Математическая Теория Игр и её Приложения, т.11, в.3, с. 95–120

УДК 519.217
ББК 22.18

СИСТЕМА ОБСЛУЖИВАНИЯ-ЗАПАСАНИЯ С ТРЕБОВАНИЯМИ СЛУЧАЙНОГО ОБЪЕМА И ПОЛОЖИТЕЛЬНЫМИ ВРЕМЕНАМИ ОБСЛУЖИВАНИЯ

Шринивас Чакраварти
Университет Кеттеринга
Флинт, США
e-mail: schakrav@kettering.edu

Рассмотрены модели обслуживания-запасания, с точечным процессом поступления клиентов, каждый из которых расходует случайное число единиц запаса, не превышающее конечную константу N. Обслуживание клиента занимает положительное случайное время. Возобновление запаса осуществляется в соответствии с политикой (s, S)-типа, при этом времена возобновления запаса являются случайными. Рассмотрено две модели. В первой модели клиент, обнаруживающий пустой запас в момент прихода, навсегда покидает систему. Во второй модели потери клиентов возможны по двум причинам. Во-первых, клиент, обнаруживающий пустой запас и простаивающий сервер в момент прихода, покидает систему. Во-вторых, систему покидают все клиенты, находящиеся в системе в такой момент окончания обслуживания, при котором запас опустошается. В обоих моделях предполагается, что заявки могут быть удовлетворены частично, в зависимости от размера требования и оставшегося запаса. Иными словами, в момент начала обслуживания клиента, запас уменьшается на размер требования,

но не более, чем на оставшийся запас. В предположении, что все случайные величины имеют экспоненциальные распределения, проведен анализ моделей в стационарном режиме с помощью классического матрично-аналитического метода. Для иллюстрации приведены примеры сравнения двух моделей.

Ключевые слова: системы обслуживания-запасания, алгоритмическая вероятность, требования случайного объема, время поставки, матрично-аналитический метод.

Поступила в редакцию: 25.05.18 *После доработки*: 28.08.19 *Принята к публикации*: 30.09.19

1. Введение

Системы обслуживания, в которых в процессе обслуживания клиента расходуется какой-либо ресурс (помимо временного ресурса на обслуживающем устройстве) принято называть системами обслуживания-запасания. Запас требуется для содержания расходуемого ресурса, который возобновляется в соответствии с какой-либо политикой, такой как политика $(s, S)-$типа. Такие системы, начиная с работы [2], являются актуальной темой для исследования (см., напр., [1–7]). Системы обслуживания-запасания, в которых запросы имеют случайный объем, а время обслуживания запросов пренебрежимо мало, хорошо изучены (см., напр., [10]). В работе [3] исследована система обслуживания-запасания, в которой клиенты (каждому из которых требуется одна единица ресурса) обслуживаются группами случайного объема. Однако, на сколько нам известно, в научной литературе отсутствуют исследования систем с требованиями случайного объема и положительными временами обслуживания. К моменту окончательного оформления статьи авторы ознакомились с недавней статьей Ю и др. [11], в которой рассмотрена система обслуживания-запасания типа $M/M/1$ с геометрически распределенными объемами требований и исследовались две модели. Одна из моделей позволяла частично удовлетворять требования пользователей в случае если к моменту окончания обслуживания объема запаса было недостаточно для полного удовлетворения требования; во второй модели происходила потеря клиента, требование которого не было удовлетворено полностью, в момент окончания обслуживания данного клиента. Данные модели были исследованы классическим матрично-

аналитическим методом. В то же время, модели, рассмотренные в данной статье отличаются от моделей в работе [11]. Далее приведены указанные отличия, более подробное описание моделей представлено в следующем разделе.

1. В модели 1 клиенты, поступающие в систему с пустым запасом, покидают систему. В этой связи условие стационарности первой модели включает вероятность потери клиента в момент прихода; при этом в работе [11] условие стационарности является достаточным, а следовательно, более ограничительным.

2. В модели 2, потери клиентов могут происходить по двум причинам. Клиент, попадающий в систему с простаивающим сервером и пустым запасом, уходит из системы; клиенты, ожидающие обслуживания (включая того, кто вошел в систему при занятом сервере и пустом запасе) покидают систему в момент окончания обслуживания, в который запас остается пустым (а потому сервер не может начать обслуживание нового клиента). В то же время, в работе [11] потери клиентов происходят либо в момент прихода в систему с пустым запасом, либо после окончания обслуживания, когда сервер не обладает достаточным запасом ресурса для обслуживания (в случае если частичное обслуживание недопустимо). Кроме того, в работе [11] предполагается, что в последнем случае теряется лишь клиент, которому не хватило ресурса, в то время как остальные клиенты (пришедшие в систему с непустым запасом) продолжают оставаться в системе. По нашему мнению, предположение об уходе клиента в момент окончания обслуживания при недостаточном ресурсе не обосновано в реальных приложениях, а также провоцирует излишнюю трату ресурса. На практике начало обслуживания в системах обслуживания-запасания происходит только при условии наличия требуемого объема ресурса (например, при обслуживании автомобиля, таком как замена масла, трансмиссионной жидкости, ремня генератора и т.п.) на момент начала обслуживания. Поэтому при пустом запасе обслуживание не начинается. Кроме того, напомним, условие стационарности в работе [11] является грубым и не включает потери клиентов в момент прихода

(когда ресурс не тратится в отличие от потерь по окончании обслуживания).

3. Мы рассматриваем модели, в которых объем требований имеет распределение общего типа с конечным носителем. В то же время, в работе [11] рассматривается усеченное геометрическое распределение, т.е. также распределение с конечным носителем.

Таким образом, за исключением работы [11], исследований моделей обслуживания-запасания с обслуживанием требований случайного объема в научной литературе не представлено, и наша работа призвана заполнить эту пустоту. В данной работе иследуется две модели, в которых клиенту требуется положительный, но конечный (не более фиксированного N) объем ресурса. Структура статьи следующая. В разделе 2 вводятся модельные предположения и произведен анализ моделей в стационарном режиме. Иллюстративные примеры для сравнения двух моделей представлены в разделе 3, заключительные замечания приведены в разделе 4.

2. Анализ модели 1 и модели 2

В этом разделе представлено описание двух рассматриваемых моделей и проведен анализ моделей в стационарном режиме матрично-аналитическим методом. Обе рассматриваемые модели не ограничены по числу клиентов в системе и являются односерверными. В дальнейшем будем обозначать $\boldsymbol{e}$ вектор-столбец из единиц (соответствующей размерности); через $\boldsymbol{e}_i$ обозначим вектор-столбец имеющий единицу на позиции i и ноль на остальных позициях. Через I обозначена единичная матрица соответствующей размерности.

