

УДК 519.872

Модель оценки времени реакции центра обработки данных

Нестеров Ю. Г.^{1,*}

^{*}ugn@bmstu.ru

¹МГТУ им. Н.Э. Баумана, Москва, Россия

В статье описана аналитическая модель для оценки математического ожидания времени реакции ЦОД на диалоговые запросы удаленных пользователей, подключенных к ЦОД через выделенные каналы связи, работающие под управлением протоколов типа HDLC для конфигураций «точка-точка» и «точка-многоточка». Взаимодействие удаленных рабочих станций с ЦОД формализовано в виде замкнутой трехфазной СеМО, в которой первая, многоканальная фаза, соответствует фазе подготовки запроса к ЦОД на рабочей станции, вторая, одноканальная фаза, отражает обработку входящих и исходящих сообщений в канале связи в соответствии с логикой протокола HDLC, а третья, многоканальная фаза, отражает собственно обработку запросов в ЦОД. Для данной СеМО построена система нелинейных алгебраических уравнений для оценки средних значений времени реакции ЦОД и обоснована единственность решения этой системы. Предложена эффективная в вычислительном отношении процедура решения полученной системы уравнений.

Ключевые слова: система массового обслуживания (СМО), сеть массового обслуживания, узел, заявка, меченая заявка, входящее сообщение, исходящее сообщение, очередь, относительный приоритет, обслуживающий аппарат (ОА), выделенный канал связи, протокол, конфигурация, центр обработки данных, время обслуживания, время ожидания, время реакции

Введение

Центры обработки данных (ЦОД) являются в последнее десятилетие одним из магистральных направлений в развитии инфраструктуры предоставления ИТ-услуг, в частности в форме облаков. Доступ к услугам ЦОД для основной массы пользователей осуществляется дистанционно через корпоративные сети (частные ЦОДы) и/или через сети интернет-провайдеров (как частные, так и публичные ЦОДы), т.е. в конечном итоге через выделенные каналы связи. При этом в корпоративных сетях и в сетях доступак опорным сетям провайдеров, построенных в 1990-е...2000-е годы и работающих до сих пор, использовался широкий спектр протоколов канального уровня, в том числе, FrameRelay [1], ATM [2], HDLC [3] и HDLC поверх ATM [5]. По интерфейсу обмена данными (DXI) ATM позволяет осуществлять доступ к сети ATM существующего оборудования (например, маршрутизаторов с HDLC) без его модернизации. Формат

данных соответствует протоколу HDLC. Мультиплексор доступа преобразует кадры HDLC в ячейки ATM, при необходимости преобразует трафик для выполнения соглашения по трафику. Сопряжение с сетью ATM производится по интерфейсу UNI [2, 4].

С конца 2000-х, в первую очередь в сетях ведущих операторов связи, технологии ATM начинают вытесняться новой технологией IP-VPN [6]. ATM-коммутаторы вытесняются маршрутизаторами IP/MPLS[7]. С 2005 года работы по развитию протокола ATM остановлены.

В свою очередь технологии, реализующие протокол HDLC, который долгое время, вплоть до конца 2000-х годов, был главным для выделенных линий связи, стали замещаться технологиями, реализующими протокол PPP, входящий в семейство HDLC [8]. Это связано в первую очередь с широким распространением оптических линий связи, обладающих высокой надежностью, что делает функциональность HDLC, предназначенную для восстановления искаженных и утерянных кадров, излишней.

Тем не менее, несмотря на прогноз 2011 г. о вытеснении ATM к 2015 году [6], в корпоративных сетях и в сетях доступа некоторых операторов связи, а также в системах специального назначения, в силу стремления корпораций к сохранению инвестиций в коммуникации, до сих пор еще широко применяются технологии, реализующие протоколы HDLC и HDLC поверх ATM.

