УДК 004.414.23

КОЛОМОЕЦ Г.П.

ИССЛЕДОВАНИЕ СТАТИСТИЧЕСКИХ ПАРАМЕТРОВ ПЕРЕДАЧИ ПАКЕТОВ С ДАННЫМИ HTTP

Выполнено исследование полученных в ходе эксперимента выборок межпакетных интервалов и длин пакетов с HTTP-запросами и HTTP-ответами с целью определения наиболее подходящих распределений плотности вероятности, которые могут быть использованы при моделировании процессов в компьютерных сетях.

Experimental research of samples of the interpacket intervals and lengths for HTTP-queries and HTTP-replies is carried out with the purpose of definition of the most approaching probability density distributions which can be used for processes simulation in computer networks.

В настоящее время достаточно распространенным является применение специализированных программных систем для моделирования процессов, происходящих при передаче данных в компьютерных сетях [1, 2]. При моделировании активно используются стандартные распределения вероятности, например, для задания времен между последовательно высылаемыми в сеть пакетами, размера пакетов и т.д. Вопросы, связанные с выбором определенного распределения вероятности в качестве входного при моделировании тех либо иных процессов, являются актуальными, поскольку непосредственно влияют на результаты моделирования.

Целью данной работы является определение наиболее подходящих распределений плотности вероятности для собранных в ходе эксперимента наборов межпакетных интервалов и длин пакетов при передаче данных в рамках HTTP-сеанса между клиентом и сервером.

В качестве клиента использовался компьютер с программой браузером, в качестве сервера – прокси-сервер, через который происходит передача всего внешнего по отношению к локальной сети HTTP-трафика. Клиент высылал запросы и получал ответы в течение примерно 124 секунд с 23-х HTTP-серверов Интернета. Следует отметить, что кроме указанного клиента в локальной сети находились и другие компьютеры, однако сетевую активность они не проявляли, за исключением относительно редкой широковещательной рассылки (доля служебных пакетов составила 6,78% от всех захваченных пакетов). В то же время исследуемый клиент проявлял активность, соответствующую обычной работе пользователя в Internet. Для захвата пакетов HTTP-клиентом и регистрации параметров сеанса использовалась свободно распространяемая программа анализатор сетевых протоколов Wireshark v. 0.99.6a [3].

Следует отметить, что из захваченных 1917 пакетов достаточно большое количество составляли служебные пакеты, устанавливающие и завершающие соединения на транспортном уровне. Поскольку в системах моделирования распределения входных параметров обычно задаются на прикладном уровне [1, 2], необходимо было отделить HTTP-запросы и ответы от служебных TCP-пакетов. Количественные характеристики захваченных пакетов приведены в табл. 1, а относительная частота HTTP-запросов и HTTP-ответов – на рис. 1. Количество ответов превышает количество запросов, поскольку если содержимое ответа превышает размер максимального переносимого кадром Ethernet блока в 1500 байт, высылаются несколько кадров с содержимым одного HTTP-ответа [4]. Количество служебных TCP-пакетов сравнимо с количеством HTTP-запросов и ответов.

Табл. 1. Количественные характеристики выборки

Общее время активности 2-х ПК, сек	123,94
Общее количество захваченных пакетов	1917
Количество пакетов только между клиентом и сервером	1787
Количество служебных пакетов	130
Количество HTTP запросов и ответов	887
Количество служебных TCP-пакетов	900
Количество HTTP-запросов	123
Количество HTTP-ответов	764

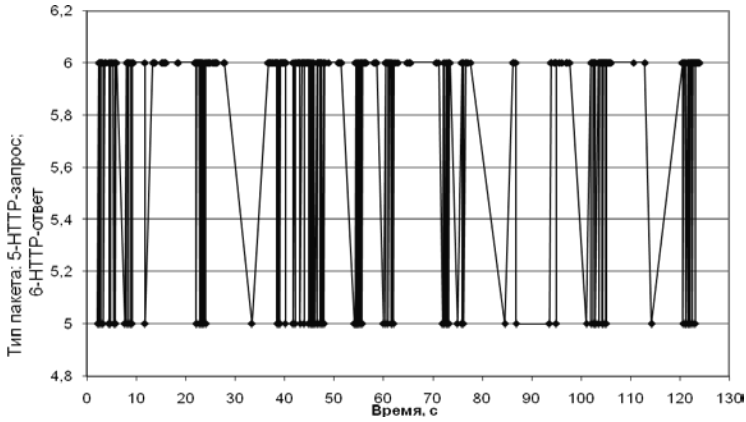


Рис. 1. Относительная частота HTTP-запросов и HTTP-ответов (без служебных TCP-пакетов)