2.1. Модель 1

В модели 1 предполагается, что клиенты с требованиями случайного объема попадают в систему в моменты пуассоновского процесса интенсивности λ. Типичный объем требования, Y, является случайной величиной с конечным носителем, т.е. объем принимает значения во множестве $\{1, 2, \cdots, N\}$, где $N < S$. При этом вероятность того, что клиенту требуется i единиц ресурса задается как $P(Y = i) = p_i,\ 1 \leq i \leq N$, где $\sum_{i=1}^{N} p_i = 1$. Обслуживание клиента занимает экспоненциально распределенное с параметром μ случайное

время. Клиент, попадающий в систему при пустом запасе, теряется. Если запас не пуст, в момент прихода клиента уровень запаса уменьшается на меньшую величину из предъявленного требования и доступного запаса. Таким образом, удовлетворение требований клиентов осуществляется полностью либо частично. В случае частичного удовлетворения требований происходит потерянный сбыт равный недостающему объему для полного удовлетворения. Как только уровень запаса понизится до s или опустится ниже, запрашивается пополнение запаса. Пополнение происходит через случайное, экспоненциально распределенное с параметром γ, время. В момент пополнения уровень запаса увеличивается на $S - s$ единиц.

2.2. Обобщенный процесс рождения-гибели для модели 1

Представленная выше модель может быть исследована как обобщенный процесс рождения-гибели. Определим состояния процесса. Пусть $N(t)$ означает число клиентов, а $J(t)$ — уровень запаса в момент t. [Заметим, что в данной модели уменьшение запаса (при условии его положительности) происходит в момент прихода клиента, поэтому уровень запаса в момент t отражает состояние запаса, уменьшенного в момент прихода последнего пришедшего к этому времени клиента]. Легко проверить, что $\{(N(t), J(t)) : t \geq 0\}$ является обобщенным процессом рождения-гибели (QBD) с пространством состояний

$$\Omega = \{(i, j) : 0 \leq j \leq S,\ i \geq 0\}.$$

Обозначим $\boldsymbol{i} = \{(i, j) : 0 \leq j \leq S\},\ i \geq 0$ множество состояний процесса, в которых в системе ровно i клиентов и запас находится на одном из $S + 1$ уровней.
Образующая матрица, Q, рассматриваемого QBD процесса имеет вид

$$Q = \begin{pmatrix} A_1 + A_2 & A_0 & & & & \\ A_2 & A_1 & A_0 & & & \\ & A_2 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \end{pmatrix},$$

где блочные подматрицы Q определены далее, с учетом обозначений

$$a = \lambda + \mu,\ \tilde{p}_i = \sum_{k=i+1}^{N} p_k,\ 1 \le i \le N-1.$$

$$A_0 = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \lambda & 0 & 0 & \cdots & 0 & 0 & 0 \\ \lambda\tilde{p}_1 & \lambda p_1 & 0 & \cdots & 0 & 0 & 0 \\ \lambda\tilde{p}_2 & \lambda p_2 & \lambda p_1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \lambda\tilde{p}_{N-1} & \lambda p_{N-1} & \lambda p_{N-2} & \cdots & 0 & 0 & 0 \\ 0 & \lambda p_N & \lambda p_{N-1} & \cdots & 0 & 0 & 0 \\ 0 & 0 & \lambda p_N & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda p_2 & \lambda p_1 & 0 \end{bmatrix}, \quad A_2 = \mu I, \quad (2.1)$$

и

$$A_1 = \begin{bmatrix} -\mu-\gamma & 0 & \cdots & 0 & 0 & \cdots & \gamma & 0 & \cdots & 0 \\ 0 & -a-\gamma & \cdots & 0 & 0 & \cdots & 0 & \gamma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -a-\gamma & 0 & \cdots & 0 & 0 & \cdots & \gamma \\ 0 & 0 & \cdots & 0 & -a & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & -a \end{bmatrix}.$$

Пусть $\boldsymbol{\pi} = (\pi_0, \cdots, \pi_S)$ означает вектор стационарных вероятностей для Марковского процесса, определяемого матрицей $A = A_0 + A_1 + A_2$. В следующей теореме получен критерий устойчивости рассматриваемой модели.

Теорема 2.1. *Модель обслуживания-запасания с образующей матрицей (2.2) является устойчивой тогда и только тогда, когда*

$$\lambda(1 - \pi_0) < \mu. \quad (2.2)$$

Доказательство. критерий следует из общего условия устойчивости QBD процесса $\boldsymbol{\pi} A_0 \boldsymbol{e} < \boldsymbol{\pi} A_2 \boldsymbol{e}$ (см., напр., [9]). □

Замечание: Критерий устойчивости становится очевидным, если показать, что $\lambda(1-\pi_0)$ является эффективной интенсивностью прихода клиентов, а π_0 определяет вероятность того, что клиент придет в систему с пустым запасом.

2.3. Стационарные вероятности состояний в модели 1

Обозначим $\boldsymbol{x}$ вектор стационарных вероятностей состояний процесса, определяемого матрицей Q, т.е. $\boldsymbol{x}$ удовлетворяет системе

$$\boldsymbol{x}Q = \mathbf{0}, \quad \boldsymbol{x}\boldsymbol{e} = 1. \tag{2.3}$$

Разделим вектор на подвектора следующим образом:

$$\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots),$$

где $\boldsymbol{x}_i$ имеет размерность $S+1$.

При удовлетворении критерия стационарности (2.2), вектор стациоарных вероятностей $\boldsymbol{x}$ можно получить (см., напр., [9]) следующим образом:

$$\boldsymbol{x}_i = \boldsymbol{\pi}(I-R)R^i, i \geq 0,$$

где матрица R является наименьшим неотрицательным решением матричного квадратного уравнения с матричным неизвестным:

$$R^2A_2 + RA_1 + A_0 = 0.$$

В случае если S не слишком велико, расчет матрицы R можно осуществлять с помощью численных методов, например, методом логарифмического спуска [8]. В противном случае, необходимо применять (блочный) метод Гаусса–Зейделя, используя особенности структуры матричных коэффициентов A_0, A_1 и A_2.

Ключевые шаги метода логарифмического спуска приведены ниже, подробности можно найти в работе [8].

Алгоритм логарифмического спуска для R:
Шаг 0: $H \leftarrow (-A_1)^{-1}A_0$, $L \leftarrow (-A_1)^{-1}A_2$, $G = L$, and $T = H$.
Шаг 1:

$$U = HL + LH, \quad M = H^2, \quad H \leftarrow (I-U)^{-1}M,$$
$$M \leftarrow L^2, \quad L \leftarrow (I-U)^{-1}M, \quad G \leftarrow G + TL, \quad T \leftarrow TH.$$

Повторять Шаг 1 до выполнения $||\boldsymbol{e} - G\boldsymbol{e}||_\infty < \epsilon$.
Шаг 2: $R = -A_0(A_1 + A_0G)^{-1}$.
Для численных экспериментов было выбрано значение $\epsilon = 10^{-10}$.

2.4. Распределение времени пребывания в системе для модели 1

В данном разделе исследуется время пребывания в системе клиента, попавшего в систему. Заметим, что клиент может быть потерян (не смотря на отсутствие ограничения на общее число клиентов в системе) в момент прихода в систему с пустым запасом. Если же клиент остается в системе, его время пребывания соответствует времени пребывания в классической односерверной системе типа $M/M/1$.

Пусть T — время пребывания в системе клиента, попавшего в систему. Справедлива следующая теорема.

