Veri Ağlarında Gecikme Modeli

Giriş

Veri ağlarındaki en önemli performans ölçütlerinden biri paketlerin ortalama gecikmesidir. Ağdaki iletişim gecikmeleri 4 farklı gecikmeden kaynaklanır:

1. **İşleme Gecikmesi:** Paketin doğru bir şekilde okunması ile paketin çıkışa verilmesi arasındaki süre
2. **Kuyruk Gecikmesi:** Paketin iletişim için bir kuyruğa eklenmesi ile iletime başlaması arasındaki süre
3. **İletim Gecikmesi:** Paketin ilk ve son bitlerinin iletilmesi arasındaki süre
4. **Yayılma Gecikmesi:** Paketin son bitinin gönderilmesi ile paketin karşı tarafta alınması arasındaki süre

Her bağlantı yolu için bir azami iletim kapasitesi mevcuttur. Eldeki spektrum farklı çoklama teknikleri kullanılarak birçok kullanıcıya aynı anda hizmet verecek hale getirilebilir (frekans bölmeli çoklama, zaman bölmeli çoklama, ...).

Kuyruk Modelleri

Bir paketin servis süresi L/C olarak ifade edilebilir. Burada L bit uzunluğu olarak paket boyutu ve C $bit/saniye$ cinsinden bağlantı kapasitesidir. Böyle bir bağlantı yolunda yer alan bir kuyruk sistemi için genellikle aşağıdaki niceliklerle ilgileniriz:

1. Sistemde ortalama kaç adet paket vardır?
2. Herhangi bir paket için ortalama gecikme süresi ne kadardır?

Sistemde n adet paket olma olasılığı P_n ile gösterilirse sistemdeki ortalama kişi sayısı

$$N = \sum_{i=0}^{\infty} nP_n$$

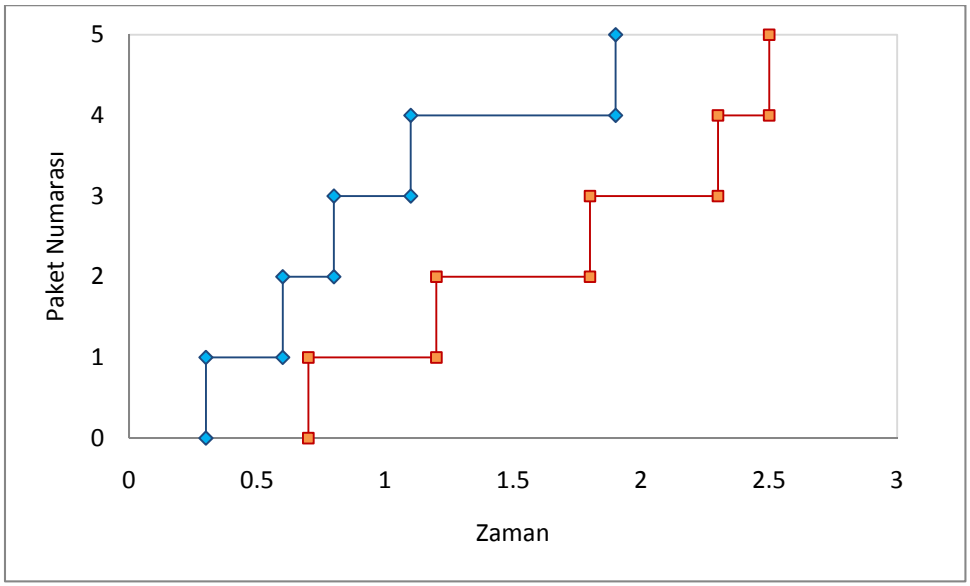
şeklinde hesaplanabilir. İlk başta boş durumda olan bir sistem için, $\alpha(t)$ ve $\beta(t)$ ifadeleri sırasıyla; sisteme $[0 - t]$ aralığında gelen paket sayısı ile sistemde $[0 - t]$ aralığında servis verilen paket sayısını göstersin. Bu durumda t anında sistemdeki paket sayısı $N(t)$,

$$N(t) = \alpha(t) - \beta(t)$$

şeklinde hesaplanabilir. Bu ifade yardımı ile Şekil-1'deki iki bağıntı arasında kalan alan iki farklı şekilde

$$\int_0^t N(\tau)d\tau = \sum_{i=1}^{\beta(t)} T_i + \sum_{i=\beta(t)+1}^{\alpha(t)} (t - t_i)$$

olarak hesaplanabilir. Burada T_i sistemde harcanan zamanları ve t_i geliş sürelerini gösteriyor.