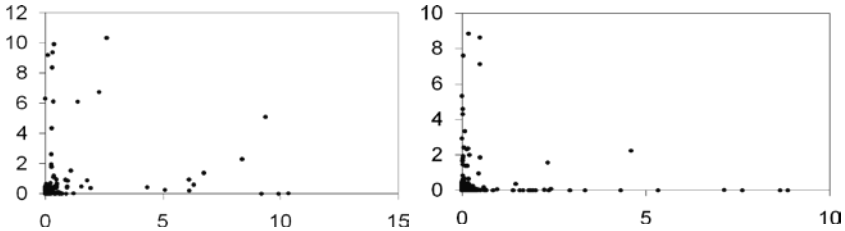
Значения статистических параметров для экспериментальных наборов времен между HTTP-запросами и времен между HTTP-ответами, а также для длин HTTP-запросов и HTTP-ответов приведены в табл. 2 (времена измеряются в секундах, длины пакетов – в байтах).

Табл. 2. Статистические параметры экспериментальных данных

Параметр	Для времен между HTTP-запросами	Для времен между HTTP-ответами	Для длин HTTP-запросов	Для длин HTTP-ответов
Количество испытаний	122	763	123	764
Минимальное значение	0,000048	0,000009	60	60
Максимальное значение	10,323789	8,849985	705	1514
Среднее значение	0,992236	0,159238	430,7724	1230,556
Медиана	0,265827	0,020460	459	1502
Дисперсия	4,703899	0,510741	11272,91	237494,709
Коэффициент вариации	2,185818	4,488008	0,24647	0,39603
Асимметрия	3,066372	8,812262	-1,03401	-1,4639203

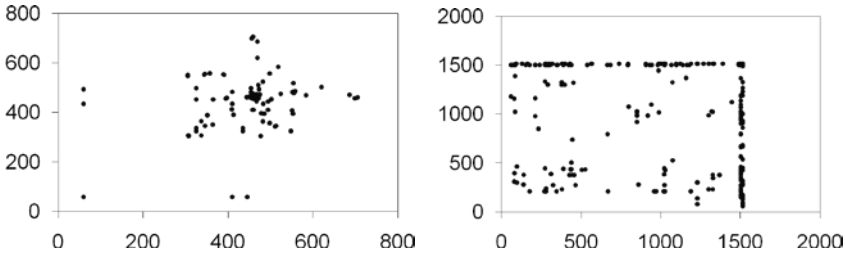
По значениям статистических параметров можно выдвинуть предположения относительно семейства наиболее подходящих распределений плотности вероятности. Так, близость значений среднего и медианы для длин HTTP-запросов и HTTP-ответов может указывать на симметричное распределение (например, нормальное). Учитывая, что для экспоненциального распределения коэффициент вариации равен 1 для любых значений масштабного параметра β, можно предположить, что вероятность ни одной из наблюдаемых величин не подчиняется этому распределению [5]. В то же время именно экспоненциальное распределение рекомендуется в качестве входного при моделировании времен между поступлениями запросов в систему, происходящими с постоянной интенсивностью [6]. Основываясь на полученных значениях асимметрии, можно предположить, что времена между HTTP-запросами и времена между HTTP-ответами имеют смещенную вправо, а их длины – смещенную влево форму распределения.