На этапах проектирования и, в дальнейшем, в процессе эксплуатации ЦОДов, при решения задач выбора конфигурации каналов связи и задач управления нагрузкой, используются модели оценки производительности (пропускной способности и времени реакции на диалоговые запросы) ЦОД совместно с сетью доступа удаленных пользователей к услугам ЦОД. Для оценки производительности и времени реакции ЦОД на диалоговые запросы пользователей чаще всего применяются разомкнутые и замкнутые системы массового обслуживания (СМО) и сети массового обслуживания (СеМО) [1,9,12]. При этом наличие аналитических моделей СМО и СеМО в силу их высокой вычислительной эффективности и легкости управления, является более предпочтительным, по сравнению с имитационными моделями, более тяжелыми и трудноуправляемыми при проведении массовых модельных экспериментов.

В работе [9] аналитические решения для средних значений времен пребывания в узлах замкнутых СМО типа «модели ремонтника» (в обозначениях Кендалла [12] - $Mr|G|Ir|m||N$), используются в качестве базовых моделей для расчета средних значений времен пребывания заявок в узлах широкого класса замкнутых сетей массового обслуживания с приоритетами и консервативными дисциплинами обслуживания. Из изученных на сегодня замкнутых СМО наиболее близкой к логике работы протокола HDLC является замкнутая СМО типа $Mr|G|Ir|1||Nr$ с дисциплиной «относительные приоритеты», для которой в работе [10] получено аналитическое решение для средних времен ожидания и пребывания. Тем не менее, отражая наличие приоритетности в обслуживании, данная СМО не учитывает особенности взаимодействия ЦОДа и удаленных рабочих станций в протоколе HDLC, а именно логику опроса вторичных узлов первичным [3].

Настоящая работа посвящена получению эффективного в вычислительном отношении аналитического решения для первого момента ФРВ времени реакции ЦОД на диалоговые запросы удаленных пользователей, подключенных к ЦОД через выделенные каналы связи, работающие под управлением протоколов типа HDLC (HDLC поверх ATM) для конфигураций выделенного канала «точка-многоточка» (point-multipoint) и «точка-точка» (point-point) в режиме нормального ответа (Normal Response Mode, NRM) [3]. При этом в логическом соединении ЦОД выполняет роль ведущего (первичного) узла, а удаленные станции – роль ведомых (вторичных) узлов. Аналогичные модели для протоколов типа IP/MPLS будут рассмотрены автором в следующих статьях.

Постановка задачи

В качестве модели для оценки первого момента функции распределения вероятностей(ФРВ) времени ответа ЦОД на запросы удаленных рабочих станций, подключенных к ЦОД по выделенному каналу связи, работающему под управлением протокола HDLC в конфигурации «точка-многоточка» и в режиме нормального ответа NRM, будем рассматривать замкнутую трехфазную СМО. Конфигурация «точка-точка», как будет показано ниже, является частным случаем конфигурации «точка-многоточка». В этой СМО первая многоканальная фаза соответствует фазе обдумывания (подготовки) запроса к ЦОД на рабочей станции. В этой фазе количество обслуживающих аппаратов (ОА), т.е. каналов обслуживания, равно количеству рабочих станций (терминалов). Вторая одноканальная фаза этой СМО отражает обработку входящих (в ЦОД) и исходящих (из ЦОД)сообщений в канале связи в соответствии с логикой протокола HDLC. Третья многоканальная фаза этой СМО отражает собственно обработку запросов в ЦОД. На этой фазе каждому активному терминалу соответствует один процесс в ЦОД, т.е. количество ОА равно количеству рабочих станций.

Структура данной СМО приведена на рис.1, где приняты следующие обозначения:

$T_1, \dots, T_R$ - ОА, соответствующие терминалам(рабочим станциям), а R – количество терминалов, подключенных к каналу связи (КС);

$C_1, \dots, C_R$ - ОА, представляющие в совокупности параллельные процессы обработки запросов в ЦОД;

A_{Tr} - время обдумывания пользователя за терминалом r (время подготовки входного сообщения);

W_{Ir} - время ожидания момента начала передачи входного сообщения от терминала r через КС;

B_{Ir} - время передачи входного сообщения от терминала r по КС;

A_{Cr} - время обработкисообщения от терминала r в ЦОД;

W_{or} - время ожиданияначала передачи по КС выходного сообщения (ответа на входное сообщение)на терминал r ;

B_{or} - время передачи по КС выходного сообщения на терминал r .