Şekil 1: Sistemdeki giriş ve çıkışlar

Bu durumda; eşitliğin her iki tarafını t 'ye böldükten sonra, sağ taraftaki ifadeyi $\alpha(t)$ ile bölüp çarpabiliriz. Son adımda ise eşitliğin her iki tarafının limitini sonsuza götürürsek:

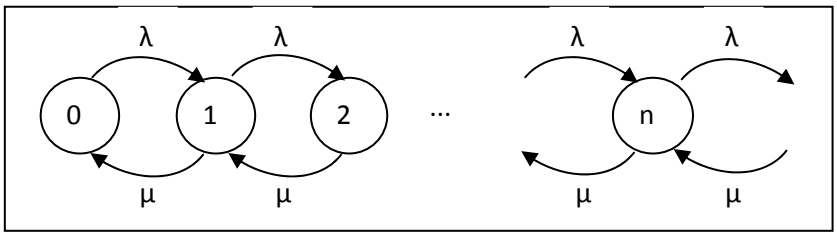
$$\lim_{t\to\infty}\frac{1}{t}\int_0^t N(\tau)d\tau = \lim_{t\to\infty}\frac{\alpha(t)}{t}\lim_{t\to\infty}\frac{\sum_{i=1}^{\beta(t)}T_i + \sum_{i=\beta(t)+1}^{\alpha(t)}(t-t_i)}{\alpha(t)}$$

sonucunu elde ederiz. Buradan da, λ ortalama paket gelme oranı olmak üzere, $N = \lambda T$ şeklinde ifade edilen **Little formülü** elde edilir. Sistemdeki toplam bekleme süresi T yerine sadece kuyruktaki bekleme süresi olan W kullanılırsa, Little formülüne benzer şekilde $N_k = \lambda W$ ifadesi elde edilir. $\bar{X}$ ortalama servis süresi olmak üzere, hattın **faydalanma oranı** $\rho = \lambda\bar{X}$ şeklinde ifade edilir.

M/M/1 Kuyruk Sistemi

M/M/1 (Giriş Dağılımı / Hizmet Dağılımı / Sunucu Sayısı) kuyruk sistemi tek sunuculu kuyruk istasyonunu ifade etmektedir. Burada, paketler λ parametresine bağlı olarak Poisson dağılımıyla sisteme gelirler. Dolayısıyla, paketlerin geliş süreleri arasındaki fark üstel olarak dağılmaktadır. Ayrıca, ortlaması μ^{-1} olan üstel dağılıma göre kullanıcılara hizmet verilir. M: belleksiz dağılım (Poisson - Üstel), G: genel dağılım, D: belirli dağılım olarak ifade edilmektedir.

Little teoremi ve sonuçları burada da geçerlidir ($N = \lambda T$). Ayrık zamanlı Markov zincirlerini kullanarak M/M/1 kuyruk sistemini Şekil-2'deki gibi ifade edebiliriz. Bu zincirde λ sıklıkla (paket/saniye) yeni bir paket gelirken, μ sıklıkla (paket/saniye) da kuyruktaki paketlerden ilkine hizmet verilmektedir. Gelen ve hizmet verilen paketlere göre kuyruktaki paket sayısı değişmektedir. Bu durumda sonsuz uzunluklu kuyruk için aşağıdaki denklemler bulunabilir.



Şekil 2: M/M/1 Kuyruk Sistemi - Markov Zinciri

Sistemin kararlılığını sağlamak için, her iki yöndeki ilerleme olasılıkları eşit olmalıdır.

$$\lambda P_{n-1} = \mu P_n$$

Faydalanma oranının $\rho = \lambda/\mu$ olduğunu kullanarak ($\mu = 1/\bar{X}$),

$$P_n = \rho P_{n-1} = \rho^n P_0$$

ifadesi elde edilir. Tüm düğümlerde bulunma olasılığının toplamının 1 olduğunu kullanalım.

$$\sum_{i=0}^{\infty} P_i = \sum_{i=0}^{\infty} \rho^i P_0 = P_0 \sum_{i=0}^{\infty} \rho^i = 1$$