Теорема 2.2. *Пусть $w^*(t)$ есть преобразование Лапласа–Стилтьеса случайной величины T. Тогда*

$$w^*(t) = \frac{\mu}{1 - \sum_{i=0}^{\infty} x_{i0}} \boldsymbol{x}_0 [tI + \mu(I - R)]^{-1}(\boldsymbol{e} - \boldsymbol{e}_1), \ Re(t) \geq 0.$$

Доказательство. результат непосредственно следует из того, что вероятность нахождения в системе i клиентов в момент прихода нового клиента равна $\dfrac{1}{1 - \sum_{i=0}^{\infty} x_{i0}} \boldsymbol{x}_i(\boldsymbol{e} - \boldsymbol{e}_1)$, а следовательно

$$w^*(t) = \frac{1}{1 - \sum_{i=0}^{\infty} x_{i0}} \sum_{i=0}^{\infty} \boldsymbol{x}_i(\boldsymbol{e} - \boldsymbol{e}_1) \Big(\frac{\mu}{t + \mu}\Big)^{i+1}, \ Re(t) \geq 0.$$

□

Следствие: среднее время пребывания в системе, μ_T, определяется формулой

$$\mu_T = \frac{1}{\mu(1 - \sum_{i=0}^{\infty} x_{i0})} \boldsymbol{\pi}(I - R)^{-1}(\boldsymbol{e} - \boldsymbol{e}_1).$$

2.5. Характеристики производительности системы в модели 1

В данном разделе представлены выражения для ключевых характеристик производительности, которые позволяют получить качественное представление о модели 1.

1. *Вероятность простоя системы* (т.е. вероятность того, что в системе нет клиентов) в случайный момент времени, $P^{(1)}_{idle}$, определяется выражением

$$P^{(1)}_{idle} = \boldsymbol{x}_0 \boldsymbol{e}.$$

2. *Вероятность занятости сервера,* $P^{(1)}_{Busy}$, определяется формулой

$$P^{(1)}_{Busy} = 1 - \boldsymbol{x}_0 \boldsymbol{e}.$$

3. *Среднее число клиентов в системе,* $\mu^{(1)}_{NS}$, определяется выражением

$$\mu^{(1)}_{NS} = \sum_{i=1}^{\infty} i \boldsymbol{x}_i \boldsymbol{e} = \boldsymbol{\pi} R (I - R)^{-1} \boldsymbol{e}.$$

4. *Средний уровень запаса,* $\mu^{(1)}_{IL}$, вычисляется следующим образом

$$\mu^{(1)}_{IL} = \sum_{i=1}^{S} i \pi_i.$$

Заметим, что формулу для $\mu^{(1)}_{IL}$ можно получить из следующих наблюдений: выражение $\boldsymbol{x}_0 (I - R)^{-1} \boldsymbol{e}_i$ определяет вероятность того, что уровень запаса равен i, для $0 \leq i \leq S$, кроме того $\boldsymbol{x}_0 (I - R)^{-1} = \boldsymbol{\pi}$.

5. *Вероятность потери клиента в момент прихода в систему с пустым запасом,* $P^{(1)}_{loss}$, определяется следующим образом:

$$P^{(1)}_{loss} = \boldsymbol{x}_0 (I - R)^{-1} \boldsymbol{e}_1 = \pi_0.$$

Заметим, что эффективная интенсивность прихода равна $\lambda(1 - \pi_0)$.

6. *Вероятность частичного удовлетворения требования клиента*, $P_{pmet}^{(1)}$, определяется формулой

$$P_{pmet}^{(1)} = \sum_{j=1}^{N-1} \pi_j \tilde{p}_j.$$

7. *Средняя длительность цикла*, $\mu_{CT}^{(1)}$, определяется как среднее время между пополнениями запаса и может быть получена из выражения

$$\mu_{CT}^{(1)} = \Big[\gamma \sum_{i=0}^{\infty} \sum_{j=0}^{s} x_{ij}\Big]^{-1}.$$

2.6. Модель 2

В модели 2 предполагается, что клиенты с требованиями случайного объема приходят в систему в соответствии с пуассоновским процессом интенсивности λ. Как и в модели 1, каждый клиент требует целочисленный объем ресурса от 1 до N единиц, где $N < S$. Вероятность того, что в систему приходит требование объема i задается как $P(Y = i) = p_i,\ 1 \le i \le N$, где $\sum_{i=1}^{N} p_i = 1$. Время обслуживания клиента экспоненциально распределено с параметром μ. Если к моменту начала обслуживания очередного клиента сервер не может начать обслуживание в связи с опустошением запаса, все ожидающие клиенты покидают систему. Аналогично, клиент, наблюдающий пустой запас и простаивающий сервер в момент прихода, теряется. Таким образом, потери клиентов могут происходить по двум причинам: в момент прихода в систему с простаивающим сервером и пустым запасом, либо в момент окончания обслуживания, опустошающего запас (при наличии ожидающих клиентов). Если уровень запаса снижается до s и ниже, запрашивается пополнение запаса. Предполагается, что время до пополнения экспоненциально распределено с параметром γ. В момент пополнения уровень запаса увеличивается на $S - s$ единиц. Как и в модели 1, в модели допускается частичное удовлетворение потребности клиента объемом запаса, имеющимся на момент начала его обслуживания.

2.7. Процесс типа $GI/M/1$ для модели 2

Анализ модели 2 можно проводить с помощью процесса $GI/M/1$-типа [9]. Аналогично модели 1, определим пространство состояний процесса. Пусть $\hat{N}(t)$ обозначает чисо клиентов в системе, а $\hat{J}(t)$ представляет уровень запаса в системе в момент t. [Заметим, что в данной модели уровень запаса убывает (при условии положительности уровня) в момент начала обслужиавния, а следовательно, уровень запаса в момент t отражает число оставшихся единиц запаса, уменьшенного в момент последнего начала обслуживания клиента, ближайшего к моменту времени t]. Легко проверить, что $\{(\hat{N}(t), \hat{J}(t)) : t \geq 0\}$ является Марковским процессом в непрерывном времени с пространством состояний

$$\hat{\Omega} = \{(i, j) : 0 \leq j \leq S,\ i \geq 0\}.$$

Пусть $\boldsymbol{i} = \{(i, j) : 0 \leq j \leq S\},\ i \geq 0$ есть множество состояний, соответствующих нахождению в системе i клиентов при уровне запаса, принимающем одно из $S + 1$ значений.
Образующая матрица, $\hat{Q}$, исследуемого процесса типа $GI/M/1$ имеет вид

$$\hat{Q} = \begin{pmatrix} B_1 & B_0 & & & & \\ \mu I & \hat{A}_1 & \hat{A}_0 & & & \\ B_2 & \hat{A}_2 & \hat{A}_1 & \hat{A}_0 & & \\ B_2 & & \hat{A}_2 & \hat{A}_1 & \hat{A}_0 & \\ \vdots & & & \ddots & \ddots & \ddots \end{pmatrix},$$

где блочные подматрицы образующей $\hat{Q}$ определены ниже.

