Prioriteettijonot

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

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

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

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

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:

$$\overline{W}_1 = \overline{R} + \overline{S}_1 \bar{N}_q^{(1)}$$

missä jälkimmäinen termi edustaa jonossa edellä olevien luokan 1 asiakkaiden palvelemiseen keskimäärin kuluvaa aikaa

Littlen lauseen perusteella pätee

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

Prioriteettiluokalle 2 saadaan samaan tapaan

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

Littlen lauseesta seuraa jälleen

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

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

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

Luokan k asiakkaan kokonaisviipymä systeemissä keskimäärin on

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

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

$$\boxed{\overline{R} = \frac{1}{2}\sum_{k=1}^{K} \lambda_k \overline{S_k^2}}$$

Havaintoja ei-syrjäyttävästä prioriteetista

- Analyysiä ei ole helppo laajentaa monen palvelimen jonoihin
  - jäännösaikaa $\overline{R}$ 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

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

  - voidaan osoittaa, että jos $\overline{S}_1 < \overline{S}_2$, niin $\overline{T}$ 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.

Syrjäyttävä prioriteetti (preemptive resume priority)

Palvelussa olevan asiakkaan palvelu keskeytyy jonkun korkeamman prioriteetin asiakkaan saapuessa ja jatkuu siitä mihin se jäi, kun korkeamman prioriteteetin 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 $\overline{T}_k$. Tämä muodostuu kolmesta osasta:

1. Asiakkaan oma keskimääräinen palveluaika $\overline{S}_k$

2. Jonossa edellä olevien, luokkiin $1, \ldots, k$ kuuluvien asiakkaiden palveluun keskimäärin kuluva aika

$$\frac{\overline{R}_k}{1 - \rho_1 - \ldots - \rho_k} \quad , \text{ missä } \overline{R}_k = \tfrac{1}{2} \sum_{i=1}^{k} \lambda_i \overline{S_i^2},$$

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

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

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

Syrjäyttävä prioriteetti (jatkoa)

Keräämällä tulokset yhteen saadaan

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

$$\boxed{\begin{aligned} \overline{T}_1 &= \frac{(1-\rho_1)\overline{S}_1+\overline{R}_1}{1-\rho_1} \\ \overline{T}_k &= \frac{(1-\rho_1-\cdots-\rho_k)\overline{S}_k+\overline{R}_k}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)} \end{aligned}}$$