Geometrik dizi açılımından yararlanarak:

$$\rho^0 + \rho^1 + \rho^2 + \cdots + \rho^n + \cdots = \frac{1}{P_0}$$

$$P_0 = 1 - \rho$$

$$P_n = \rho^n (1 - \rho)$$

ifadesini elde ederiz. Buradan da kararlı durumda kuyruk sistemindeki ortalama paket sayısını aşağıdaki şekilde hesaplayabiliriz.

$$N = \lim_{i \to \infty} E\{N(t)\} = \sum_{i=0}^{\infty} i P_i = \sum_{i=0}^{\infty} i \rho^i (1 - \rho)$$

$$= \rho(1 - \rho) \sum_{i=0}^{\infty} i \rho^{i-1} = \rho(1 - \rho) \frac{\partial}{\partial \rho} \sum_{i=0}^{\infty} \rho^i$$

$$= \rho(1 - \rho) \frac{\partial}{\partial \rho} \left(\frac{1}{1 - \rho} \right) = \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda}$$

Little teoreminden yararlanarak herhangi bir paketin sistemdeki ortalama bekleme süresi:

$$T = \frac{N}{\lambda} = \frac{1}{\mu - \lambda}$$

olarak hesaplanır. Bu süreden hizmet alma süresini çıkartarak kuyrukta harcanan ortalama süreyi aşağıdaki gibi hesaplayabiliriz.

$$W = T - T_{servis} = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}$$

Buradan da kuyruktaki ortalama paket sayısı aşağıdaki şekilde hesaplanabilir.

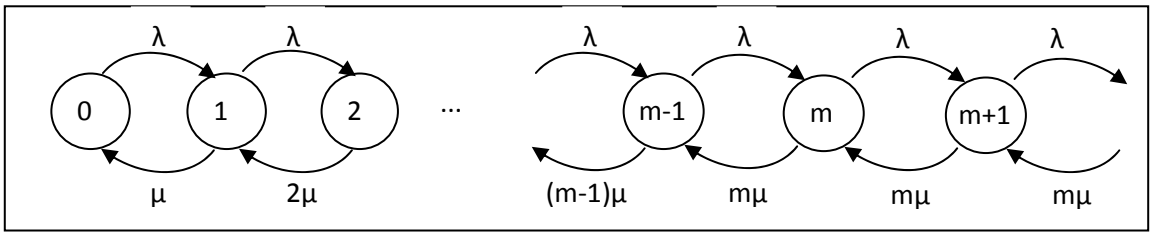
$$N_k = \lambda W = \frac{\rho^2}{1 - \rho}$$

Görüldüğü üzere; sistemin yalnızca faydalanılma oranı kullanılarak, kuyruktaki ortalama paket sayısı hesaplanabilmektedir.

1. $N \neq N_k + 1$ acaba neden?
   Çünkü, kuyruk sunucusu her zaman hizmet vermiyor. Aradaki farkın ρ olduğunu görebilirsiniz.
2. Sisteme paket geliş hızını ve sunucunun hizmet verme hızını K katına çıkarsak ($\lambda_i = K\lambda$ ve $\mu_i = K\mu$) sistemin davranışı nasıl değişir?
   $N_i = \frac{\lambda}{\mu-\lambda}$ olduğundan sistemdeki ortalama paket sayısı değişmedi. Ancak, $T_i = \frac{1}{K(\mu-\lambda)}$ olduğundan paketlerin ortalama bekleme süreleri azaldı.
3. Sistemin band genişliğini K eşit parçaya bölerek, her birindeki kuyruğa bu veriyi dağıtsak ($\lambda_i = \frac{\lambda}{K}$ ve $\mu_i = \frac{\mu}{K}$), sistemin davranışı nasıl değişir?
   $N_i = \frac{\lambda}{\mu-\lambda}$ olduğundan sistemdeki ortalama paket sayısı değişmedi. Ancak, $T_i = \frac{K}{\mu-\lambda}$ olduğundan paketlerin ortalama bekleme süreleri arttı.

M/M/m Kuyruk Sistemi

Bankadaki sistemlere benzer şekilde, kuyruğumuza toplam m adet sunucunun hizmet verdiğini düşünelim. Her bir kasiyer kuyrukta biri olduğu sürece hizmet sunsun.