Поскольку формальное определение соответствия стандартных распределений плотности вероятности экспериментальным наборам данных предполагает выборочную независимость данных, необходимо было определить степень такой независимости [6]. Для этого были построены диаграммы разброса наблюдений для пар времен между HTTP-запросами и пар времен между HTTP-ответами (Рис. 2), а также для пар длин HTTP-запросов и пар длин HTTP-ответов (Рис. 3).



а) б)

Рис. 2. Диаграмма разброса наблюдений для пар времен между HTTP-запросами (а) и пар времен между HTTP-ответами (б)



а) б)

Рис. 3. Диаграмма разброса наблюдений для пар длин HTTP-запросов (а) и пар длин HTTP-ответов (б)

Характер разброса для межпакетных времен позволяет заключить, что выборочные данные обладают определенной степенью независимости, хотя, судя по графикам, типы распределений для времен между HTTP-запросами и времен между HTTP-ответами должны отличаться. Выборочные данные для длин HTTP-запросов также указывают на их относительную независимость, однако длины HTTP-ответов назвать таковыми нельзя, поскольку среди ответов преобладают пакеты максимального размера в 1514 байт (133 значения) и близкого к нему в 1502 байта (404 значения) (их суммарная доля составляет 70%). Этот факт отражает работу драйвера сетевого адаптера прокси-сервера, пытающегося передать в кадрах канального уровня пакет максимально возможного размера с целью уменьшения накладных расходов на передачу заголовков пакетов. Если же не учитывать длины пакетов в 1502 и 1514 байт, то диаграмма разброса для оставшихся длин HTTP-ответов демонстрирует (Рис. 4) достаточную степень независимости значений длин, очевидно связанную со случайным значением длины передаваемых с WWW-серверов файлов.

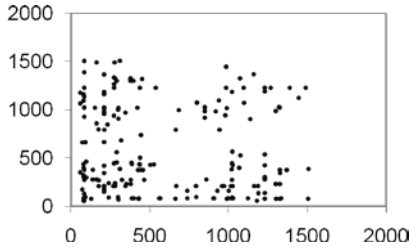


Рис. 4. Диаграмма разброса наблюдений для пар длин HTTP-ответов без учета длин пакетов в 1502 и 1514 байт

Анализ полученных выборок HTTP-запросов клиента к серверу, HTTP-ответов сервера клиенту, а также выборок длин пакетов с HTTP-запросами и длин пакетов с HTTP-ответами на предмет соответствия стандартным непрерывным распределениям плотности вероятности выполнялся в системе статистического анализа данных Statistica 6.0 разработки StatSoft Inc. [7].

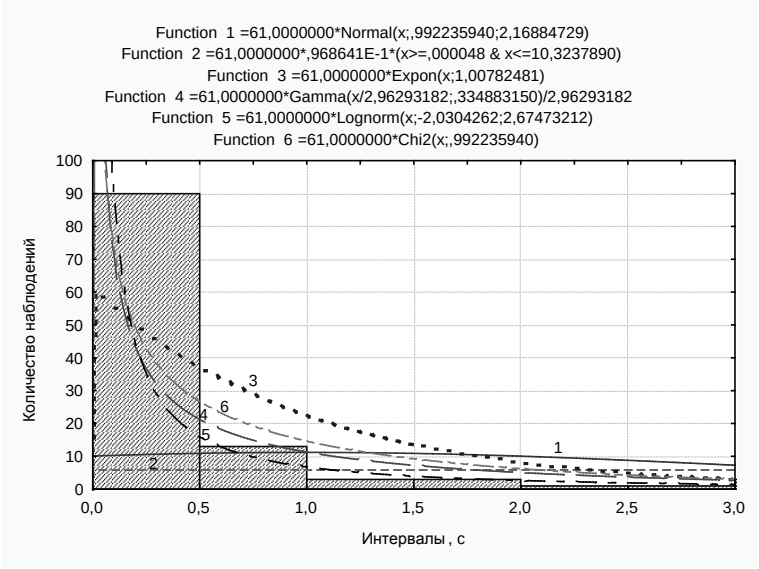


Рис. 5. Гистограмма интервалов времени между HTTP-запросами и стандартные плотности распределения вероятности

