



О законе распределения вероятностей случайной величины



Марина ЗОТОВА

Marina A.ZOTOVA

В статье исследуется закон распределения вероятностей случайной величины на примере времени реакции на поступающие запросы в центре обработки данных (ЦОД), представленном как система массового обслуживания.

Для решения поставленной задачи автор проверяет гипотезы об экспоненциальном распределении, использует в своем исследовании положения теории массового обслуживания и методы имитационного моделирования. Полученные при этом результаты дают вполне обоснованные ориентиры пользователям ЦОД и возможность соотносить расчетные величины с действующими стандартами и растущими потребностями системы.

Ключевые слова: теория массового обслуживания, теория вероятностей, центр обработки данных, распределение вероятностей случайной величины, время реакции, моделирование.

Зотова Марина Александровна — технолог 1-й категории проектно-конструкторско-технологического бюро по системам информатизации (ПКТБ ЦКИ) ОАО «РЖД», аспирант кафедры «Вычислительная техника» Российской открытой академии транспорта Московского государственного университета путей сообщения (МИИТ), Москва, Россия.

Центр обработки данных (ЦОД) работает при случайном потоке запросов на обработку и случайном времени их выполнения. Исследования, проведенные ранее, позволяют представить ЦОД как систему массового обслуживания (СМО) типа $M/M/1/\infty$. Это условное обозначение означает простейший входной поток, экспоненциальное распределение времени обслуживания, одно обслуживающее устройство, бесконечный входной поток, при котором вероятность поступления нового запроса не зависит от того, сколько уже имеется запросов в системе. Расчет времени реакции ЦОД — времени от момента поступления запроса в систему из внешней среды — достаточно согласуется с экспериментальными данными, поэтому искать ответ на вопрос о соответствии работы ЦОД требованиям стандарта ITIL [1] следует, опираясь на разнообразные модели СМО.

Исследования периода создания самых первых человеко-машинных систем, в которых человек решал свои задачи в режиме диалога с машиной, показали, что у пользователя возникают отрицательные эмоции в том случае, если время ответа превышает 2 секунды [2, 3]. Стандарт ЦОД подразуме-

вает, что величину допустимого значения времени реакции разработчик согласовывает с заказчиком системы [1].

В силу того, что время реакции ЦОД — величина случайная, оценивать его на модели можно лишь как среднее значение. Причем необходимо использовать методы теории массового обслуживания и имитационного моделирования, чтобы получить вероятность того, что время ответа превысит наперед заданную величину максимально допустимого времени реакции, предусмотренную стандартами. Если применить модель $M/M/1/\infty$, то тогда закон распределения вероятностей случайной величины — времени пребывания запроса в системе, определяемой как сумма случайных времен пребывания в очереди и собственно времени обслуживания, может быть применен, исходя из следующих соображений.

Если запрос поступает в СМО и застаёт её свободной с вероятностью p_0 , то тогда время пребывания в системе есть случайная величина, имеющая экспоненциальное распределение с μ -параметром времени обслуживания, поскольку имеет место обработка поступившего запроса.

Если запрос поступает в СМО и застаёт в ней один запрос на обслуживание с вероятностью p_1 , то тогда время пребывания в системе есть случайная величина, представляющая из себя сумму двух экспоненциальных величин с μ -параметром времени обслуживания. Первая — остаток времени дообслуживания того запроса, что был на обслуживании в момент поступления, вторая — обслуживание вновь поступившего запроса. Из теории вероятностей известно [4], что в случае экспоненциального распределения остаток времени дообслуживания имеет то же самое экспоненциальное распределение с тем же самым параметром μ . Следовательно, распределение времени пребывания в СМО здесь есть распределение Эрланга 2-го порядка [5].

Если запрос поступает в СМО и застаёт в ней i запросов (один на обслуживании, $(i-1)$ в очереди) с вероятностью p_i , то тогда время пребывания в системе есть случайная величина, представляющая из себя сумму $(i+1)$ экспоненциально распределённых случайных величин с μ — параметром времени обслуживания. Первая — остаток

времени дообслуживания того запроса, что был на обслуживании в момент поступления, остальные — обслуживание запросов, выбранных из очереди, и вновь поступившего запроса. И дообслуживание, и обслуживание принятых запросов из очереди есть случайные величины, имеющие экспоненциальное распределение с одним и тем же параметром μ . Следовательно, распределение времени пребывания в СМО в этом случае есть распределение Эрланга $(i+1)$ -го порядка.