Şekil 3: M/M/m Kuyruk Sistemi - Markov Zinciri

İlk bölümde elde ettiğimiz sonuçları benzer bir yoldan elde edelim.

$$n\mu P_n = \lambda P_{n-1}, \quad n \leq m$$

$$m\mu P_n = \lambda P_{n-1}, \quad n > m$$

Buradan da, $\rho = \frac{\lambda}{m\mu} \leq 1$ olmak üzere:

$$P_n = \begin{cases} P_0 \dfrac{(m\rho)^n}{n!}, & n \leq m \\ P_0 \dfrac{m^m \rho^n}{m!}, & n > m \end{cases}$$

olarak bulunur. Bir sonraki adımda, $\sum_{i=0}^{\infty} P_n = 1$ kullanılırsa:

$$P_0 = \left[\sum_{i=0}^{m-1} \frac{(mp)^i}{i!} + \frac{(mp)^m}{m!\,(1-\rho)} \right]^{-1}$$

ifadesini elde ederiz. Sisteme yeni gelen bir paketin, m sunucunun hiçbirisinin boşta olmaması sebebiyle kuyrukta bekleme olasılığı:

$$P_k = \sum_{i=m}^{\infty} P_i = \sum_{i=m}^{\infty} \frac{P_0 m^m \rho^i}{m!} = \frac{P_0 (m\rho)^m}{m!} \sum_{i=m}^{\infty} \rho^{i-m} = \frac{P_0 (m\rho)^m}{m!\,(1-\rho)}$$

olarak hesaplanır. Elde edilen bu sonuç, **Erlang C formülü** olarak da bilinmektedir. Gelinen son aşamada, kuyrukta bekleyen ortalama paket sayısı aşağıdaki gibi hesaplanabilir.

$$N_k = \sum_{i=0}^{\infty} iP_{m+i} = \sum_{i=0}^{\infty} iP_0 \frac{m^m \rho^{m+i}}{m!} = \frac{P_0 (m\rho)^m}{m!} \sum_{i=0}^{\infty} i\rho^i = \frac{\rho P_0 (m\rho)^m}{(1-\rho)^2 m!}$$

Buradan da,

$$\frac{N_k}{P_k} = \frac{\rho}{1-\rho}$$

kuyrukta bekleyen ortalama paket sayısının, sisteme yeni gelen bir paketin boşta sunucu bulamama olasılığına göre değişimini göstermektedir. Yukarıdaki denklem kuyrukta bekleyen paketler olduğu zaman; M/M/m sistemlerinin, servis hızı $m\mu$ olan M/M/1 sistemleri gibi davrandığını göstermektedir. Diğer önemli denklemleri de aşağıdaki gibi elde edebiliriz.

$$W = \frac{N_k}{\lambda} = \frac{\rho P_k}{\lambda(1-\rho)}$$

$$T = \frac{1}{\mu} + W = \frac{1}{\mu} + \frac{P_k}{m\mu - \lambda} \qquad (\rho = \lambda/m\mu)$$

$$N = \lambda T = \frac{\lambda}{\mu} + \frac{\lambda P_k}{m\mu - \lambda} = m\rho + \frac{\rho P_k}{(1-\rho)}$$

M/M/∞ Kuyruk Sistemi

$$n\mu P_n = \lambda P_{n-1}$$

$$P_n = P_0 \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}$$

$$P_0 = \left[1 + \sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}\right]^{-1} = e^{-\lambda/\mu}$$

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \frac{e^{-\lambda/\mu}}{n!}$$

Görüldüğü üzere, kararlı durumda paket sayısı λ/μ parametreli Poisson dağılımı gibi davranmaktadır. Sistemdeki ortalama paket sayısı $N = \lambda/\mu$ ve ortalama gecikme $T = 1/\mu$ şeklindedir.