В каждой из очередей $L_1, \dots, L_R$ может находиться либо 0, либо 1 заявка, а в очереди L_0 не более $(R-1)$ заявок. Всего в СМО циркулирует R заявок, по одной для каждого терминала. Алгоритм обслуживания заявок из очередей $L_1, \dots, L_R, L_0$ в КС следующий. Очереди $L_1, \dots, L_R$ обслуживаются циклически, заявки из очереди L_0 имеют более высокий приоритет, чем заявки из очередей $L_1, \dots, L_R$ (приоритет передачи). Это соответствует алгоритму управления КС по протоколу HDLC в конфигурации «точка-многоточка» в режиме нормального ответа NRM [3].

Первые два момента ФРВ времен $A_{Tr}, A_{Cr}, B_{Ir}, B_{Or}, r = \overline{1, R}$ заданы.

Первые моменты соответствующей случайной величины будем обозначать строчными буквами, соответствующими ее имени.

Время реакции Tr ЦОД на запросы r -го терминала есть сумма вида:

$$Tr = W_{Ir} + B_{Ir} + A_{Cr} + W_{Or} + B_{Or}.$$

Переходя к средним значениям, получим:

$$\tau_r = w_{Ir} + b_{Ir} + a_{Cr} + w_{Or} + b_{Or}$$

Т.е. для определения τ_r - среднего значения времени реакции ЦОД на запросы r -го терминала, требуется определить w_{Ir} и w_{Or} - средние значения времен W_{Ir} и W_{Or} , соответственно, $r = \overline{1, R}$.

Решение задачи

Для определения w_{Ir} и w_{Or} - средних значений W_{Ir} и W_{Or} воспользуемся т.н. методом «меченой заявки», предложенным Кобхэмом [11].

Заметим заявку от терминала r , поступающую из C_r в очередь L_0 . Для случайного времени W_{Or} ожидания этой заявкой начала передачи по КС можем записать:

$$W_{Or} = D_{Or} + \sum_{k=1}^{r-1} q_{Ok} \cdot B_{Ok} + \sum_{k=1}^{r-1} \sum_{v^k=0}^{l_{Ok}} B_{Ok}^{v^k}, \quad r = \overline{1, R} \quad (1)$$

Здесь:

D_{Or} - остаточное время обслуживания заявки, находившейся в КС в момент прихода меченой заявки;

$$q_{Ok} = \begin{cases} 0, & \text{если в очереди } L_0 \text{ не было заявки на вывод от терминала } k \\ \text{в момент прихода меченой заявки от терминала } r; \\ 1, & \text{в противном случае.} \end{cases}$$

$l_{Ok} = 0$ или 1 - случайное число заявок терминала k , которые придут в очередь за время W_{Or} , причем $B_{Ok}^0 \neq 0$.

Переходя к средним из (1), получим:

$$w_{Or} = \overline{D}_{Or} + \sum_{k=1}^{r-1} \overline{q}_{Ok} \cdot b_{Ok} + \sum_{k=1}^{r-1} \overline{l}_{Ok} b_{Ok}, \quad r = \overline{1, R} \quad (2)$$

Здесь:

$$\overline{D}_{Or} = \sum_{\substack{k=1 \\ k \neq r}}^R \left[\lambda_k^* b_{Ok} \frac{b_{Ok}^{(2)}}{2b_{Ok}} + \lambda_k^* b_{Ik} \frac{b_{Ik}^{(2)}}{2b_{Ik}} \right] = \sum_{\substack{k=1 \\ k \neq r}}^R \left[\frac{\lambda_k^* (b_{Ok}^{(2)} + b_{Ik}^{(2)})}{2} \right] \quad (3)$$

$$\overline{q}_{Ok} = \lambda_k^* w_{Ok} = \frac{w_{Ok}}{w_{Ok} + b_{Ok} + a_{Tk} + w_{Ik} + b_{Ik} + a_{Ck}} \quad (4)$$

$$\overline{l}_{Ok} = \lambda_k^* w_{Or} = \frac{w_{Or}}{w_{Ok} + b_{Ok} + a_{Tk} + w_{Ik} + b_{Ik} + a_{Ck}} \quad (5)$$