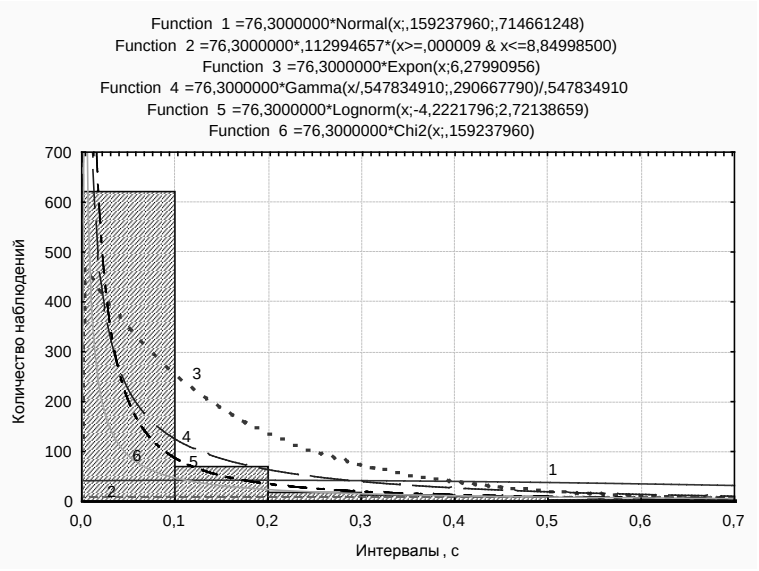


Рис. 6. Гистограмма интервалов времени между HTTP-ответами и стандартные плотности распределения вероятности

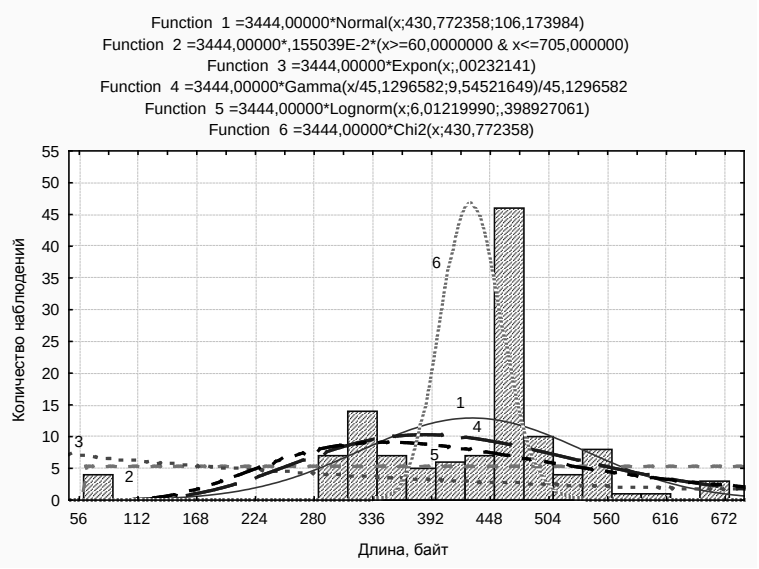


Рис. 7. Гистограмма длин HTTP-запросов и стандартные плотности распределения вероятности