$$B_1 = \begin{bmatrix} -\gamma & 0 & \cdots & 0 & 0 & \cdots & \gamma & 0 & \cdots & 0 \\ 0 & -\lambda-\gamma & \cdots & 0 & 0 & \cdots & 0 & \gamma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\lambda-\gamma & 0 & \cdots & 0 & 0 & \cdots & \gamma \\ 0 & 0 & \cdots & 0 & -\lambda & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & -\lambda \end{bmatrix}.$$

$$B_0 = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \lambda & 0 & 0 & \cdots & 0 & 0 & 0 \\ \lambda\tilde{p}_1 & \lambda p_1 & 0 & \cdots & 0 & 0 & 0 \\ \lambda\tilde{p}_2 & \lambda p_2 & \lambda p_1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \lambda\tilde{p}_{N-1} & \lambda p_{N-1} & \lambda p_{N-2} & \cdots & 0 & 0 & 0 \\ 0 & \lambda p_N & \lambda p_{N-1} & \cdots & 0 & 0 & 0 \\ 0 & 0 & \lambda p_N & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda p_2 & \lambda p_1 & 0 \end{bmatrix},$$

$$B_2 = \mu \boldsymbol{e_1}(\boldsymbol{S}+\boldsymbol{1})\boldsymbol{e_1}(\boldsymbol{S}+\boldsymbol{1})',$$

$$\hat{A}_0 = \lambda I, \quad \hat{A}_2 = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \mu & 0 & 0 & \cdots & 0 & 0 & 0 \\ \mu\tilde{p}_1 & \mu p_1 & 0 & \cdots & 0 & 0 & 0 \\ \mu\tilde{p}_2 & \mu p_2 & \mu p_1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \mu\tilde{p}_{N-1} & \mu p_{N-1} & \mu p_{N-2} & \cdots & 0 & 0 & 0 \\ 0 & \mu p_N & \mu p_{N-1} & \cdots & 0 & 0 & 0 \\ 0 & 0 & \mu p_N & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mu p_2 & \mu p_1 & 0 \end{bmatrix},$$

и, наконец,

$$\hat{A}_1 = \begin{bmatrix} -a-\gamma & 0 & \cdots & 0 & 0 & \cdots & \gamma & 0 & \cdots & 0 \\ 0 & -a-\gamma & \cdots & 0 & 0 & \cdots & 0 & \gamma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -a-\gamma & 0 & \cdots & 0 & 0 & \cdots & \gamma \\ 0 & 0 & \cdots & 0 & -a & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & -a \end{bmatrix}.$$

Заметим, что данная система обслуживания всегда стационарна. В случае $\gamma \to \infty$, который соответствует немедленному пополнению запаса (при котором не происходит потерь клиентов), необходимо наложить дополнительное условие $\lambda < \mu$ как необходимое и достаточное

для стационарности. Заметим, что при этом система вырождается в классическую систему $M/M/1$. Данный факт может быть использован как дополнительная проверка при проведении численных экспериментов. В дальнейшем анализе предполагается конечность γ, а следовательно, стационарность системы.

2.8. Стационарное распределение в модели 2

Обозначим $\hat{\boldsymbol{x}}$ вектор стационарных вероятностей состояний, соответствующий матрице $\hat{Q}$. Это означает, что $\hat{\boldsymbol{x}}$ удовлетворяет системе

$$\hat{\boldsymbol{x}}\hat{Q} = \mathbf{0}, \quad \hat{\boldsymbol{x}}\boldsymbol{e} = 1.$$

Разделим вектор следующим образом:

$$\hat{\boldsymbol{x}} = (\hat{\boldsymbol{x}}_0, \hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \cdots),$$

где $\hat{\boldsymbol{x}}_i$ имеет размерность $S+1$.

Вектор стационарных вероятностей $\hat{\boldsymbol{x}}$ можно получить следующим образом (см., напр. [9]):

$$\hat{\boldsymbol{x}}_0 B_1 + \hat{\boldsymbol{x}}_1[\mu I + \hat{R}(I-\hat{R})^{-1}B_2] = \mathbf{0},$$

$$\hat{\boldsymbol{x}}_0 B_0 + \hat{\boldsymbol{x}}_1[\hat{A}_1 + \hat{R}\hat{A}_2) = \mathbf{0},$$

$$\hat{\boldsymbol{x}}_i = \hat{\boldsymbol{x}}_1 \hat{R}^{i-1}, i \geq 1,$$

при выполнении условия баланса

$$[\hat{\boldsymbol{x}}_0 + \hat{\boldsymbol{x}}_1(I-\hat{R})^{-1}]\boldsymbol{e} = 1,$$

где матрица $\hat{R}$ есть минимальное неотрицательное решение матричного квадратного уравнения:

$$\hat{R}^2\hat{A}_2 + \hat{R}\hat{A}_1 + \hat{A}_0 = 0.$$

Вычисление матрицы $\hat{R}$ можно осуществлять известными численными методами, в частности, блочным итеративным методом Гаусса–Зейделя, использующим особенности структуры матричных коэффициентов $\hat{A}_0$, $\hat{A}_1$ и $\hat{A}_2$.

2.9. Время пребывания в системе для модели 2

В данном разделе получены выражения для преобразования Лапласа–Стилтьеса времени пребывания в системе, $\hat{T}$, произвольного отмеченного клиента, принятого в систему.

Заметим, что случайная величина $\hat{T}$ не зависит от последующих приходов. Пусть вектор $\boldsymbol{z}$, представленный в виде $\boldsymbol{z} = (z_0, z_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \cdots)$, означает стационарное распределение состояний системы в момент прихода указанного клиента. Именно, z_0 означает вероятность того, что указанный клиент будет потерян в связи с пустым уровнем запаса при простаивающем сервере; z_1 есть вероятность того, что клиент обнаружит в момент прихода простаивающий сервер с непустым запасом, а следовательно, немедленно поступит на обслуживание; j-я компонента вектора $\boldsymbol{z}_i, i \geq 2$ размерности $S+1$ соответствует вероятности того, что отмеченный клиент наблюдает i клиентов в системе и уровень запаса j в момент сразу после прихода. Легко убедиться в том, что

$$\begin{aligned} z_0 &= x_{0,0}, \\ z_1 &= \hat{\boldsymbol{x}}_0 \boldsymbol{e} - z_0, \\ \boldsymbol{z}_i &= \hat{\boldsymbol{x}}_{i-1}, \ i \geq 2. \end{aligned}$$

Время пребывания клиента, принятого в систему, $\hat{T}$, может рассматриваться как время до поглощения Марковской цепи в непрерывном времени с поглощающим состоянием. Образующая матрица такой цепи, $\tilde{Q}$, имеет следующий вид:

$$\tilde{Q} = \begin{array}{c} * \\ 1 \\ \underline{2} \\ \underline{3} \\ \underline{4} \\ \underline{5} \\ \vdots \end{array} \begin{bmatrix} 0 & 0 & & & & & \\ \mu & -\mu & & & & & \\ \boldsymbol{a}_0 & \boldsymbol{a}_1 & C_1 & & & & \\ \boldsymbol{a}_0 & & C_0 & C_1 & & & \\ \boldsymbol{a}_0 & & & C_0 & C_1 & & \\ \boldsymbol{a}_0 & & & & C_0 & C_1 & \\ \vdots & & & & & \ddots & \ddots \end{bmatrix},$$

где

$$\boldsymbol{a}_0 = \mu \boldsymbol{e}_1(S+1), \ \ \boldsymbol{a}_1 = \mu(\boldsymbol{e}(S+1) - \boldsymbol{e}_1(S+1)), \ \ C_0 = \hat{A}_2, \ \ C_1 = \hat{A}_1 + \hat{A}_2.$$

Пусть $\hat{w}^*(s), Re(s) \geq 0$, есть преобразование Лапласа–Стилтьеса случайной величины $\hat{T}$. Следующая теорема определяет выражение для нахождения $\hat{w}^*(s)$.