В выражениях (3), (4) и (5) $\lambda_k^* = (w_{Ok} + b_{Ok} + a_{Tk} + w_{Ik} + b_{Ik} + a_{Ck})^{-1}$ есть интенсивность поступления заявки от терминала k в очередь L_0 , а $b_{Ok}^{(2)}$ и $b_{Ik}^{(2)}$ суть вторые начальные моменты ФРВ времен B_{Ok} и B_{Ik} соответственно. Для остаточного времени обслуживания D_{Or} использованы результаты работы [13].

Подставляя (3), (4) и (5) в (2), получим R уравнений с 2R неизвестными – (w_{Ir}, w_{Or}), $r = \overline{1, R}$.

$$w_{Or} = \sum_{\substack{k=1 \\ k \neq r}}^R \left[\frac{b_{Ok}^{(2)} + b_{Ik}^{(2)}}{2(w_{Ok} + b_{Ok} + a_{Tk} + w_{Ik} + b_{Ik} + a_{Ck})} \right] + \sum_{k=1}^{r-1} \frac{b_{Ok} (w_{Ok} + w_{Or})}{w_{Ok} + b_{Ok} + a_{Tk} + w_{Ik} + b_{Ik} + a_{Ck}}; \quad r = \overline{1, R} \quad (6)$$

Еще R уравнений получим, рассматривая поступление меченой заявки от терминала r из T_r в очередь L_r (очередь входных сообщений).

Для случайного времени W_{Ir} ожидания меченой заявкой момента начала ее передачи по КС, учитывая тот факт, что исходящие сообщения имеют более высокий приоритет перед входящими, запишем:

$$W_{Ir} = D_{Ir} + \sum_{k=1}^{r-1} q_{Ik} \cdot B_{Ok} + \sum_{k=1}^{r-1} \sum_{\nu^k=0}^{l_{Ik}} B_{Ok}^{\nu^k} + D_{\Pi O, r} \quad (7)$$

Здесь:

D_{Ir} - остаточное время обслуживания заявки, находившейся в КС в момент прихода меченой заявки, причем $D_{Ir} = D_{Or}$.

$q_{Ir} = q_{Or}$, см. (4);

$l_{Ik} = 0$ или 1 - случайное число заявок терминала k, которые придут в очередь L_0 за время W_{Ir} , причем $B_{Ok}^0 \geq 0$.

$D_{\Pi O, r}$ - случайное время задержки начала ввода меченой заявки, связанное с циклическим опросом терминалов $T_1, \dots, T_R$ ведущим узлом.

Для $D_{\Pi O, r}$ можно записать:

$$D_{\Pi O, r} = \sum_{k=1}^R \Pr \{k | r\} \cdot D_{kr} \quad (8)$$

Здесь:

1. $\Pr\{k|r\}$ - вероятность того, что предыдущий обслуженный из очередей $L_1, \dots, L_R$ запрос был от терминала k при условии, что поступает меченая заявка от терминала r , $k = \overline{1, R}$, $r = \overline{1, R}$, $k \neq r$, $\Pr\{r|r\} = 0$.

Очевидно, что

$$\Pr\{k|r\} = \lambda_k^* / \Lambda_r^* \quad (9)$$

где λ_k^* см. (5), а $\Lambda_r^* = \sum_{\substack{k=1 \\ k \neq r}}^R \lambda_k^*$, $r = \overline{1, R}$.

2. D_{kr} - случайная задержка, связанная с обслуживанием заявок на ввод из очередей с номерами, лежащими между k -м и r -м по ходу опроса.

