}} \qquad (N+1)\text{:n Exp}(\mu)\text{-jakautuneen sm:n summa}$$

$$\begin{cases} S_i \sim \text{Exp}(\mu) & \text{riippumattomia} \\ N \sim \text{Geom}_0(\rho) & \text{jononpituuden tasapainojakauma (alkaa nollasta), PASTA!} \end{cases}$$

$$\begin{aligned} f_T(t) &= \sum_{n=0}^{\infty} f_{T|N}(t,n)\text{P}\{N=n\} = \sum_{n=0}^{\infty} \overbrace{\mu\frac{(\mu t)^n}{n!}e^{-\mu t}}^{\text{Erlang}(n+1,\mu)}(1-\rho)\rho^n \\ &= \mu(1-\rho)e^{-\mu t}\sum_{n=0}^{\infty}\frac{(\mu\rho t)^n}{n!} = \mu(1-\rho)e^{-\mu(1-\rho)t} \end{aligned}$$

$$\boxed{f_T(t) = (\mu-\lambda)e^{-(\mu-\lambda)t}} \qquad \text{eksponenttijakauma Exp}(\mu-\lambda)$$

Viipymäajan jakauma (jatkoa)

Saman tuloksen voi johtaa käyttäen satunnaissumman Laplace-muunnokselle aikaisemmin johdettua tulosta.

$$\begin{cases} \mathcal{G}_{N+1}(z) = \dfrac{(1-\rho)z}{1-\rho z} & N+1 \sim \text{Geom}(1-\rho), \text{ alkaa ykkösestä} \\ f_S^*(s) = \dfrac{\mu}{\mu+s} \end{cases}$$

$$\begin{aligned} f_T^*(s) &= \mathcal{G}_{N+1}(f_S^*(s)) = \frac{(1-\rho)\frac{\mu}{\mu+s}}{1-\rho\frac{\mu}{\mu+s}} = \frac{\mu-\lambda}{(\mu+s)-\lambda} = \frac{(\mu-\lambda)}{(\mu-\lambda)+s} \\ &\Rightarrow \quad \sim \text{Exp}(\mu-\lambda) \end{aligned}$$

Odotusajan jakauma

Odotusaika W muodostuu jonossa saapumishetkellä olevien asiakkaiden palveluajoista

$$W = S_1 + \cdots + S_N \ , \quad \text{missä } S_i \sim \text{Exp}(\mu) \text{ ja } N \sim \text{Geom}_0(1-\rho) \text{ (alkaa nollasta)}$$

Jos $N = 0$ niin summassa ei ole yhtään termiä ja $W = 0$.

Lasketaan W:n häntäjakauma ehdollistamalla

$$\begin{aligned} \text{P}\{W > t\} &= \underbrace{\text{P}\{W > t \,|\, N = 0\}}_{0} \text{P}\{N = 0\} + \text{P}\{W > t \,|\, N > 0\} \underbrace{\text{P}\{N > 0\}}_{\rho} \\ &= \rho \cdot \text{P}\{W > t \,|\, N > 0\} \end{aligned}$$

Geometrisen jakauman muistittomuuden johdosta N ehdolla $N > 0$ on jakautunut kuten $\text{Geom}(\rho)$ (alkaa ykkösestä).

Siten summa $S_1 + \cdots + S_N$ ehdolla $N > 0$ on jakautunut täsmälleen kuten $S_1 + \cdots + S_{N+1}$ edellä eli noudattaa $\text{Exp}(\mu - \lambda)$-jakaumaa.

$$\boxed{\text{P}\{W > t\} = \rho \, e^{-(\mu-\lambda)t}}$$

Odotusaika on 0 äärellisellä todennäköisyydellä $\text{P}\{W = 0\} = 1 - \text{P}\{W > 0\} = 1 - \rho$.
Tämä on tietenkin sama kuin jonon tyhjänäolotodennäköisyys $\text{P}\{N = 0\}$.