Распределение Эрланга с порядком $v=(i+1)$ и параметром μ имеет плотность распределения вероятностей [5]:

$$f(t) = \mu (\mu t)^{v-1} e^{-\mu t} / (v-1)!, t \geq 0, v > 0. \quad (1)$$

И значит, плотность распределения вероятностей случайной величины — времени пребывания запроса в СМО $f_{\text{reaction}}(t)$ выражается формулой:

$$f_{\text{reaction}}(t) = \sum_{i=1}^{\infty} p_{i-1} \mu (\mu t)^{v-1} e^{-\mu t} / (v-1)!, t \geq 0. \quad (2)$$

Суммирование проводится от 1 до бесконечности, поскольку СМО $M/M/1/\infty$ получает бесконечный входной поток, следовательно, теоретически имеет место ненулевая вероятность нахождения сколь угодно большого количества запросов.

Вероятности $p_i, i=0,1,2,3,\dots$ вычисляются по формулам, приводимым в теории массового обслуживания [5]. Когда речь идёт о поступлении запроса в систему, то следует говорить об условных вероятностях $\tilde{p}_i, i=0,1,2,3,\dots$ и той ситуации, что запрос встретит там i запросов в момент поступления. В случае замкнутых СМО эти условные вероятности отличаются от стационарных вероятностей нахождения системы в состоянии i . Однако, как показано в [6], для $M/M/1/\infty$ вероятности $\tilde{p}_i$ совпадают со стационарными вероятностями нахождения СМО в состоянии i .

Нетрудно увидеть, что использовать громоздкую формулу (2) для вычисления вероятности того, что случайная величина превысит наперед заданную, довольно сложно при инженерных расчетах, так как необходимо выполнить численное интегрирование. Поэтому применим метод имитационного (статистического) моделирования [7]. А именно возьмем общецеле-





* МОДЕЛЬ РАЗОМКНУТОЙ Системы массового ОБСЛУЖИВАНИЯ
типа M/M/1/бесконечность
* 1 ЕДИНИЦА МОДЕЛЬНОГО ВРЕМЕНИ = 0,1 МИЛЛИСЕКУНДЕ РЕАЛЬНОГО ВРЕМЕНИ
SIMULATE
EXPON FUNCTION RN1,C24 ЭКСПОНЕНЦИАЛЬНОЕ РАС ПРЕДЕЛЕНИЕ (СРЕ ДНЕ=1)
0,0/1,104/2,222/3,355/4,509/5,69/
0,6,915/7,12/75,138/8,16/84,183/88,212/
0,9,23/92,252/94,281/95,299/96,32/97,35/
0,98,39/99,46/99,53/99,62/99,7/99,78
XTIME TABLE M1,0,1000,50
INPUTGENERATE 189 0, FNS EXPON ПУАССОНОВСКИЙ ВХОДНОЙ ПОТОК
QUEUE SYSTEM ИЗМЕРЕНИЕ ВРЕМЕНИ В СИСТЕМЕ
* ИМИТИРУЕМ ОБСЛУЖИВАНИЕ В СМО
QUEUE PHASE1 ИЗМЕРЕНИЕ ВРЕМЕНИ В ФАЗЕ 1
QUEUE LINE1 ИЗМЕРЕНИЕ ВРЕМЕНИ В ОЧЕРЕДИ 1
SEIZE SMO1 ЗАНЯТИЕ ОДНОГО КАНАЛА ОБСЛУЖИВАНИЯ В СМО1
DEPART LINE1 ОКОНЧАНИЕ ИЗМЕРЕНИЯ ВРЕМЕНИ В ОЧЕРЕДИ 1
OBSLU ADVANCE 1 89 FNS EXPON ИМИТАЦИЯ ОБСЛУЖИВАНИЯ В ФАЗЕ 1
RELEASE SMO1 ОСВОБОЖДЕНИЕ ОДНОГО КАНАЛА ОБСЛУЖИВАНИЯ В СМО1
DEPART PHASE1 ОКОНЧАНИЕ ИЗМЕРЕНИЯ ВРЕМЕНИ В ФАЗЕ 1
DEPART SYSTEM ОКОНЧАНИЕ ИЗМЕРЕНИЯ ВРЕМЕНИ В СИСТЕМЕ
TABULATE XTIME ГИСТОГРАММА СУММАРНОГО ВРЕМЕНИ
TERMINATE 1
START 10000,NP РАЗГОН МОДЕЛИ
RESET СБРОС НАКОПЛЕННЫХ СТАТИСТИК
START 10000 ОСНОВНОЕ МОДЕЛИРОВАНИЕ
CLEAR
OBSLU ADVANCE 900,FNS EXPON ИМИТАЦИЯ ОБСЛУЖИВАНИЯ В ФАЗЕ 1
START 10000,NP РАЗГОН МОДЕЛИ
RESET СБРОС НАКОПЛЕННЫХ СТАТИСТИК
START 10000 ОСНОВНОЕ МОДЕЛИРОВАНИЕ
END