Теорема 2.3. *Верно равенство*

$$\hat{w}^*(t) = \frac{1}{1-z_0}\Big[z_1 b_1(t) + \sum_{i=2}^{\infty} \boldsymbol{z}_i \boldsymbol{b}_i(t)\Big], Re(t) \geq 0,$$

где

$$b_1(t) = \mu(t+\mu)^{-1},$$
$$\boldsymbol{b}_2(t) = (tI - C_1)^{-1}\boldsymbol{a}_0 + (tI - C_1)^{-1}\boldsymbol{a}_1 b_1(t),$$
$$\boldsymbol{b}_i(t) = (tI - C_1)^{-1}\boldsymbol{a}_0 + (tI - C_1)^{-1}\boldsymbol{a}_1 \boldsymbol{b}_{i-1}(t),\ Re(t) \geq 0.$$

Доказательство. непосредственно следует из формулы полной вероятности с учетом следующих ключевых фактов: (а) $\dfrac{z_1}{1-z_0}$ определяет условную вероятность того, что клиент попадает на обслуживание без ожидания, а следовательно, преобразование Лапласа–Стилтьеса определяется выражением $\dfrac{t}{t+\mu}$; (б) вектор $\dfrac{1}{1-z_0}\boldsymbol{z}_i,\ i \geq 2$ определяет вероятность того, что запас находится на одном из $S+1$ уровней, и клиенту необходимо ожидать $i-1$ окончаний обслуживания до попадания на сервер либо ухода в связи с опустошением запаса. □

Следствие: среднее время пребывания, $\mu_{\hat{T}}$, определяется формулой

$$\mu_{\hat{T}} = \frac{1}{1-z_0}\left[\frac{z_1}{\mu} + \sum_{i=2}^{\infty} \boldsymbol{z}_i \hat{\boldsymbol{b}}_i\right],$$

где $\hat{\boldsymbol{b}}_i = -\dfrac{d}{ds}\boldsymbol{b}_i(s)\Big|_{s=0},\ i \geq 2$, удовлетворяет следующей системе, пригодной для рекуррентного вычисления. Обозначим

$$D = (-C_1)^{-1}C_0,\ \boldsymbol{d}_1 = (-C_1)^{-1}\boldsymbol{e},\ \boldsymbol{d}_2 = (-C_1)^{-1}(\boldsymbol{e} - \boldsymbol{e}_1),$$

тогда несложно убедиться в том, что

$$\hat{\boldsymbol{b}}_2 = \boldsymbol{d}_1 + \boldsymbol{d}_2,$$
$$\hat{\boldsymbol{b}}_i = \boldsymbol{d}_1 + D\hat{\boldsymbol{b}}_{i-1},\ i \geq 3.$$

2.10. Характеристики производительности системы в модели 2

В данном разделе представлены выражения для ключевых характеристик производительности, которые позволяют получить качественное представление о модели 2, по аналогии с моделью 1.

1. *Вероятность простоя системы* (т.е. вероятность того, что в системе нет клиентов) в случайный момент времени, $P^{(2)}_{idle}$, равна
$$P^{(2)}_{idle} = \hat{\boldsymbol{x}}_0 \boldsymbol{e}.$$

2. *Вероятность занятости сервера,* $P^{(2)}_{Busy}$, определяется выражением
$$P^{(2)}_{Busy} = 1 - \hat{\boldsymbol{x}}_0 \boldsymbol{e}.$$

3. *Среднее число клиентов в системе,* $\mu^{(2)}_{NS}$, определяется формулой
$$\mu^{(2)}_{NS} = \sum_{i=1}^{\infty} i \hat{\boldsymbol{x}}_i \boldsymbol{e} = \hat{\boldsymbol{x}}_1 (I - \hat{R})^{-2} \boldsymbol{e}.$$

4. *Средний уровень запаса,* $\mu^{(2)}_{IL}$, равен
$$\mu^{(2)}_{IL} = \sum_{i=1}^{S} i \Big[(\hat{\boldsymbol{x}}_0 + \hat{\boldsymbol{x}}_1 (I - \hat{R})^{-1}) \boldsymbol{e}_{i+1} \Big].$$

5. *Вероятность потери клиента при пустом запасе,* $P^{(2)}_{losA}$, в момент прихода клиента, определяется выражением
$$P^{(2)}_{losA} = \hat{\boldsymbol{x}}_0 \boldsymbol{e}_1$$

6. *Вероятность потери ожидающего клиента,* $P^{(2)}_{losD}$, в связи с опустошением запаса в момент окончания обслуживания определяется как отношение интенсивности потерь клиентов к интенсивности входного потока $\frac{R_{losD}}{\lambda}$, и равна
$$P^{(2)}_{losD} = \frac{\mu}{\lambda} \sum_{i=1}^{\infty} (i-1) \hat{\boldsymbol{x}}_i \boldsymbol{e}_1 = \frac{\mu}{\lambda} \Big[\hat{\boldsymbol{x}}_1 \hat{R} (I - \hat{R})^{-1} \boldsymbol{e}_1 \Big].$$

7. *Вероятность потери клиента в связи с опустошением запаса*, $P_{loss}^{(2)}$, представлена двумя составляющими:

$$P_{loss}^{(2)} = P_{losA}^{(2)} + P_{losD}^{(2)} = 1 - \frac{\mu}{\lambda}[1 - \hat{\boldsymbol{x}}_0 \boldsymbol{e}].$$

8. *Вероятность частичного удовлетворения требования*, $P_{pmet}^{(2)}$, определяется (с учетом того, что $\hat{x}_{0,j}$ есть вероятность попадания в систему с простаивающим сервером и j единицами запаса, а $\frac{\mu}{\lambda}\hat{\boldsymbol{x}}_1\hat{R}(I-\hat{R})^{-1}\boldsymbol{e}_{j+1}$ есть вероятность наблюдать хотя бы одного ожидающего клиента и j единиц запаса в момент окончания обслуживания) следующим образом:

$$P_{pmet}^{(2)} = \sum_{j=1}^{N-1} \left[\hat{x}_{0,j} + \frac{\mu}{\lambda}\hat{\boldsymbol{x}}_1\hat{R}(I-\hat{R})^{-1}\boldsymbol{e}_{j+1} \right] \tilde{p}_j.$$

9. *Средняя длительность цикла*, $\mu_{CT}^{(2)}$, определяемая как среднее время между пополнениями запаса, может быть вычислена по формуле

$$\mu_{CT}^{(2)} = \left[\gamma \sum_{i=0}^{\infty} \sum_{j=0}^{s} \hat{x}_{ij} \right]^{-1}.$$

3. Численные примеры

В данном разделе представлены результаты сравнения двух моделей на двух примерах. Для обоих моделей используются четыре перечисленные ниже распределения объема требований.