M/M/m/m Kuyruk Sistemi: Kayıplı m-Sunucu Sistemi

Bu sistemin M/M/m sisteminden farkı, sisteme yeni bir paket geldiğinde eğer boş bir sunucu bulamıyorsa, paket düşmektedir. Bu yapı telefon sistemlerinde oldukça yoğun şekilde kullanılmaktadır.

$$n\mu P_n = \lambda P_{n-1}$$

$$P_n = P_0 \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}$$

$$P_0 = \left[\sum_{i=0}^{m} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}\right]^{-1}$$

Bunlara göre, yeni gelen bir paketin tüm sunucuları dolu bulup düşme olasılığı, diğer adıyla **Erlang B formülü** aşağıdaki gibi bulunur.

$$P_m = \frac{(\lambda/\mu)^m/m!}{P_0}$$

M/G/1 Kuyruk Sistemleri

Paketlerin gelişleri Poisson dağılımına uyarken, sunucuların verdikleri hizmetlerin üstel yerine genel dağılımda olduğu durumu inceleyelim. Bu durumda servis sürelerinin $(X_1, X_2, \dots)$ şeklinde olduğunu düşünelim. Bu durumda

$$\bar{X} = E\{X\} = \frac{1}{\mu} = \text{Ortalama servis süresi}$$

$$\overline{X^2} = E\{X^2\} = \text{Servis süresinin ikinci momenti}$$

olmak üzere, i. paketin kuyrukta ortalama bekleme süresi aşağıdaki gibi hesaplanır.

$$W_i = R_i + \sum_{j=i-N_i}^{i-1} X_j$$

W_i: i.paketin kuyruktaki bekleme süresi; R_i: hizmet alan bir paketin varlığından dolayı, i. paketin kuyrukta fazladan beklediği süre; X_i: i. paketin servis süresi; N_i: kuyrukta i.paketin önünde yer alan paketlerin sayısını ifade etmektedir. Rasgele değişkenlerin birbirlerinden bağımsız olduğu varsayımıyla, yukarıdaki formül için beklenen değerleri hesaplayalım.

$$E\{W_i\} = E\{R_i\} + E\left\{\sum_{j=i-N_i}^{i-1} E\{X_j/N_i\}\right\} = E\{R_i\} + \bar{X}E\{N_i\}$$

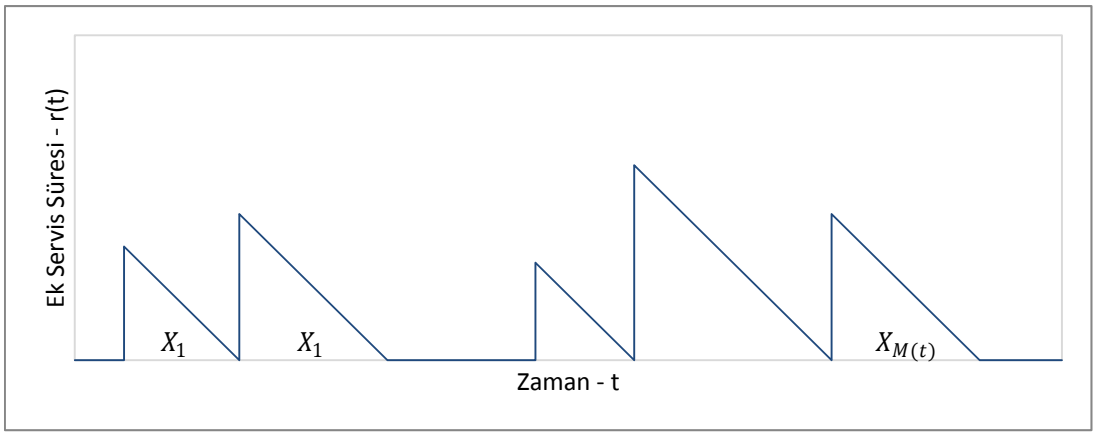
Bu ifadenin $i \to \infty$ için limitini alırsak, aşağıdaki formülü elde ederiz.

$$W = R + \frac{1}{\mu} N_k$$

Burada R ortalama artık bekleme süresini ifade eder ve $R = \lim_{i\to\infty} E\{R_i\}$ şeklinde hesaplanır. Little teoreminden ($N_k = \lambda W$) yararlanarak, $W = R + \rho W$ ($\rho = \lambda/\mu$) ifadesini elde ederiz. Buradan da W yalnız bırakılırsa aşağıdaki sonuç elde edilir.

$$W = \frac{R}{1-\rho}$$

R değerini Şekil-4'ten yararlanarak hesaplayalım.