Рис. 1. GPSS-модель для получения гистограммы времени реакции ЦОД.

вую систему моделирования дискретных сложных систем GPSS [8].

Модель СМО на языке GPSS выглядит относительно просто и представлена на рис.1.

Среднее время выполнения запроса в мэйнфрейме ЦОД составляет 0,0189 с для варианта с производительностью 5939 Mips. В модели за единицу времени принято 0,1 мс, поэтому первый параметр блока *GENERATE*, имитирующего простейший поток запросов в СМО, равен 1890 единицам модельного времени, а первый параметр блока *ADVANCE*, имитирующего время обслуживания – 189 единицам. При таких параметрах коэффициент загрузки обслуживающего устройства с именем СМО1 теоретически должен быть равен 0,1.

Результаты моделирования показывают, что полученная экспериментальным путем (моделированием) загрузка устройства FACILITY/UTIL составляет 0,099, то есть практически равна теоретической.

Гистограмма случайной величины – времени пребывания запроса в системе описывается строками:

TABLE	MEAN	STD. DEV.	RANGE	RETRY	FREQUENCY	CUM. %
XTIME	209,869	206,684		0		
	0,000	-	100,000		3808	38,08
	100,000	-	200,000		2271	60,79
	200,000	-	300,000		1512	75,91
	300,000	-	400,000		898	84,89
	400,000	-	500,000		585	90,74
	500,000	-	600,000		381	94,55
	600,000	-	700,000		207	96,62
	700,000	-	800,000		129	97,91
	800,000	-	900,000		64	98,75
	900,000	-	1000,000		46	99,21
	1000,000	-	1100,000		34	99,55
	1100,000	-	1200,000		20	99,75
	1200,000	-	1300,000		10	99,85
	1300,000	-	1400,000		6	99,91
	1400,000	-	1500,000		3	99,94
	1500,000	-	1600,000		3	99,97
	1600,000	-	1700,000		1	99,98
	1700,000	-	1800,000		0	99,98
	1800,000	-	1900,000		0	99,98
	1900,000	-	2000,000		1	99,99
	2000,000	-	2100,000		0	99,99
	2100,000	-	2200,000		0	99,99
	2200,000	-	2300,000		1	100,00

Таблица 1

Вероятности того, что время реакции ЦОД будет не больше Т, при коэффициенте нагрузки 0,1

Коэффициент загрузки	Граница, Т	Вероятность
0,1	< 0,0100 с.	0,3443
	< 0,0200 с.	0,5647
	< 0,0300 с.	0,7145
	< 0,0400 с.	0,8114
	< 0,0500 с.	0,8770
	< 0,0600 с.	0,9212
	< 0,0700 с.	0,9462
	< 0,0800 с.	0,9647
	< 0,0900 с.	0,9763
	< 0,1000 с.	0,9834
	< 0,1100 с.	0,9897
	< 0,1200 с.	0,9930
	< 0,1300 с.	0,9959
	< 0,1400 с.	0,9975
	< 0,1500 с.	0,9986
	< 0,1700 с.	0,9994

Среднее время пребывания в системе – 209,869, средне-квадратическое отклонение (корень квадратный из дисперсии) этой величины – 206,684. Таким образом, коэффициент вариации случайной величины практически равен 1, что позволяет (основываясь на частотах попадания в разряды гистограммы) сделать предположение о том, закон распределения вероятностей случайной величины – времени пребывания в СМО – экспоненциальный.

Проверка гипотезы об экспоненциальном распределении с помощью χ^2 – квадрата критерия показала, что гипотеза может быть принята с доверительной вероятностью 0,62269368. Этот результат ожидаем, так как основной вклад в случайную величину – время реакции ЦОД вносит первое слагаемое в формуле (2), при этом среднее число запросов в системе (по результатам моделирования) составляет 0,1, а среднее время в системе – 209,869, что незначительно отличается от чистого времени обслуживания.