Распределение объема требований: пусть Y озачает типичный случайный объем требований. Четыре распределения для Y таковы:
1. Требование постоянного объема (C): здесь $P(Y = C) = 1$.

2. Усеченное распределение Пуассона (P):

$$P(Y=i) = \begin{cases} e^{-\theta}\frac{\theta^{i-1}}{(i-1)!}, \ i = 1, \cdots, N-1, \\ \\ \sum_{k=N}^{\infty} e^{-\theta}\frac{\theta^{k-1}}{(k-1)!}, \ i = N. \end{cases}$$

3. Усеченное геометрическое распределение (G):

$$P(Y=i)=\begin{cases}(1-\zeta)\zeta^{i-1},\ i=1,\cdots,N-1,\\ \\ \zeta^{N-1},\ i=N.\end{cases}$$

4. Равномерное (U):

$$P(Y=i)=\frac{1}{N},\ i=1,\cdots,N.$$

Для корректного сравнения нормализуем параметры указанных распределений таким образом, чтобы средний объем требований совпадал для заданного N, т.е. выберем θ и ζ так, чтобы средний объем требований был равен $0.5(N+1)$. Заметим также, что при $N=1$ применим только сценарий с требованиями постоянного объема. На графиках обозначим через $C1, C2, C3$ и $C4$ сценарии, соответствующие постоянным объемом требований размера 1, 2, 3 и 4 соответственно. Аналогично, $P1, P2, P3$ и $P4$ соответствуют усеченному Пуассоновскому распределению; $G1, G2, G3$ и $G4$ — усеченному геометрическому распределению; $U1, U2, U3$ и $U4$ — равномерному распределению.

Пример 1: в данном примере используются фиксированные значения параметров $\lambda=1, \mu=1.1, \gamma=0.1$, при этом $N=1,3,5$ и $N=7$. Заметим, что при этом средний случайный объем принимает значения 1, 2, 3 и 4. На Рис. 1 и 2, соответственно, представлены случаи $s=20, S=80$ и $s=30, S=80$, и приведено сравнение характеристик производительности первой и второй моделей.

Анализ графиков позволяет сделать следующие выводы:

1. В отношении вероятности потери клиента наблюдается неравенство $P_{loss}^{(1)} < P_{loss}^{(2)}$, которое имеет следующее интуитивное объяснение. В то время как в модели 1 клиенты теряются при пустом запасе, в модели 2 клиенты могут быть потеряны как в момент прихода, так и в момент окончания обслуживания. В то же время, анализ графиков вероятности потери только в момент прихода (данные графики не приводятся в статье)

показал, что, как и ожидается, вероятность потери в момент прихода в модели 1 выше, чем в модели 2. Это справедливо для всех рассматриваемых четырех типов распределений случайного объема. Заметим также, что для обоих моделей вероятность потери возрастает с ростом среднего объема требования.

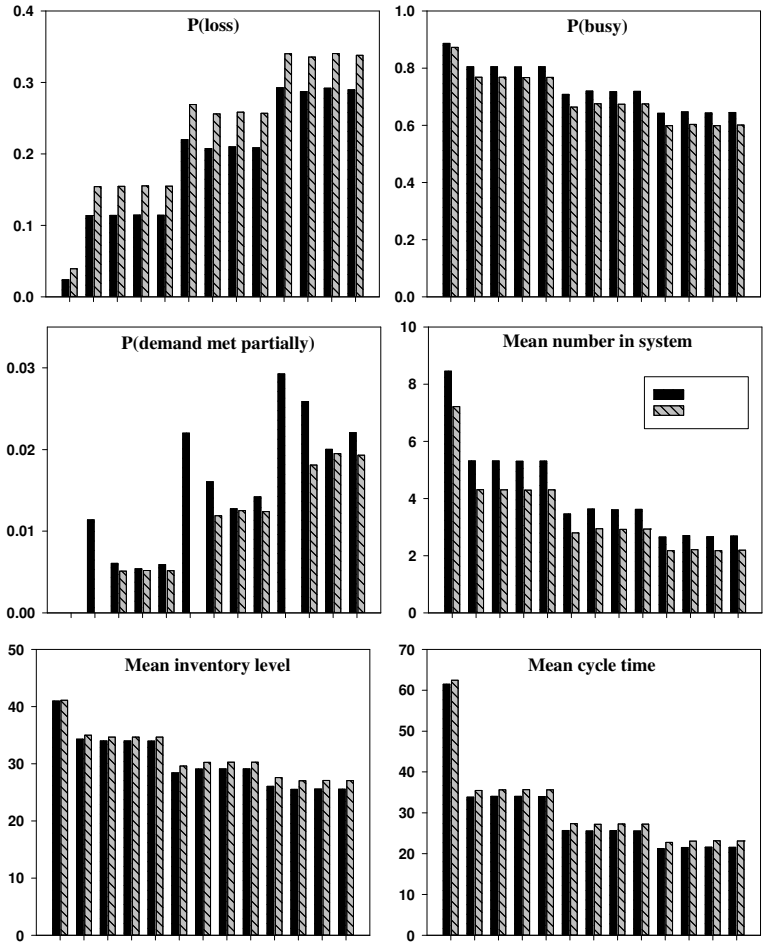

Рисунок 1. Сравнение двух моделей при $S = 80$ и $s = 20$

Наблюдаются также значимые различия в зависимости от типа распределения случайного объема, подобные различия наблюдаются как при $s = 20$, так и при $s = 30$. В то же время, вероятность потерь меньше при $s = 30$ по сравнению с $s = 20$, как и ожидается в связи с более ранним пополнением запаса

при $s = 30$.

2. В соответствии с предыдущим наблюдением, $P_{busy}^{(1)} > P_{busy}^{(2)}$ для всех сценариев. Как и ожидается, вероятность занятости выше при $s = 30$ по сравнению с $s = 20$ для всех сценариев.

3. В отношении вероятности частичного удовлетворения требования в начальный момент обслуживания, можно отметить следующие интересные наблюдения. Во-первых, для всех сценариев $P_{pmet}^{(2)} < P_{pmet}^{(1)}$, что неудивительно, поскольку в модели 2 пополнение запаса возможно даже для таких клиентов, которые прибыли в систему с пустым запасом (при условии, что сервер был занят все время до пополнения). В то же время можно отметить, что в случае запросов постоянного объема, в модели 1 вероятность частичного удовлетворения значительно превышает таковую в модели 2. Кроме того, при большом значении среднего объема, распределение объема играет значительную роль для значения указанной характеристики в модели 1. Это справедливо как для случая $s = 20$, так и при $s = 30$.

4. Для среднего числа клиентов в системе выполнено $\mu_{NS}^{(2)} < \mu_{NS}^{(1)}$ для всех сценариев. Интересно отметить, что в обоих моделях данная характеристика убывает с ростом среднего объема требования. Интуитивно это неожиданное наблюдение можно объяснить частичным удовлетворением требований. Указанное соотношение верно как при $s = 20$, так и при $s = 30$.

