

Prioriteettijonot

Tarkastellaan $M/G/1$ -jonojärjestelmää, jossa asiakkaat on jaettu K :hon prioriteettiluokkaan, $k = 1, \dots, K$:

- luokalla 1 on korkein prioriteetti ja luokalla K matalin prioriteetti,
- eri luokkien saapumisnopeudet ovat $\lambda_1, \dots, \lambda_K$,
- palveluajan odotusarvo ja toinen momentti eri luokissa: \bar{S}_k ja \bar{S}_k^2 , $k = 1, \dots, K$.

Tavoitteena on johtaa tällaiselle jonojärjestelmälle Pollaczek-Khinchinin kaavan tyyppisiä keskiarvotuloksia.

- Prioriteettisysteemit ovat tulossa yhä tärkeämmiksi myös tietoliikennejärjestelmissä
 - tietokonejärjestelmissä (esim. käyttöjärjestelmät) niitä on käytetty jo pitkään
- Nyt kiinnitetään huomio ns. aikaprioriteettiin, joka määrittelee palvelujärjestyksen
 - antamalla asiakkaalle korkeampi prioriteetti halutaan vähentää viivettä ja viiveen vaihtelua
- Kun jonosysteemillä on äärellinen koko (esim. puskurin koko), on erillinen kysymys, miten kontrolloidaan estymisiä (ylivuoto); tällöin puhutaan tilaprioriteetista.

Ei-syrjäyttävä prioriteetti (nonpreemptive priority)

- Palvelussa olevan asiakkaan palvelu suoritetaan loppuun vaikka jonoon tulisikin korkeamman prioriteetin asiakkaita.
- Jokaisella prioriteetilla on oma (looginen) jononsa.
- Palvelimen vapautuessa palveluun otetaan asiakas korkeimman prioriteetin ei-tyhjästä jonosta.

Merkinnät

$$\left\{ \begin{array}{l} \bar{N}_q^{(k)} = \text{luokan } k \text{ odottavien asiakkaiden keskimääräinen lukumäärä jonossa} \\ \bar{W}_k = \text{luokan } k \text{ asiakkaiden keskimääräinen odotusaika} \\ \rho_k = \text{luokan } k \text{ kuorma, } \rho_k = \lambda_k \bar{S}_k \\ \bar{R} = \text{palvelimen keskimääräinen jäljellä oleva palveluaika} \end{array} \right.$$

Jonon stabiilisuusehto:

$$\rho_1 + \cdots + \rho_K < 1$$

Jos ehto ei ole voimassa, jotakin luokkaa k alempien prioriteettien (suurempi k) jonot kasvavat rajatta.

Ei-syrjäyttävä prioriteetti (jatkoa)

Samaan tapaan kuin Pollaczek-Khinchinin keskiarvokaavan johdossa päätellään korkeimmalle prioriteettiluokalle 1:

$$\bar{W}_1 = \bar{R} + \bar{S}_1 \bar{N}_q^{(1)} \quad \text{missä jälkimmäinen termi edustaa jonossa edellä olevien luokan 1 asiakkaiden palvelemiseen keskimäärin kuluva aika}$$

Littlen lauseen perusteella pätee

$$\bar{N}_q^{(1)} = \lambda_1 \bar{W}_1 \quad \Rightarrow \quad \bar{W}_1 = \bar{R} + \rho_1 \bar{W}_1 \quad \Rightarrow \quad \boxed{\bar{W}_1 = \frac{\bar{R}}{1 - \rho_1}}$$

Prioriteettiluokalle 2 saadaan samaan tapaan

$$\bar{W}_2 = \bar{R} + \underbrace{\bar{S}_1 \bar{N}_q^{(1)} + \bar{S}_2 \bar{N}_q^{(2)}}_{\text{edellä olevien luokkien 1 ja 2 asiakkaiden palveluun kuluva aika}} + \underbrace{\bar{S}_1 \lambda_1 \bar{W}_2}_{\text{luokan 2 asiakkaan odotusaikana saapuvien ylemmän luokan asiakkaiden palveluun keskimäärin kuluva aika}}$$

Littlen lauseesta seuraa jälleen

$$\bar{N}_q^{(2)} = \lambda_2 \bar{W}_2 \quad \Rightarrow \quad \bar{W}_2 = \bar{R} + \rho_1 \bar{W}_1 + \rho_2 \bar{W}_2 + \rho_1 \bar{W}_2 \quad \Rightarrow \quad \bar{W}_2 = \frac{\bar{R} + \rho_1 \bar{W}_1}{1 - \rho_1 - \rho_2}$$