Результаты в графическом виде представлены на рис. 2. Прямоугольники – экспериментальная гистограмма, полученная путем имитационного моделирования. Кривая плотности распределения вероятностей показана пунктиром. По оси х

Таблица 2
Вероятности того, что время реакции ЦОД
будет не больше Т для коэффициента
нагрузки 0,225537

Коэффициент загрузки	Граница, Т	Вероятность
0,225537	< 0,0100 с.	0,3355
	< 0,0200 с.	0,5516
	< 0,0300 с.	0,7030
	< 0,0400 с.	0,7990
	< 0,0500 с.	0,8675
	< 0,0600 с.	0,9148
	< 0,0700 с.	0,9402
	< 0,0800 с.	0,9600
	< 0,0900 с.	0,9716
	< 0,1000 с.	0,9802
	< 0,1100 с.	0,9866
	< 0,1200 с.	0,9911
	< 0,1300 с.	0,9944
	< 0,1400 с.	0,9968
	< 0,1500 с.	0,9981
	< 0,1600 с.	0,9993

Таблица 3
Вероятности того, что время реакции ЦОД
будет не больше Т для коэффициента
нагрузки 0,9

Коэффициент загрузки	Граница, Т	Вероятность
0,9	< 0,0400 с.	0,2295
	< 0,0800 с.	0,3942
	< 0,1200 с.	0,5194
	< 0,1600 с.	0,6093
	< 0,2000 с.	0,6929
	< 0,2400 с.	0,7566
	< 0,2800 с.	0,8193
	< 0,3200 с.	0,8633
	< 0,3600 с.	0,8977
	< 0,4000 с.	0,9230
	< 0,4400 с.	0,9423
	< 0,4800 с.	0,9610
	< 0,5200 с.	0,9739
	< 0,5600 с.	0,9824
	< 0,6000 с.	0,9877
	< 0,6400 с.	0,9913
	< 0,6800 с.	0,9932
	< 0,7200 с.	0,9953
	< 0,7600 с.	0,9978
	< 0,8000 с.	0,9992

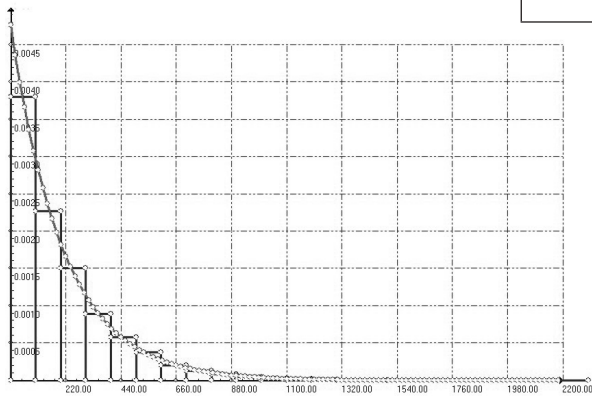


Рис. 2. Нагрузка 0,1. Гистограмма и экспоненциальное распределение.

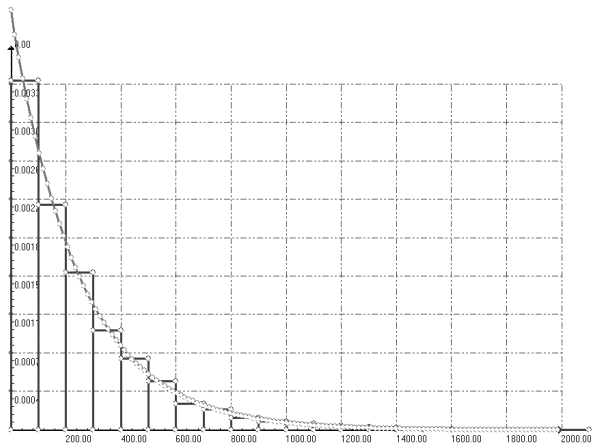
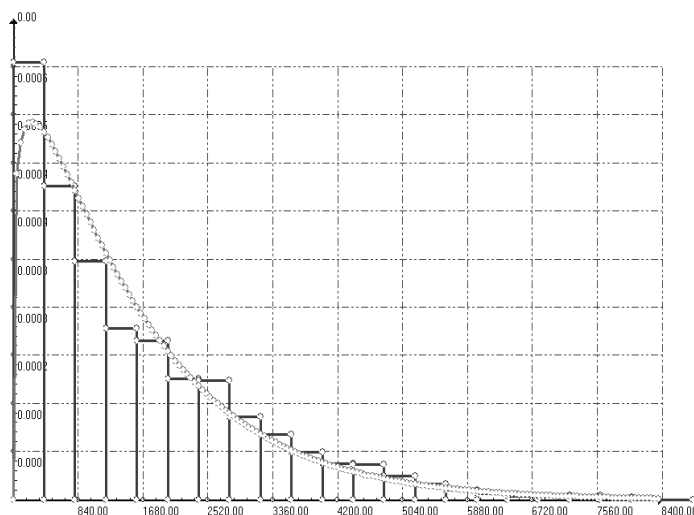