5. Наконец, в ходе экспериментов не было отмечено значимых различий в двух характеристиках: среднем уровне запаса и средней длительности цикла, в случаях $s = 20$ и $s = 30$.

Поскольку эффективная интенсивность приходов $\lambda_e = \lambda(1 - P_{loss}^{(1)})$ в модели 1 зависит как от типа входного процесса, так и от других параметров модели, необходимо провести сравнение двух моделей при заданной интенсивности входного потока λ_e в модели 2. Данный эксперимент обсуждается ниже в Примере 2.

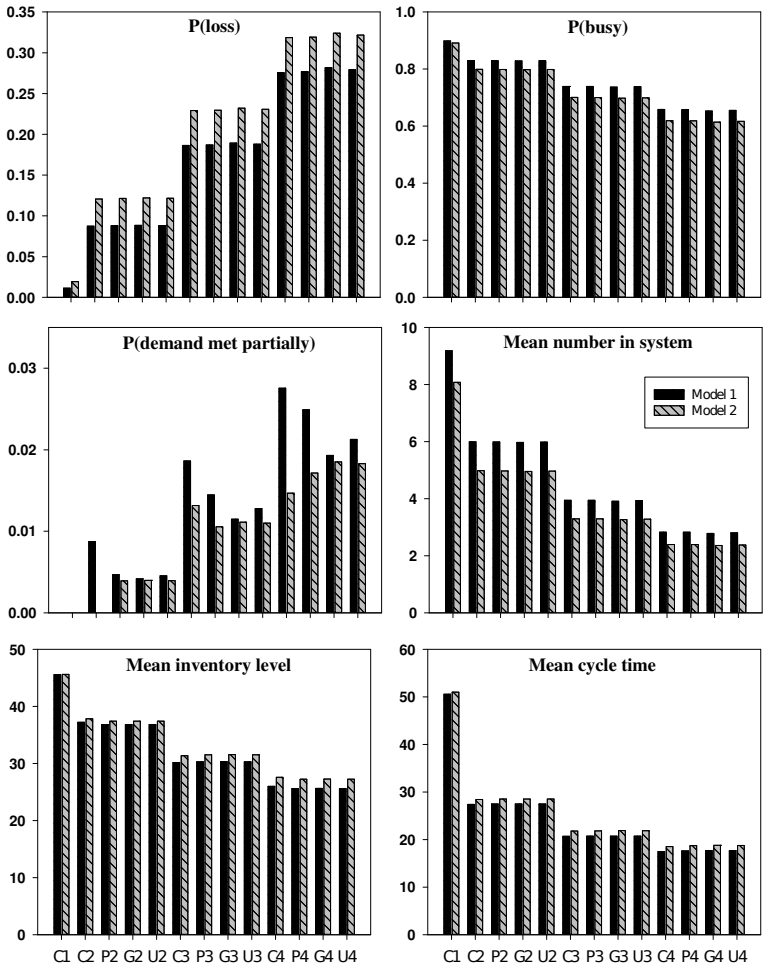

Рисунок 2. Сравнение двух моделей при $S = 80$ и $s = 30$

Пример 2: Данный эксперимент аналогичен Примеру 1, но интенсивность прихода клиентов в модели 2 выбирается равной соответствующей эффективной интенсивности приходов в модели 1. На Рис. 3 и 4, соответственно, для случаев $s = 20, S = 80$ и $s = 30, S = 80$, приведены сравнительные графики характеристик моделей, наблюдение которых позволяет сделать следующие выводы.

1. При $s = 20$ вероятности потерь в обоих моделях очень близки для всех сценариев, за исключением случая среднего объема требования равного 1. В последнем случае $P_{loss}^{(1)} < P_{loss}^{(2)}$. В то же время, при $s = 30$ можно заметить, что $P_{loss}^{(2)} < P_{loss}^{(1)}$. Как

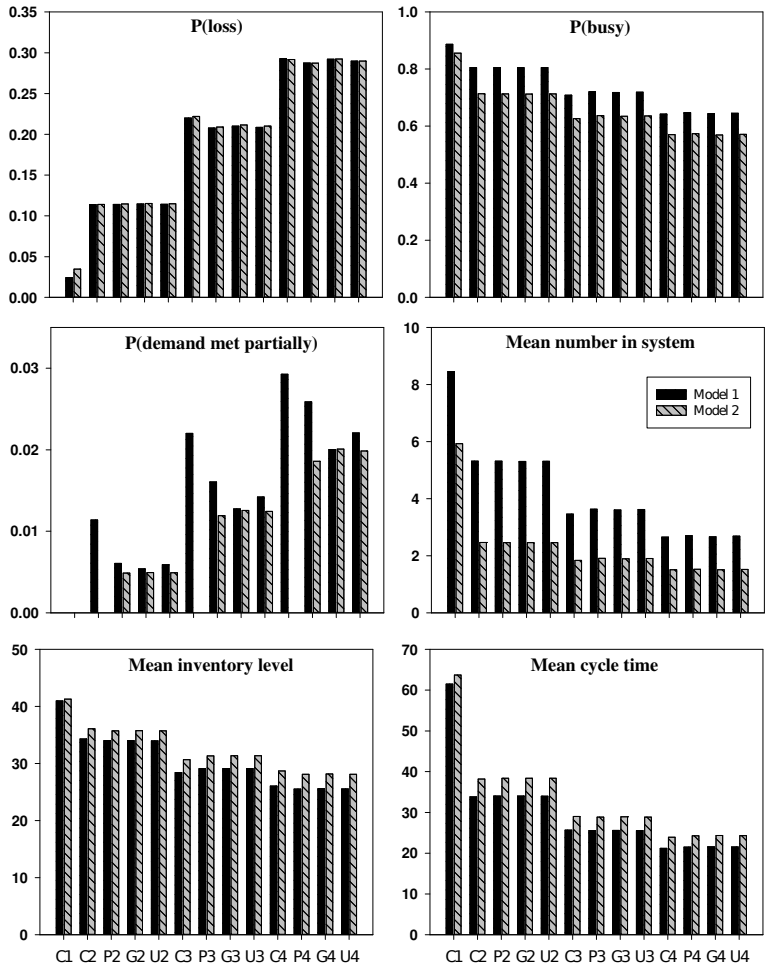

Рисунок 3. Сравнение двух моделей при $S = 80$ и $s = 20$

было отмечено в Примере 1, данная характеристика возрастает с ростом среднего объема. Кроме того, данная характеристика нечувствительна к типу распределения объема требований.

2. Неравенство $P^{(2)}_{busy} < P^{(1)}_{busy}$ выполнено для всех сценариев, как при $s = 20$, так и при $s = 30$. Кроме того, наблюдается существенное уменьшение $P^{(2)}_{busy}$ при переходе от $s = 20$ к $s = 30$.