Для D_{kr} можно записать:

$$D_{kr} = \begin{cases} \sum_{j=k+1}^{r-1} (m_j + l_j') B_{lj}, & \text{если } k < r, \\ \sum_{j=k+1}^R (m_j + l_j'') B_{lj} + \sum_{j=1}^{r-1} (m_j + l_j''') B_{lj}, & \text{если } k > r \end{cases} \quad (10)$$

Здесь:

$$m_j = \begin{cases} 1, & \text{если в момент прихода меченой заявки в очереди } L_j \\ & \text{была заявка на ввод через КС;} \\ 0, & \text{противном случае.} \end{cases}$$

$$l_j' = \begin{cases} 1, & \text{если в течение времени } W_{lr} \text{ в очередь } L_j \text{ поступит заявка от терминала } j, \\ & \text{причем раньше, чем начнет обслуживаться любая из заявок терминалов } l \\ & (r > l > j) \text{ из очередей } L_l \text{ на ввод;} \\ 0, & \text{в противном случае.} \end{cases}$$

$$l_j'' = \begin{cases} 1, & \text{если в течение времени } W_{lr} \text{ в очередь } L_j \text{ поступит заявка от терминала } j, \\ & \text{причем раньше, чем начнет обслуживаться любая из заявок от терминалов } l \\ & (R \geq l > j) \text{ или } (1 \leq l < r); \\ 0 & \text{в противном случае.} \end{cases}$$

Переходя к средним из (7) с учетом (8) и (9), получим:

$$w_{lr} = \bar{D}_{lr} + \sum_{k=1}^{r-1} \bar{q}_{lk} b_{Ok} + \sum_{k=1}^{r-1} \bar{l}_{lk} b_{Ok} + \sum_{k=1}^R \frac{\lambda_k^*}{\Lambda_r^*} \bar{D}_{kr} \quad (11)$$

Здесь:

$$\bar{D}_{lr} = \bar{D}_{Or}, \text{ см. (3);}$$

$$\bar{q}_{lk} = \bar{q}_{Ok}, \text{ см. (4).}$$

$$\bar{l}_{lk} = \lambda_k^* w_{lr} = \frac{w_{lr}}{w_{ok} + b_{ok} + a_{Tk} + w_{lk} + b_{lk} + a_{Ck}} \quad (12)$$

Для $\bar{D}_{kr}$ имеем из (10):

$$\bar{D}_{kr} = \begin{cases} \sum_{j=k+1}^{r-1} (\bar{m}_j + \bar{l}_j) b_{lj}, & \text{если } k < r \\ \sum_{j=k+1}^R (\bar{m}_j + \bar{l}_j) b_{lj} + \sum_{j=1}^{r-1} (\bar{m}_j + \bar{l}_j) b_{lj}, & \text{если } k > r \end{cases} \quad (13)$$

Здесь:

$$\bar{m}_j = \lambda_j^* w_{lj} = \frac{w_{lj}}{w_{oj} + b_{oj} + a_{Tj} + w_{lj} + b_{lj} + a_{Cj}} \quad (14)$$

$$\bar{l}_j' = \left(\frac{\lambda_j^*}{\Lambda_{j,(r-1)}^{**}} \right) \cdot \lambda_j^* w_{lr}, \quad \Lambda_{j,(r-1)}^{**} = \sum_{l=j}^{r-1} \lambda_l^* \quad (15)$$

$$\bar{l}_j'' = \left(\frac{\lambda_j^*}{\Lambda_{j,(r-1)}^{**}} \right) \cdot \lambda_j^* w_{lr}, \quad \Lambda_{j,(r-1)}^{**} = \sum_{l=j}^R \lambda_l^* + \sum_{l=1}^{r-1} \lambda_l^* \quad (16)$$

$$\bar{l}_j''' = \left(\frac{\lambda_j^*}{\Lambda_{j,(r-1)}^{**}} \right) \cdot \lambda_j^* w_{lr}, \quad \text{т.е. } \bar{l}_j''' = \bar{l}_j' \quad (17)$$