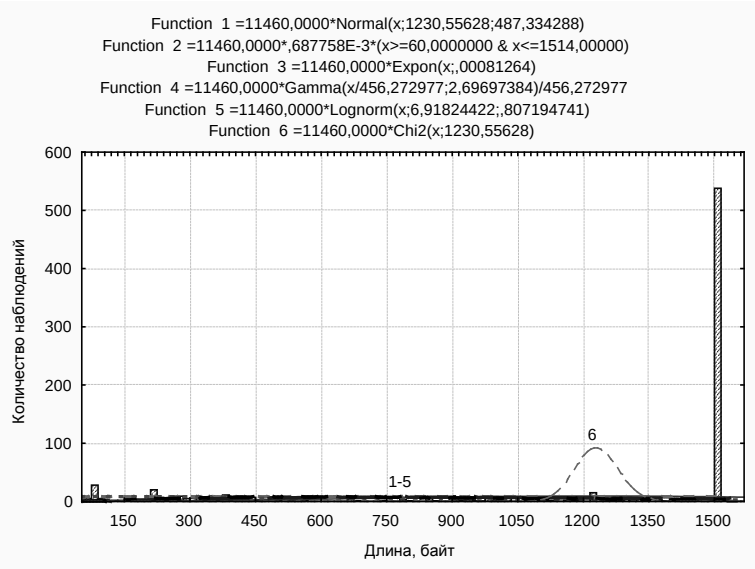


Рис. 8. Гистограмма длин HTTP-ответов и стандартные плотности распределения вероятности:
1 - нормальное, 2 -равномерное, 3 - экспоненциальное,
4 - гамма, 5 - логнормальное, 6 - Хи-квадратичное

Гистограммы времен между HTTP-запросами и времен между HTTP-ответами, длин HTTP-запросов и длин HTTP-ответов, а также кривые доступных в программе плотностей вероятности нормального, равномерного, экспоненциального, гамма, логнормального и Хи-квадратичного распределений приведены на рис. 5 – рис. 8. Согласие стандартной плотности вероятности с экспериментальными значениями оценивалось с помощью критерия "Хи-квадрат" и критерия Колмогорова-Смирнова [6]. Кроме численных значений критериев согласия, естественно, оценивалось совпадение форм плотностей вероятности и гистограмм. Учитывая, что проверка по критерию "Хи-квадрат" чувствительна к выбору количества и размеров интервалов, используемых при построении гистограмм, эти параметры выбирались в соответствии с рекомендациями, состоящими в выборе интервалов одинакового размера и их количества такого, чтобы произведение доли единичного интервала на количество данных в выборке было не меньше 5 [6]. Этим рекомендациям удовлетворяют 22 интервала для 122 значений времен между HTTP-запросами (122/22≈5,55) и 100 интервалов для 763 значений времен между HTTP-ответами (763/100=7,63). При построении гистограмм для длин пакетов по тем же соображениям были выбраны 22 интервала для 123 длин HTTP-запросов и 100 интервалов для 764 длин HTTP-ответов.

Табл. 3. Результаты тестов на соответствие стандартным распределениям

Распределение	Тест Хи-квадрат				Тест Колмогорова-Смирнова	
	χ^2	df	p	χ^2(df, 0,99)	d	$(n^{1/2}+0,12+0,11/ n^{1/2})\cdot d$
1	2	3	4	5	6	7
Параметры для времен между HTTP-запросами (22 интервала)						
Нормальное	68,91287	6	0,00000	16,812	0,35518	3,97
Равномерное	1284,3637	17	0,00000	33,409	0,75247	8,41
Экспоненциальное	67,69580	4	0,00000	13,277	0,34647	3,88
Гамма	21,52803	4	0,00025	13,277	0,15209	1,70
Логнормальное	15,03783	4	0,00462	13,277	0,16294	1,82
Хи-квадратичное	32,00457	5	0,00001	15,086	0,21599	2,41
Параметры для времен между HTTP-ответами (100 интервалов)						
Нормальное	485,35127	15	0,00000	30,578	0,41184	11,43
Равномерное	44614,687	85	0,00000	≈120	0,90405	25,08
Экспоненциальное	420,18642	6	0,00000	16,812	0,54846	15,22
Гамма	142,73452	10	0,00000	23,209	0,24999	6,94
Логнормальное	42,51791	14	0,00010	29,141	0,31854	8,84
Хи-квадратичное	68,27119	13	0,00000	27,688	0,50593	14,04
Параметры для распределений длин HTTP-запросов (22 интервала)						
Нормальное	118,32683	9	0,00000	21,666	0,18899	2,121
Равномерное	386,86109	19	0,00000	36,191	0,34732	3,897
Экспоненциальное	533,26119	11	0,00000	24,725	0,47487	5,328
Гамма	192,89382	11	0,00000	24,725	0,21137	2,371
Логнормальное	256,61483	12	0,00000	26,217	0,21329	2,393
Хи-квадратичное	187,46842	3	0,00000	11,345	0,37837	4,246
Параметры для распределений длин HTTP-ответов (100 интервалов)						
Нормальное	36207,672	70	0,00000	≈100	0,41542	11,53
Равномерное	38862,774	94	0,00000	≈125	0,69594	19,32
Экспоненциальное	52503,057	72	0,00000	≈102	0,40913	11,36
Гамма	31873,700	80	0,00000	≈107	0,40949	11,37
Логнормальное	43231,961	75	0,00000	106,393	0,39246	10,90
Хи-квадратичное	43285,040	15	0,00000	30,578	0,70839	19,67