3. Среднее число клиентов в системе значимо убывает в модели 2. Это не удивительно, поскольку интенсивность приходов в данном эксперименте значительно ниже по сравнению с Приме-

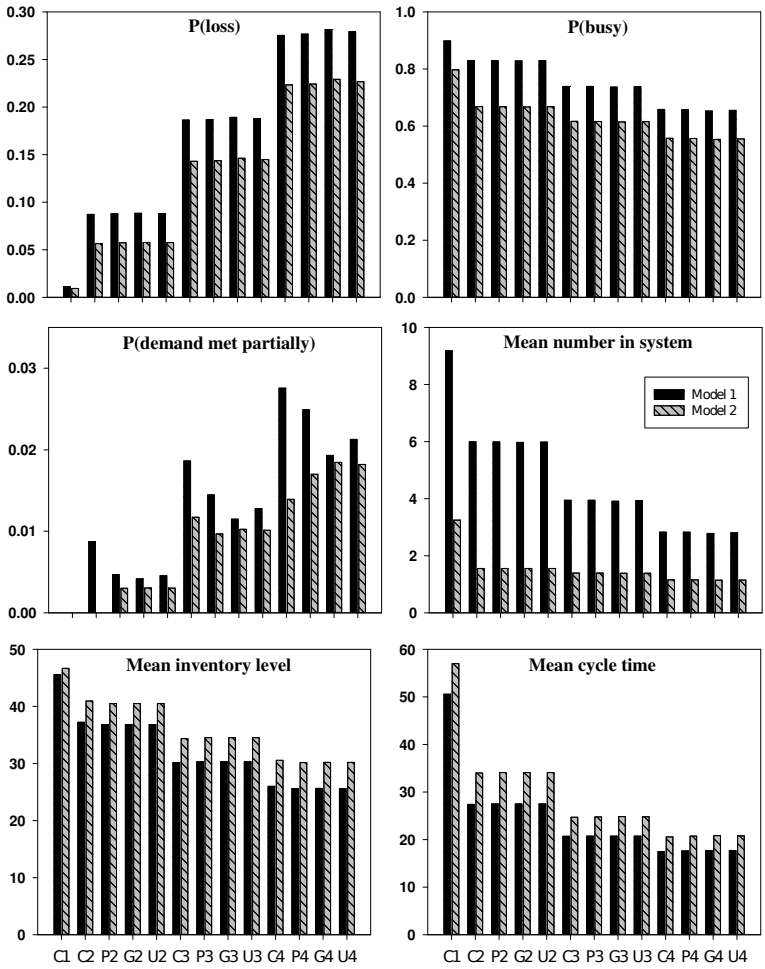

Рисунок 4. Сравнение двух моделей при $S = 80$ и $s = 30$

ром 1. В то же время, скорость убывания данной характеристики значительно превышает скорость убывания интенсивности приходов. Это справедливо и при $s = 20$, и при $s = 30$.

4. Обе характеристики, средний уровень запаса и средняя длительность цикла, существенно выросли в модели 2 по сравнению с их значениями в Примере 1.

5. Как и в Примере 1, наблюдается выполнение неравенства

$$P_{pmet}^{(2)} < P_{pmet}^{(1)}.$$

6. Таким образом, модель 2 имеет лучшие характеристики, чем модель 1, при равной интенсивности (эффективного) входного потока.

4. Заключение

Научная литература по системам обслуживания-запасания многочисленна и ее объем продолжает увеличиваться. В то же время, модели с требованиями случайного объема и положительным временем обслуживания мало изучены. Интерес к данному типу моделей возник в научном сообществе сравнительно недавно. В данной работе представлен анализ двух моделей с обслуживанием требований случайного объема и приведено сравнение моделей для практической демонстрации их преимуществ. Модели, представленные в статье, могут быть обобщены за счет применения более гибкого входного процесса и распределения времени обслуживания фазового типа, однако данное исследование выходит за рамки этой статьи.

Благодарности

Автор благодарит анонимных рецензентов за внимательное прочтение и конструктивные замечания, которые позволили улучшить представление результатов в статье.

СПИСОК ЛИТЕРАТУРЫ

1. Berman O., Kaplan E.H., Shimshak D.G. *Deterministic approximations for inventory management at service facilities* // IIE Trans. 1993. V. 25. N. 5. P. 98–104.

2. Bradley E. L. *Queues with balking and their application to an inventory problem.* In Technical Report, DTIC Document. 1969.

3. Chakravarthy S.R., Maity A., Gupta U.C. *Modeling and Analysis of Bulk Service Queues with an Inventory under "*(s, S)*" Policy* // Annals of Operations Research. 2017. V. 258. P. 263–283.

4. Choi K. H., Yoon B. K. *A Survey on the Queueing Inventory Systems with Phase-type Service Distributions* // QTNA '16, December 13-15, 2016, Wellington, New Zealand, Proceedings. 2016.

5. Karthikeyan K., Sudhesh R. *Recent review article on queueing inventory systems* // Research J. Pharm. and Tech. 2016. V. 9. N. 11. P. 1451–1461.

6. Krishnamoorthy A., Lakshmy B., Manikandan R. *A survey on inventory models with positive service time* // OPSEARCH. 2011. V. 48, N. 2, 153–169.

7. Krishnamoorthy A., Manikandan R., Lakshmy B. *A revisit to queueing-inventory system with positive service time* // Annals of Operations Research. 2015. V. 233, P. 221–236.

8. Latouche G., Ramaswami V. *Introduction to matrix analytic methods in stochastic modeling.* SIAM. 1999.

9. Neuts M.F. *Matrix-geometric solutions in stochastic models: An algorithmic approach.* The Johns Hopkins University Press, Baltimore, MD. 1981.

10. Saidane S., Babai M., Aguir M. *On the performance of the base-stock inventory system under a compound erlang demand distribution* // Computers and Industrial Engineering. 2013. V. 66. P. 548–554.

11. Yue D., Zhao G., Qin Y. *An $M/M/1$ Queueing-Inventory System with Geometric Batch Demands and Lost Sales* // Journal of Systems Science and Complexity. 2018. V. 31. P. 1024–1041.

QUEUEING-INVENTORY MODELS WITH BATCH DEMANDS AND POSITIVE SERVICE TIMES

Srinivas Chakravarthy, Departments of Industrial and Manufacturing Engineering & Mathematics
Kettering University (schakrav@kettering.edu).

Abstract: We consider queueing-inventory models in which customers arrive according to a point process and each customer demands for one or more items but not exceeding a pre-determined (finite) value, say, N. The demands of the customers require positive service times. Replenishments are based on (s, S)-type policy and the lead times are assumed to be random. We consider two models. In Model 1, any arriving customer finding the inventory level to be zero will be lost. In Model 2, the loss of customers occur in two ways. First, an arriving customer finding the inventory level to be zero with the server being idle will be lost, and secondly, the customers, if any, present at a service completion with zero inventory will all be lost. We assume that in both the models the demands may be met partially based on the requests and the available inventory levels at that time. The inventory level is reduced by the amount to meet (fully or partially) the requisite demand of the customer at the beginning of the service. Under the assumption that all underlying random variables are exponential, we perform the steady-state analysis of the models using the classical matrix-analytic methods. Illustrative examples comparing the two models are presented.

Keywords: queueing-inventory systems, algorithmic probability, batch demands, lead times, matrix-analytic methods.