Рис. 3. Нагрузка 0,225537. Гистограмма и экспоненциальное распределение.





**Рис. 4. Нагрузка 0,9.
Гистограмма и гамма-
распределение.
Порядок
распределения
1,18370930, параметр
 $1/\mu = 1333,70355714$
единиц модельного
времени.**



указаны единицы модельного времени.

Моделирование ЦОД как СМО при уровне нагрузки 0,1 позволило получить вероятности того, что время реакции будет меньше некоторой величины T — таблица 1.

Результаты оптимизации режима ЦОД показали, что центр работает при коэффициенте нагрузки 0,225537 (таблица 2). Моделирование при 838 единицах модельного времени дало среднее время в системе 248,047, средне-квадратическое отклонение этой величины — 247,601. Таким образом, получаем распределение, близкое к экспоненциальному. Результаты в графическом виде представлены на рисунке 3, вероятности того, что время реакции будет меньше некоторой величины T в таблице 2.

При оптимальном режиме ЦОД время выполнения практически всех запросов не превысит 0,12 с.

Было проведено моделирование ЦОД при высокой нагрузке 0,9. Среднее время пребывания в системе составило 1578,7173 единиц модельного времени, средне-квадратическое отклонение — 1451,0482 единиц модельного времени. Результаты по вычислению вероятностей того, что время реакции не превысит некоторой границы, представлены в таблице 3.

График функции плотности распределения вероятностей времени реакции ЦОД при загрузке 0,9 представлен на рисунке 4.

Проверялась гипотеза о том, что случайная величина — время пребывания запроса

в СМО имеет гамма-распределение с плотностью:

$$f(t) = \mu (\mu t)^{v-1} e^{-\mu t} / \Gamma(v-1), t \geq 0, \quad (4)$$

где $\Gamma(v)$ — гамма-функция:

$$\Gamma(v) = \int_0^{\infty} e^{-t} t^{v-1} dt. \quad (5)$$

Для целых значений $\Gamma(v)$ есть простой факториал $(v-1)!$.

Проведенные исследования показали, что в случае использования мэйнфрейма с производительностью 5939 Mips время реакции ЦОД практически не превысит величины 0,64 с даже при очень большой нагрузке 0,9.

ЛИТЕРАТУРА

1. Ян Ван Бон, Георгес Кеммерлинг, Дик Пондман. Введение в ИТ-сервис-менеджмент: IT InfrastructureLibrary. Перевод на русский язык под ред. Поточного М. Ю., гл. редактор английской версии: Ян Ван Бон (Jan van Bon), 2003, 228 с. Электронный ресурс: [http://www.yessoft.ru/ITIL/ITIL-Introduction%20to%20itSMF \(rus\) -2003.pdf](http://www.yessoft.ru/ITIL/ITIL-Introduction%20to%20itSMF%20(rus)-2003.pdf). Дата доступа 20.06.2014.
2. Шерр А. Анализ вычислительных систем с разделением времени. — М.: Мир, 1970. — 135 с.
3. Сакман Г. Решение задач в системе человек—ЭВМ. — М.: Мир, 1973, 351 с.
4. Вентцель Е. С., Овчаров Л. А. Теория вероятностей и ее инженерные приложения. — М.: Высшая школа, 2000. — 480 с.
5. Вентцель Е. С. Теория вероятностей. — М.: Наука, Физматгиз, 1969. — 576 с.
6. Джейсуол Н. Очереди с приоритетами. — М.: Мир, 1973. — 280 с.
7. Вентцель Е. С. Исследование операций. — М.: Советское радио, 1972. — 552 с.
8. Кудрявцев Е. М. GPSS World. Основы имитационного моделирования различных систем. — М.: ДМК Пресс, 2004. — 320 с.