Соотношения (6) и (11)÷(17) совместно с (3), (4) и (9) определяют систему из $2R$ нелинейных алгебраических уравнений с $2R$ неизвестными (w_{lr}, w_{or}) , $r = \overline{1, R}$. В векторной форме эта система запишется следующим образом:

$$w = \varphi(w) \quad (18)$$

где $w = (w_{l1}, \dots, w_{lR}, w_{o1}, \dots, w_{oR})$, а оператор φ задается соотношениями (6) и (9)÷(17).

Поскольку оператор φ определяет трехузловую замкнутую СеМО с консервативной дисциплиной обслуживания, то, как показано в работе [9], существует единственное решение системы уравнений типа (18), и для отыскания решения пригоден модифицированный метод Ньютона-Канторовича [9]. Начальное приближение $w^0 = (0, 0, \dots, 0)$ - нулевой вектор.

Теперь рассмотрим вариант для конфигурации КС «точка-точка». При этом допустимы режим нормального ответа NRM и асинхронный сбалансированный режим (Asynchronous Balanced Mode, ABM). Этот вариант управления каналом связи применяется для соединения двух маршрутизаторов на концах выделенного канала. На одном конце канала – ЦОД, на другом – маршрутизатор локальной сети, к которой подключены терминалы (рабочие станции). Предполагается, что скорость передачи сообщений по локальной сети существенно выше, чем скорость передачи по выделенному КС, поэтому временем задержки в локальной сети можно пренебречь.

Нетрудно видеть, что аналитическая модель для этого варианта является частным случаем полученной выше модели для конфигурации «точка-многоточка». Нужно только в выражении (7) вместо $D_{цО,г}$ использовать выражение (19):

$$D_r = \sum_{\substack{k=1 \\ k \neq r}}^R m_k B_{rk} \quad (19)$$

Здесь:

D_r - случайное время ожидания меченой заявкой начала обслуживания, связанное с обслуживанием заявок на ввод от других терминалов, которые (заявки) уже находились во входных очередях в момент прихода меченой заявки;

m_k - случайное число (0 или 1) заявок от терминала k в очереди Lk в момент прихода меченой заявки.

Переходя к средним из (19) получим:

$$\bar{D}_r = \sum_{\substack{k=1 \\ k \neq r}}^R \bar{m}_k b_{rk} \quad (20)$$

Здесь $\bar{m}_k$ определяется соотношением (14).

Заключение

1. Получена формализация процесса взаимодействия с ЦОД удаленных рабочих станций, подключенных по выделенному каналу связи, работающему под управлением протокола HDLC, в виде замкнутой трехфазной СеМО, в которой первая фаза соответствует этапу обдумывания (подготовки запроса) на рабочей станции, вторая фаза соответствует этапу передачи сообщений по КС и третья фаза соответствует этапу собственно обработки сообщения в ЦОД.
2. Разработана аналитическая модель в виде системы нелинейных алгебраических уравнений для оценки средних значений времени реакции ЦОД для двух конфигураций КС под управлением протокола HDLC - «точка - многоточка» и «точка - точка». Обосновано существование единственного решения полученной системы уравнений.
3. Предложена эффективная в вычислительном отношении процедура решения полученной системы уравнений.
4. Разработанный подход применим также и для оценки дисперсий времени реакции ЦОД.

Список литературы

1. Вишневский В.М. Теоретические основы проектирования компьютерных сетей. М.: Техносфера, 2003. 512 с.
2. Назаров А.Н., Разживин И.А., Симонов М.В. АТМ: Технические решения создания сетей: Справочное издание. М.: Горячая Линия-Телеком, 2001. 376 с.
3. ISO/IEC 13239:2002. Information technology -- Telecommunications and information exchange between systems -- High-level data link control (HDLC) procedures // ISO: organi-

zation website. Режим доступа:

http://www.iso.org/iso/catalogue_detail.htm?csnumber=37010 (дата обращения 28.07.2015).