Şekil 4: Ek Servis Süresi Hesabı

$$\frac{1}{t}\int_0^t r(\tau)d\tau = \frac{1}{t}\sum_{i=1}^{M(t)}\frac{1}{2}X_i^2 = \frac{1}{2}\frac{M(t)}{t}\frac{\sum_{i=1}^{M(t)}X_i^2}{M(t)}$$

Bu ifade sonsuza götürüldüğünde elde edilen $R = \frac{1}{2}\lambda\overline{X^2}$ denkleminden yararlanarak, literatürde Pollaczek-Khinchin formülü olarak bilinen aşağıdaki formül elde edilebilir.

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)}$$

Bu formül yardımıyla aşağıdakiler hesaplanabilir.

$$T = \bar{X} + \frac{\lambda\overline{X^2}}{2(1-\rho)}$$

$$N_k = \frac{\lambda^2\overline{X^2}}{2(1-\rho)}$$

$$N = \rho + \frac{\lambda^2\overline{X^2}}{2(1-\rho)}$$

Bu formüllerde üstel dağılım için $\overline{X^2} = 2/\mu^2$ ve belirli dağılım için $\overline{X^2} = 1/\mu^2$ ifadeleri geçerlidir.

Tatilli M/G/1 Kuyruk Sistemleri

Her yoğun çalışmanın ardından sunucunun bir süreliğine tatile çıktığını varsayalım. Bu durum gerçek hayatta, sistemler arası kontrol paketlerinin gönderilmesine karşı düşmektedir. Eğer tatil bittiğinde sistem hala boşsa, hemen yeni bir tatil başlamaktadır. Bu durumda sisteme yeni gelen bir paket ya kendinden önceki paketlere hizmet verilmesini ya da tatildeki sunucuyu ve önündeki paketlere hizmet verilmesini bekler. Bu durumda R'nin hesaplanmasında kullanılan denklem aşağıdaki şekli alır.

$$\frac{1}{t}\int_0^t r(\tau)d\tau = \frac{1}{t}\sum_{i=1}^{M(t)}\frac{1}{2}X_i^2 + \frac{1}{t}\sum_{i=1}^{L(t)}\frac{1}{2}V_i^2$$

Burada $M(t)$, t anına kadar tamamlanmış hizmet sayısını; $L(t)$, t anına kadar görülen tatil sayısını ifade eder. Bu durumda ortalama kuyrukta bekleme süresi aşağıdaki gibi hesaplanır.

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{\overline{V^2}}{2\bar{V}}$$

Yukarıdaki formülde $\bar{V}$ ve $\overline{V^2}$ sıraısyla tatil sürelerinin birinci ve ikinci momentini ifade etmektedir.

Öncelikli Kuyruklar

M/G/1 sistemleri için n farklı öncelik sınıfının tanımlı olduğu durumu inceleyelim. k. öncelik sınıfı için gelme oranı ve servis süresinin ilk iki momenti sırasıyla $\lambda_k, \overline{X_k} = 1/\mu_k, \overline{X_k^2}$ şeklinde verilebilir.

Sırasını Bekleyen Öncelik

Düşük öncelikli servisteki paketin tamamlanmasının beklendiği durumdur. Daha önceden P-K formülünü elde ettiğimiz yapıda birinci sınıftaki paketlerin kuyruktaki ortalama bekleme süreleri aşağıdaki gibi bulunur.

$$W_1 = R + \frac{1}{\mu_1} N_k^1$$

ve buradan da $N_k^1 = \lambda_1 W_1$ kullanılarak aşağıdaki son hali elde edilir.

$$W_1 = \frac{R}{1-\rho_1}.$$

İkinci öncelikli sınıftaki paketler için de benzer bir bekleme süresi hesabı yapılabilir. İlk hesaptan farklı olarak ikinci sınıftaki kuyrukta yer alan paketler ilk sınıftaki paketleri de bekleyeceklerdir. Bu durum aşağıdaki şekilde ifade edilebilir.

$$W_2 = R + \frac{1}{\mu_1} N_k^1 + \frac{1}{\mu_2} N_k^2 + \rho_1 W_2$$

Buradan da ikinci öncelikteki paketlerin ortalama bekleme süresi aşağıdaki gibi hesaplanır.

$$W_2 = \frac{R}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

Herhangi bir paket için ortalama gecikme:

$$T_k = \frac{1}{\mu_k} + W_k$$

olarak hesaplanmaktadır. Bu durumlar için $R = \frac{1}{2}\sum_{i=1}^{n} \lambda_i \overline{X_i^2}$ olarak hesaplanabilir.