Äärellinen jono: $M/M/1/K$-järjestelmä

Oletetaan, että systeemipaikkoja on K (odotuspaikat + palvelin)

Tasapainoyhtälöt leikkauksissa ovat täsmälleen samat kuin ennenkin

$$\pi_n = \rho^n \pi_0, \qquad n = 0, 1, \ldots, K$$

Ero on vain normituksessa

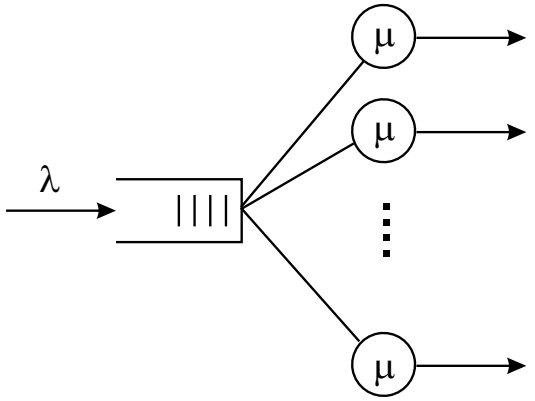
$$\sum_{n=0}^{K} \pi_n = 1 \quad \Rightarrow \quad \pi_0 = (1 + \rho + \cdots + \rho^K)^{-1} = \frac{1-\rho}{1-\rho^{K+1}}$$

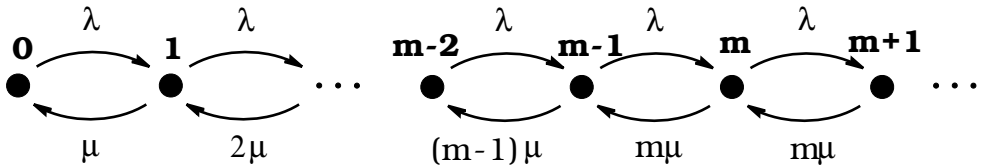
$$\boxed{\pi_n = \frac{\rho^n}{1 + \rho + \cdots + \rho^K} = \frac{1-\rho}{1-\rho^{K+1}}\,\rho^n} \qquad n = 0, 1, \ldots, K \quad \text{katkaistu geom. jakauma}$$

- Tilan K todennäköisyys π_K on todennäköisyys sille, että saapuva asiakas näkee järjestelmän täynnä ("puskuri vuotaa yli").
- Kun $K = 1$, kysymyksessä on yhden palvelimen menetysjärjestelmä,
  $\pi_n = \dfrac{\rho^n}{1+\rho},\ n = 0, 1.$

***M/M/m*-jono** (Erlangin odotusjärjestelmä)

- m rinnakkaista palvelijaa
- Poisson-saapumiset
- Eksponentiaalinen palveluaika





- Tilakaavio on tilaan m asti samanlainen kuin estojärjestelmässä.
- Siitä eteenpäin identtinen sellaisen $M/M/1$-jonon tilakaavion kanssa, jossa palvelimen kapasiteetti on $m\mu$.

Tasapainoyhtälöt voidaan jälleen kirjoittaa leikkausmenetelmää käyttäen:

$$\begin{cases} \lambda\pi_{n-1} &= n\mu\pi_n, & n \le m \\ \lambda\pi_{n-1} &= m\mu\pi_n, & n > m \end{cases}$$

Vakiotekijää π_0 vaille määrätty ratkaisu on

$$\begin{cases} \pi_n &= \pi_0 \dfrac{(m\rho)^n}{n!}, & n \le m \\ \pi_n &= \pi_0 \dfrac{m^m\rho^n}{m!}, & n > m \end{cases}$$

$a = \lambda/\mu$ liikenneintensiteetti
$\rho = \lambda/m\mu = a/m$ liikenneintensiteetti palvelinta kohden.

Tilan 0 todennäköisyys π_0 määräytyy normiehdosta $\Sigma_n \pi_n = 1$,

$$\pi_0 = \Big(\underbrace{\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!}}_{u} + \underbrace{\frac{(m\rho)^m}{m!(1-\rho)}}_{v} \Big)^{-1}$$

Todennäköisyys P_q sille, että asiakkaan saapuessa kaikki palvelijat ovat varattuja ja asiakas joutuu odottamaan on

$$\boxed{P_q = C(m,a) = \sum_{n=m}^{\infty} \pi_n = \sum_{n=m}^{\infty} \frac{\pi_0 m^m \rho^n}{m!} = \frac{\pi_0 (m\rho)^m}{m!(1-\rho)} = \frac{v}{u+v}}$$