4. Руководство по технологиям объединённых сетей. Настольный справочник специалиста по сетевым технологиям: пер. с англ. 4-е изд. М.: Вильямс, 2005. 1040 с.
5. Forrest L. Pierson, Jr. Method and system for processing an HDLC message: patent US 6195346 B1. 2001.
6. From ATM to IP/Ethernet: Three Strategies for Cost-Effective Network Convergence. White Paper. Tellabs, 2011.
7. IP/MPLS, ATM and Frame Relay Forum Specifications. Режим доступа: <https://www.broadband-forum.org/technical/ipmplstechspec.php> (дата обращения 28.07.2015).
8. The Point-to-Point Protocol (PPP). RFC 1661, STD 51 / ed. by W. Simpson. Режим доступа: <https://tools.ietf.org/html/rfc1661> (дата обращения 28.07.2015).
9. Нестеров Ю.Г. Декомпозиционный метод анализа замкнутых сетей массового обслуживания // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2014. № 2. С. 263-276. DOI: [10.7463/0214.0700018](https://doi.org/10.7463/0214.0700018)
10. Нестеров Ю.Г. Анализ характеристик замкнутой системы массового обслуживания с относительными приоритетами // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2014. № 3. С. 242-254. DOI: [10.7463/0314.0702664](https://doi.org/10.7463/0314.0702664)
11. Гнеденко Б.В., Коваленко И.Н. Введение в теорию массового обслуживания. М.: Изд-во ЛКИ, 2007. 400 с.
12. Клейнрок Л. Теория массового обслуживания: пер. с англ. М.: Машиностроение, 1979. 432 с.
13. Кёниг Д., Штойян Д. Методы теории массового обслуживания: пер. с нем. М.: Радио и связь, 1981. 128 с.

Model for Estimation of the Data Center Response Time

Yu.G. Nesterov^{1,*}

^{*}ugn@bmstu.ru

¹Bauman Moscow State Technical University, Moscow, Russia

Keywords: queuing system, queuing network, node, transaction, labeled transaction, input message, output message, queue, non-preemptive priority, facility, leased line, protocol, configuration, data center, service time, waiting time, response time

The article presents an analytical model to estimate the mathematical expectation of the data center (DC) response time to the interactive requests from remote users connected to the DC via leased lines (LL), which run under control of HDLC-type protocols for configurations "point-point" and "point - multipoint » in Normal Response Mode, NRM. In the logic connection DC acts as the host (primary node) and remote stations play the role of the slave units (secondary nodes).

The objective of research is to obtain analytical expressions to define the following:

- Mathematical expectation of waiting time required for input and output messages of each class to start transmitting via leased line;
- Mathematical expectation of total DC response time to requests of each class.

The process of interaction between remote workstations and DC is formalized as a closed three-node queuing network (QN), wherein the first multi-channel node corresponds to the phase of request deliberation (preparation) to DC on the workstation. In this phase, the number of facilities, i.e. service channels, is equal to the number of workstations (terminals), R . The second single-channel node of this QN reflects processing of input and output messages to the LL in accordance with the logic of HDLC protocol. The third multi-channel node of this NQ reflects, as a matter of fact, processing query to DC. In this node one process in DC corresponds to each active terminal, i.e., the number of facilities is equal to the number of workstations R .

It is shown that the problem of determining the DC average response time to requests of remote users is reduced to the problem of determining the average waiting time for the input and output messages to start transmitting messages by LL - that sort of unknown average waiting time turns out to be $2R$. The first two points of the probability distribution functions of reflection time, LL transmission time, and time of message processing, as a matter of fact, in DC are assumed to be specified.

To create a system of algebraic equations in regard to the unknown average waiting time for the input and output messages to start LL sending a so-called method of "labeled transaction"

is used. The system of equations is constructed, initially, to provide the "point-multipoint configuration." Then it is shown how to obtain a system for configuration "point-point" as a special case.