В табл. 3 приведены значения параметров теста Хи-квадрат и теста Колмогорова-Смирнова для сравниваемых теоретических распределений и экспериментальных данных. Здесь χ^2 – значение подобранного параметра теста, df – количество степеней свободы, p – относительная величина достоверности используемого распределения, χ^2(df, 0,99) – табличное значение χ^2 для указанного числа степеней свободы, не превышение которого позволяет заключить о достоверности используемого приближения на уровне 0,99 (минимально возможный уровень достоверности), d – наибольшее вертикальное расстояние между теоретической и экспериментальной функциями распределения, в последнем столбце приведена выверенная статистика критерия Колмогорова-Смирнова (минимальным значением этой статистики для выбора с достоверностью на уровне 0,99 является 1,628), n – количество испытаний случайной величины [6].

Как видно из таблицы 3, в соответствии с результатами обоих тестов ни одно из исследованных распределений не может быть выбрано для описания экспериментальных данных даже с минимально возможной достоверностью, однако наилучшими кандидатами для задания времен между HTTP-запросами могут быть гамма или логнормальное распределение. Для определения возможных кандидатов для моделирования межпакетных интервалов HTTP-ответов необходимо исследование других стандартных распределений плотности вероятности. Что касается длин пакетов, то для них также ни одно из исследованных распределений формально не может быть выбрано. Мы удалили детерминированную составляющую, соответствующую максимальным длинам HTTP-ответов в 1502 и 1514 байт и повторили анализ. Соответствующая гистограмма и графики исследованных стандартных распределений плотностей вероятности приведены на рис.9. В этом случае выборка составила 228 наблюдений, распределенных между 60 и 1507 байтами, и при построении гистограммы было выбрано 30 интервалов (228/30=7,6).

В табл. 4 приведен анализ соответствия экспериментального набора данных исследованным стандартным распределениям плотности вероятности. И хотя значения статистики Колмогорова-Смирнова в этом случае находятся ближе к граничному критерию (для экспоненциального и гамма распределений), позволяющему осуществить выбор стандартного распределения, все же формально мы этого сделать не можем. Учитывая полученные результаты, можно рекомендовать в качестве входного распределения плотности вероятности HTTP-ответов выбирать экспоненциальное или гамма распределение для длин 30% пакетов вместе с фиксированной длиной в 1502 (или 1514) байт для 70% пакетов.