Erlangin C-kaava

$a = m\rho;\ \rho = a/m$

Odottavien asiakkaiden keskimääräinen lukumäärä $\bar{N}_q$

$$\bar{N}_q = \sum_{n=0}^{\infty} n\pi_{m+n} = \sum_{n=0}^{\infty} n\pi_0 \frac{m^m \rho^{m+n}}{m!} = P_q \sum_{n=0}^{\infty} n(1-\rho)\rho^n$$

Summa on muodoltaan samanlainen kuin $M/M/1$-jonon keskipituuden lauseke, joten

$$\boxed{\bar{N}_q = P_q \frac{\rho}{1-\rho}} \qquad \bar{N} = m\rho + \bar{N}_q \quad \Rightarrow \quad \boxed{\bar{N} = m\rho + P_q \frac{\rho}{1-\rho}}$$

Soveltamalla Littlen tulosta saadaan keskimääräinen odotusaika sekä aika systeemissä:

$$\boxed{\begin{cases} \bar{W} = \dfrac{\bar{N}_q}{\lambda} = P_q \cdot \dfrac{1}{m\mu - \lambda} \\ \bar{T} = \dfrac{\bar{N}}{\lambda} = \dfrac{1}{\mu} + \bar{W} = \dfrac{1}{\mu} + P_q \cdot \dfrac{1}{m\mu - \lambda} \end{cases}}$$

Odotusajan jakauma

$$\mathrm{P}\{W > t\} = \underbrace{\mathrm{P}\{W > t \,|\, N < m\}}_{0}\,\mathrm{P}\{N < m\} + \mathrm{P}\{W > t \,|\, N \geq m\}\,\underbrace{\mathrm{P}\{N \geq m\}}_{P_q}$$

$$\boxed{\mathrm{P}\{W > t\} = P_q e^{-(m\mu-\lambda)t}}$$

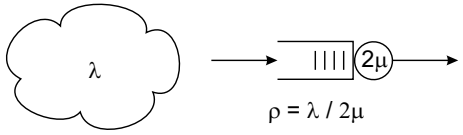
Kun $N \geq m$, systeemi käyttäytyy kuten $M/M/1$-jono, jonka kapasiteetti on $m\mu$.

Esimerkki 1

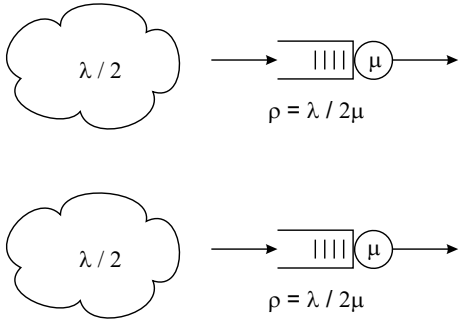
Osaston lähiverkkoon on liitetty yhteisessä käytössä oleva tulostin. Tulostustöiden oletetaan saapuvan poissonisesti intensiteetillä λ ja tulostustöiden keston oletetaan noudattavan eksponentiaalista jakaumaa $\mathrm{Exp}(\mu)$.

Tulostimen kapasiteetti on käynyt riittämättömäksi kasvaneeseen kuormaan nähden. Tulostuspalvelun parantamiseksi, tarjolla on kolme vaihtoehtoa:

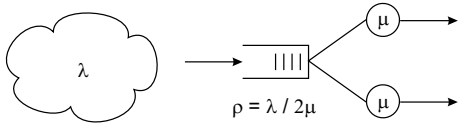
1. Hankitaan uusi kaksi kertaa nopeampi tulostin, jonka palvelunopeus 2μ.



2. Hankitaan rinnalle toinen samanlainen tulostin (palvelunopeus μ), ja jaetaan käyttäjäkunta kahteen yhtäsuureen osaan, joiden kummankin työt ohjataan omalle tulostimelleen. Kummallekin tulostimelle saapuu töitä intensiteetillä $\lambda/2$.



3. Sama kuin edellisessä kohdassa, mutta kaikki työt ohjataan tulostimille yhteisen tulostinjonon kautta, josta jonon ensimmäinen työ lähetetään aina vapautuvalle tulostimelle.



Esimerkki 1 (jatkoa)

Verrataan vaihtoehtojen suoritukykyä eri kuormilla. Suorituskyvyn mittana käyteään tässä tulostustyön keskimääräistä valmistumisaikaa $\bar{T}$ (aika systeemissä eli aika työn saapumisesta jonoon siihen, kun työ valmistuu).