The first phase considers an arrival of the "labelled transaction" submitting output message in QN and builds R equations in regard to the unknown average waiting time for the output message to start its sending by LL from DC to the workstation.

The second phase considers an arrival of the "labelled transaction" submitting input message in QN and builds R equations in regard to the unknown average waiting time for the input message to start its sending by LL from the workstation to DC.

The result is a system of the $2R$ algebraic equations in regard to $2R$ unknown average waiting time for a start of message transmission by LL. The existence of a unique solution of the resulting system of equations is proved, and a computationally efficient procedure for determining solution as a modified Newton-Kantorovich method is proposed.

The paper shows how to obtain a system of equations for the "point-point" configuration, as a special case, from the system of equations for the "point-multipoint" configuration. It presents the formulas for calculating the DC average response time to remote users' requests for each workstation.

References

1. Vishnevskiy V.M. *Teoreticheskie osnovy proektirovaniya komp'yuternykh setey* [Theoretical bases of designing of computer networks]. Moscow, Tekhnosfera Publ., 2003. 512 p. (in Russian).
2. Nazarov A.N., Razzhivin I.A., Simonov M.V. *ATM: Tekhnicheskie resheniya sozdaniya setei* [ATM: Technical solutions of networking]. Moscow, Goryachaya Liniya-Telekom Publ., 2001. 376 p. (in Russian).
3. ISO/IEC 13239:2002. Information technology -- Telecommunications and information exchange between systems -- High-level data link control (HDLC) procedures // ISO: organization website. Available at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=37010, accessed 28.07.2015.
4. Internetworking Technologies Handbook. Cisco Press, 2003. 1128 p. (Russ. ed.: *Rukovodstvo po tekhnologiyam ob"edinennykh setei. Nastol'nyi spravochnik spetsialista po setevym tekhnologiyam*. Moscow, Williams Publishing House, 2005. 1040 p.).
5. Forrest L. Pierson, Jr. *Method and system for processing an HDLC message*. Patent US 6195346 B1. 2001.
6. From ATM to IP/Ethernet: Three Strategies for Cost-Effective Network Convergence. White Paper. Tellabs, 2011.
7. IP/MPLS, ATM and Frame Relay Forum Specifications. Available at: <https://www.broadband-forum.org/technical/ipmplstechspec.php>, accessed 28.07.2015.

8. Simpson W., ed. *The Point-to-Point Protocol (PPP)*. RFC 1661, STD 51. Available at: <https://tools.ietf.org/html/rfc1661> , accessed 28.07.2015.
9. Nesterov Yu.G. Decomposition method for analysis of closed queuing networks. *Nauka i obrazovanie MGTU im. N.E. Bauman = Science and Education of the Bauman MSTU*, 2014, no. 2, pp. 263-276. DOI: [10.7463/0214.0700018](https://doi.org/10.7463/0214.0700018) (in Russian).
10. Nesterov Yu.G. Feature analysis for closed queuing system with non-preemptive priorities. *Nauka i obrazovanie MGTU im. N.E. Bauman = Science and Education of the Bauman MSTU*, 2014, no. 3, pp. 242-254. DOI: [10.7463/0314.0702664](https://doi.org/10.7463/0314.0702664) (in Russian).
11. Gnedenko B.V., Kovalenko I.N. *Vvedenie v teoriyu massovogo obsluzhivaniya* [Introduction to queuing theory]. Moscow, LKI Publ., 2007. 400 p. (in Russian).
12. Kleinrock L. *Queueing Systems. Vol. 1. Theory*. New York, John Wiley and Sons, 1975. (Russ. ed.: Kleinrock L. *Teoriia massovogo obsluzhivaniia*. Moscow, Mashinostroenie Publ., 1979. 432 p.).
13. Konig D., Stoyan D. *Metody teorii massovogo obsluzhivaniya* [Methods of queuing theory]. Transl. from German. Moscow, Radio i svyaz' Publ., 1981. 128 p. (in Russian).