Ei-syrjäyttävä prioriteetti (jatkoa)

Sijoittamalla saatuun \overline{W}_2 :n lausekkeeseen \overline{W}_1 :n kaava saadaan

$$\overline{W}_2 = \frac{\overline{R}}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}$$

Jatkamalla samalla tavalla alempiin prioriteettiluokkiin (suurempiin k :n arvoihin) saadaan yleinen tulos

$$\overline{W}_k = \frac{\overline{R}}{(1 - \rho_1 - \dots - \rho_{k-1})(1 - \rho_1 - \dots - \rho_k)}$$

Luokan k asiakkaan kokonaisviipymä systeemissä keskimäärin on

$$\overline{T}_k = \overline{W}_k + \overline{S}_k$$

\overline{W}_k :n lausekkeessa esiintyvä jäljelläolevan palveluajan odotusarvo \overline{R} voidaan johtaa samanlaisen "kolmiotarkastelun" avulla kuin Pollaczek-Khinchinin keskiarvokaavan tapauksessa:

$$\overline{R} = \frac{1}{2} \sum_{k=1}^K \lambda_k \overline{S}_k^2$$

Havaintoja ei-syrjäyttävästä prioriteetista

- Analyysiä ei ole helppo laajentaa monen palvelimen jonoihin
 - jäännösaikaa \bar{R} on vaikeaa määrätä
 - onnistuu kuitenkin, jos kaikilla luokilla palveluaika on samalla keskiarvolla eksponentiaalisesti jakautunut
- Asiakkaan keskimääräistä odotusaikaa voidaan säädellä prioriteettiluokkien valinnalla
 - jos lyhyen palveluajan asiakkaille annetaan etusija, niin koko asiakaspopulaation yli laskettu keskimääräinen odotusaika lyhenee
 - vrt. kopiokonejono, jossa väliin päästetään asiakas, jolla on vain yksi kopioitava sivu
 - kahden luokan tapauksessa keskimääräinen viipymäaika on

$$\bar{T} = \frac{\lambda_1 \bar{T}_1 + \lambda_2 \bar{T}_2}{\lambda_1 + \lambda_2}$$

- voidaan osoittaa, että jos $\bar{S}_1 < \bar{S}_2$, niin \bar{T} on pienempi kuin tapauksessa, jossa prioriteetit vaihdettaisiin (tai jos prioriteetteja ei käytetä ollenkaan)
- Myös korkeimman luokan 1 odotusaika riippuu alempien luokkien liikenteestä (λ_k -arvoista), koska ei-syrjäytyvyyden vuoksi alemmat luokat eivät ole täysin “näkymättömiä” ylemmille luokille.

Kleinrockin säilymislause ei-syrjäyttävälle prioriteettijonoille

Kleinrockin säilymislauseen mukaan pätee

$$\boxed{\sum_{k=1}^K \rho_k \bar{W}_k = \frac{\rho \bar{R}}{1 - \rho}}$$

missä $\rho = \rho_1 + \dots + \rho_K$ on jonon kokonaiskuorma.

- Kuormilla painotettua odotusaikojen summaa ei voida muuttaa määriteltiinpä prioriteetit miten hyvänsä.
- Jonosysteemin muuttaminen jonkin \bar{W}_k :n pienentämiseksi johtaa väistämättä jonkun tai joidenkin muiden odotusaikojen kasvuun.

Jos erityisesti kaikilla luokilla on sama keskim. palveluaika, $\bar{S}_k = s$, saadaan jakamalla s :llä

$$\frac{\lambda \bar{R}}{1 - \rho} = \sum_{k=1}^K \lambda_k \bar{W}_k \stackrel{\text{(Little)}}{=} \sum_{k=1}^K \bar{N}_q^{(k)} = \bar{N}_q$$

Toisin sanoen jonottavien asiakkaiden kokonaismäärän keskiarvo \bar{N}_q on prioriteettijaosta riippumaton vakio (saman keskimääräisen palveluajan tapauksessa).

Kleinrockin säilymislauseen todistus

Keskimääräinen tekemätön työ \bar{V} systeemissä voidaan jakaa kahteen osaan

$$\bar{V} = \bar{V}_q + \bar{R}$$

Jonossa keskimäärin oleva tekemätön työ \bar{V}_q voidaan kirjoittaa Littlen lauseen avulla

$$\bar{V}_q = \sum_k \bar{N}_q^{(k)} \bar{S}_k = \sum_k \lambda_k \bar{W}_k \bar{S}_k = \sum_k \rho_k \bar{W}_k$$

Siten on

$$\sum_k \rho_k \bar{W}_k = \bar{V} - \bar{R}$$

\bar{V} on tunnetusti palvelujärjestyksestä riippumaton (pätee koko prosessille $V(t)$). Niin on myös \bar{R} , koska kaikki asiakkaat lopulta kulkevat palvelimen kautta. Tämä osoittaa, että $\sum_k \rho_k \bar{W}_k$ on vakio.