1. Tässä tapauksessa kysymyksessä on $M/M/1$-jono, parametrit λ ja 2μ.

$$\rho = \frac{\lambda}{2\mu} \qquad \bar{T}_1 = \frac{1}{2\mu - \lambda} = \frac{1}{1-\rho} \cdot \frac{1}{2\mu}$$

2. Tässä tapauksessa kysymyksessä on kaksi erillistä $M/M/1$-jonoa, parametrit $\lambda/2$ ja μ.

$$\rho = \frac{\lambda/2}{\mu} = \frac{\lambda}{2\mu} \qquad \bar{T}_2 = \frac{1}{\mu - \lambda/2} = \frac{1}{1-\rho} \cdot \frac{1}{\mu}$$

   Kuorma palvelinta kohden on sama kuin edellä, nyt vain kaikki tapahtuu kaksi kertaa hitaammin (sekä saapumiset että palvelu).

3. Yhteisen tulostusjono tapauksessa sopiva jonomalli on $M/M/2$, parametrit λ ja μ.

$$\rho = \frac{\lambda}{2\mu} \qquad \bar{T}_3 = \frac{1}{\mu} + P_q \frac{1}{2\mu - \lambda} \approx \begin{cases} \dfrac{1}{\mu} & \rho \ll 1 \\ \dfrac{1}{1-\rho} \cdot \dfrac{1}{2\mu} & \rho \approx 1 \end{cases}$$

Esimerkki 1: Yhteenveto vertailusta

Otetaan tapaus 1 referenssiksi: lasketaan tapausten 2 ja 3 valmistumisaikojen suhde vastaavaan aikaan tapauksessa 1.

	T_2/T_1	T_3/T_1
$\rho \ll 1$	2	2
$\rho \approx 1$	2	1

- Vaihtoehto 1 eli yksi nopea tulostin on paras.
- Vaihtoehdossa 2 suoritusaika on aina kaksinkertainen tapaukseen 1 nähden.
- Tapauksessa 3 toisesta tulostimesta ei ole apua pienellä kuormalla: työ pääsee aina suoraan tulostukseen (ilman jonotusta) mutta itse tulostus on kaksi kertaa hitaampaa kuin tapauksen 1 nopealla tulostimella.
- Raskaalla kuormalla tapauksen 3 keskimääräinen valmistumisaika on sama kuin tapauksessa 1. Kaksi yhteisen jonon kautta syötettyä hitaampaa rinnakkaista tulostinta purkaa jonossa olevaa työtä yhtä tehokkaasti kuin yksi nopea.
- Näin ei ole asian laita tapauksessa 2. Kahden eri jonon tapauksessa on aina mahdollista, että toinen tulostin seisoo tyhjänä vaikka toisella tulostimella töitä on jonossa. Tämä heikentää suorituskykyä niin, että raskaallakin kuormalla vaihtoehto 2 on kaksi kertaa hitaampi kuin vaihtoehto 1.

Esimerkki 2

- Puhelinkeskusta mallinnetaan $M/M/m$-järjestelmänä (kaikkien linjojen ollessa varattuja, soittajaa odotutetaan antamalle hänelle soittoääntä).
- Kuinka monta johtoa (m) tarvitaan, jotta todennäköisyys sille, että asiakas joutuu odottamaan kauemmin kuin ajan $t_{\max}$ on alle 1 % ?

$$P_q e^{-(m\mu-\lambda)t_{\max}} < 0.01 \qquad \Rightarrow \qquad m > \frac{\log(100P_q) + \lambda t_{\max}}{\mu t_{\max}}$$

P_q on m:n funktio (monotonisesti pienenevä), joten epäyhtälö on vielä implisiittinen.

Se voidaan ratkaista yksinkertaisimmin kokeilemalla järjestyksessä arvoja $m = 1, 2, 3, \ldots$, kunnes epäyhtälö toteutuu.

Sallimalla asiakkaiden jonkin verran odottaa mahdollista resurssien vapautumista, voidaan estyvien asiakkaiden määrää merkittävästi pienentää tai kääntäen järjestelmän kuormitusastetta ρ voidaan nostaa verrattuna puhtaaseen menetysjärjestelmään, jossa on sama estotodennäköisyys.