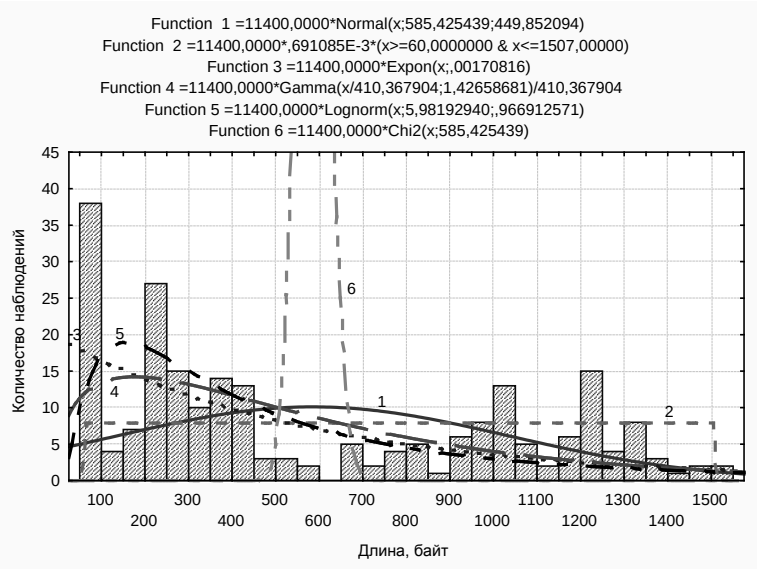


Рис. 9. Гистограмма длин HTTP-ответов, не включающих пакеты длиной 1502 и 1514 байт и стандартные плотности распределения вероятности: 1 - нормальное, 2 -равномерное, 3 - экспоненциальное, 4 - гамма, 5 - логнормальное, 6 - Хи-квадратичное

Табл. 4. Результаты тестов на соответствие стандартных распределений и экспериментальных наборов длин HTTP-ответов без учета HTTP-ответов длиной 1502 и 1514 байт

Распределение	**Тест Хи-квадрат**				**Тест Колмогорова-Смирнова**	
	χ^2	**df**	**p**	**χ^2(df, 0,99)**	**d**	**$(n^{1/2}+0,12+0,11/n^{1/2})\cdot d$**
Параметры для распределений длин HTTP-ответов без 1502 и 1514 байт (30 интервалов)						
Нормальное	145,71012	22	0,00000	40,289	0,18394	2,801
Равномерное	285,43559	26	0,00000	45,642	0,29581	4,504
Экспоненциальное	154,89075	20	0,00000	37,566	0,12115	1,845
Гамма	154,26182	19	0,00000	36,191	0,12819	1,952
Логнормальное	211,13851	17	0,00000	33,409	0,14033	2,137
Хи-квадратичное	1429,8538	2	0,00000	9,210	0,57448	8,748

Выводы

При моделировании передачи пакетов в сети следует учитывать достаточно большое количество служебных пакетов, которые добавляются к запросам и ответам пользовательских приложений, поэтому задание входных распределений статистических параметров передачи должно выполняться для разных уровней сетевого взаимодействия и в общем случае должно отличаться.

Наилучшими кандидатами из исследованных стандартных распределений для выбора в качестве входных распределений плотности вероятности при моделировании интервалов между HTTP-запросами являются гамма и логнормальное распределение. Для определения возможных кандидатов для моделирования временных интервалов между HTTP-ответами необходимо дополнительное исследование.

Ни одно из исследованных распределений не подходит для моделирования распределения длин HTTP-запросов и HTTP-ответов, что связано, очевидно, с достаточно большой детерминированной составляющей в них. Эта составляющая возникает вследствие определенного количества полей в заголовках пакетов с относительно мало варьирующейся длиной для HTTP-запросов и с стремлением драйвера сетевого адаптера передавать HTTP-ответы в кадрах максимально возможного размера с целью уменьшения доли накладных расходов.

Анализ выборки длин HTTP-ответов без учета кадров максимального размера показывает, что возможным решением для выбора входных распределений является использование экспоненциального или гамма распределения для длин 30% пакетов вместе с фиксированной длиной в 1502 (или 1514) байт для 70% пакетов.

Список использованной литературы

9. Система моделирования компьютерных сетей OPNET Modeler – OPNET Technologies,Inc. - http://www.opnet.com/solutions/network_rd/modeler.html.
10. Система имитационного моделирования компьютерных сетей OMNET++ – http://www.omnetpp.org/index.php.
11. Анализатор сетевых протоколов Wireshark – http://www.wireshark.org.
12. Таненбаум Э. Компьютерные сети. 4-е изд. – СПб.: Питер, 2003. - 992 с.
13. Гмурман В.Е. Теория вероятностей и математическая статистика. Учеб. пособие для вузов.-Изд. 7-е, стер.-М.: Высш. шк., 2001.-479с.
14. Лоу А. М., Кельтон В. Д. Имитационное моделирование. Классика CS. 3-е изд. - СПб.: Питер; Киев: Издательская группа BHV, 2004. – 847 с.
15. Система статистического анализа данных Statistica v.6.- StatSoft, Inc. – http://www.statsoft.com.