Arvoa muuttamatta \bar{V} voidaan laskea millä palvelujärjestyksellä tahansa. Lasketaan se erityisesti tavalliselle FIFO-jonolle ($M/G/1$ -jono), jossa on vain yksi luokka.

$$\bar{V} = \sum_k \rho_k \bar{W}_k + \bar{R} = \rho \bar{W} + \bar{R} \stackrel{\text{(PASTA)}}{=} \rho \bar{V} + \bar{R} \quad \Rightarrow \quad \bar{V} = \frac{\bar{R}}{1 - \rho}$$

Sijoittamalla aikaisempaan yhtälöön seuraa haluttu tulos $\sum_k \rho_k \bar{W}_k = \frac{\rho \bar{R}}{1 - \rho}$

Syrjäyttävä prioriteetti (preemptive resume priority)

Palvelussa olevan asiakkaan palvelu keskeytyy jonkun korkeamman prioriteetin asiakkaan saapuessa ja jatkuu siitä mihin se jäi, kun korkeamman prioriteetin jonot ovat tyhjentyneet

- Alemmat prioriteetit ovat tässä tapauksessa täysin “näkymättömiä” eivätkä vaikuta mitenkään ylempien prioriteettien toimintaan.
- Pakettijonon tapauksessa syrjäyttävää prioriteettia lähestytään, kun pakettien lähetys tapahtuu pieninä paloina, esim. ATM-soluina, joilla on prioriteetit
 - korkeamman prioriteetin paketin saapuessa alemman prioriteetin solujen lähetys keskeytyy ja jatkuu vasta, kun korkeamman prioriteetin paketit on kokonaan lähetetty

Syrjäyttävä prioriteetti (jatkoa)

Lasketaan luokan k asiakkaan keskimääräinen viipymä \bar{T}_k . Tämä muodostuu kolmesta osasta:

1. Asiakkaan oma keskimääräinen palveluaika \bar{S}_k
2. Jonossa edellä olevien, luokkiin $1, \dots, k$ kuuluvien asiakkaiden palveluun keskimäärin kuluva aika

$$\frac{\bar{R}_k}{1 - \rho_1 - \dots - \rho_k}, \text{ missä } \bar{R}_k = \frac{1}{2} \sum_{i=1}^k \lambda_i \bar{S}_i^2,$$

joka on sama kuin keskimääräinen odotusaika $M/G/1$ -jonossa, jonka muodostavat vain luokkien $1, \dots, k$ asiakkaat. Tämä johtuu siitä, että luokkien $1, \dots, k$ tekemätön työ systeemissä $V_k(t)$ (johon luokat $k+1, \dots, K$ eivät vaikuta) ei riipu luokkien $1, \dots, k$ keskinäisestä palvelujärjestyksestä (tekemätön työ on “anonyymiä”),

$$E^*[V_k(t)]_{\text{pr.jono}} = E^*[V_k(t)]_{M/G/1} = E[W_k]_{M/G/1}$$

3. Niiden ylempiin prioriteettiluokkiin $1, \dots, k-1$ kuuluvien asiakkaiden palveluun keskimäärin kuluva aika, jotka saapuvat k -luokan asiakkaan systeemissäoloajan kuluessa

$$\sum_{i=1}^{k-1} \bar{S}_i \lambda_i \bar{T}_k = \sum_{i=1}^{k-1} \rho_i \bar{T}_k, \quad k > 1 \quad (0, \text{ jos } k = 1)$$

Syrjäyttävä prioriteetti (jatkoa)

Keräämällä tulokset yhteen saadaan

$$\bar{T}_k = \bar{S}_k + \frac{\bar{R}_k}{1 - \rho_1 - \cdots - \rho_k} + \left(\sum_{i=1}^{k-1} \rho_i \right) \bar{T}_k$$

$$\bar{T}_1 = \frac{(1 - \rho_1)\bar{S}_1 + \bar{R}_1}{1 - \rho_1}$$
$$\bar{T}_k = \frac{(1 - \rho_1 - \cdots - \rho_k)\bar{S}_k + \bar{R}_k}{(1 - \rho_1 - \cdots - \rho_{k-1})(1 - \rho_1 - \cdots - \rho_k)}$$