Труды Карельского научного центра РАН
№ 5. 2011. С. 75–85

УДК 519.872.8: 004.382.2

МОДЕЛИ МНОГОСЕРВЕРНЫХ СИСТЕМ ДЛЯ АНАЛИЗА ВЫЧИСЛИТЕЛЬНОГО КЛАСТЕРА

Е. В. Морозов, А. С. Румянцев

Институт прикладных математических исследований Карельского научного центра РАН

Работа посвящена изучению свойств процесса загрузки в многосерверных системах обслуживания, в том числе при наличии тяжелых хвостов. Проведен анализ моментных свойств процесса загрузки. Предложена модель вычислительного кластера как стохастической многосерверной системы обслуживания. Подробно обсуждается связь модели кластера с классическими моделями многосерверных систем. Приведены результаты статистических экспериментов, которые показывают хорошее согласие модели с работой кластера. Работа поддержана РФФИ, проект 07-10-00017.

Ключевые слова: вычислительный кластер, многосерверные системы, стационарность, имитационное моделирование, распределения с тяжелым хвостом.

E. V. Morozov, A. S. Rumyantsev. MULTI-SERVER MODELS TO ANALYZE HIGH PERFORMANCE CLUSTER

The paper is dedicated to the properties of the workload process in multi-server systems, i.a. in the presence of heavy tails. The moment properties of the workload process are analyzed. A model of a high performance cluster as a multi-server stochastic system is suggested. A connection between the model of the cluster and the classical multi-server system models is discussed in detail. The results of numerical experiments are provided, demonstrating an agreement between the model and the cluster performance.

Key words: high performance cluster, multi-server systems, stability, simulation, heavy-tailed distributions.

Введение

Анализу стохастических моделей многосерверных систем обслуживания посвящено большое число работ. Теоретическое изучение мотивировано запросами производителей вычислительных мощностей, так как в последнее десятилетие имеет место повсеместная тенденция внедрения многоядерных и многопроцессорных архитектур. В то же время точный анализ таких моделей затруднен и аналитические результаты доступны только для узкого класса простейших моделей.

С точки зрения практики, исследователя могут интересовать характеристики системы, влияющие на качество обслуживания (QoS), например, величина средней задержки в системе, величина дисперсии задержки. Важной проблемой является определение моментных свойств сетевых процессов, в частности, так называемого моментного индекса, т. е. максимального конечного момента. Другим важным аспектом анализа является зависимость моментных свойств сетевых процессов от

дисциплины обслуживания клиентов [Borst et al., 2002]. Особенно сложен анализ в случае, когда заявки в систему обслуживания поступают в виде пачек (см., напр., работы [Keilson, Seidmann, 1987; Wolff, 1991; Liu, 1993]), или в виде регенеративного процесса [Huang, Sigman, 1999]. Вычислительные кластеры и Грид представляют в этом смысле особую сложность, так как обладают вышеперечисленными свойствами, такими как многопроцессорность, сложные дисциплины обслуживания потока заявок, возможность поступления пачек заявок, необходимость в больших временах обслуживания на процессорах (тяжелые хвосты).

Теоретический интерес последних лет к исследованию моделей систем обслуживания с распределением с тяжелыми хвостами связан с обнаружением этого эффекта в траффиках компьютерных сетей (см., напр., [Crovella, Bestavos, 2002]). Тяжелый хвост означает более медленное, чем экспоненциальное, убывание хвостовой вероятности $\overline{F}(x) := P(X > x)$ случайной величины X, т. е.

$$\lim_{x\to\infty} e^{-ax}\overline{F}(x) = \infty, \quad \forall a > 0.$$

Это свойство радикально отличается от свойства экспоненциального распределения, попрежнему широко используемого для моделирования процессов в современных компьютерных системах. Присутствие тяжелых хвостов эмпирически обнаружено в распределениях размеров файлов на жестких дисках, времен вычисления задач на процессоре, времен передачи файлов между сервером и клиентом в высокоскоростных сетях и т. д. [Crovella, Bestavos, 2002]. Важный подкласс тяжелохвостых распределений составляют *субэкспоненциальные распределения*, которые (для неотрицательных с.в.) определяются как

$$\lim_{x\to\infty} \frac{P(X_1 + X_2 > x)}{P(X > x)} = 2,\ x \geqslant 0. \qquad (1)$$

Другой важный подкласс составляют *правильно меняющиеся распределения*, характеризуемые соотношением

$$\lim_{x\to\infty} \frac{\overline{F}(xt)}{\overline{F}(x)} = t^{-\alpha},\ \alpha \geqslant 0.$$

(При $\alpha = 0$ функция называется *медленно меняющейся*). Наиболее часто используемым распределением такого рода является распределение Парето, имеющее (стандартный) вид

$$F(x) = 1 - x^{-\alpha},\ x \geqslant 1. \qquad (2)$$

В то же время, обнаружение тяжелых хвостов в компьютерных системах обозначило ряд проблем, недооценивание которых приводит к серьезным последствиям как для компаний-производителей сетевого оборудования, так и для потребителей. (Частичный анализ такого рода ошибок приведен в работе [Morozov et al., 2008].) Одна из причин возникающих проблем состоит в том, что многие алгоритмы управления сетевыми устройствами и процессами в вычислительных системах строились с учетом предположения об экспоненциальности соответствующих распределений [Harchol-Balter, 1999]. Все это требует тщательного анализа многосерверных систем обслуживания и построения подходящих моделей.

Модели многосерверных систем обслуживания

В этом разделе дан обзор ряда важных известных результатов для многосерверных систем обслуживания, в том числе с бесконечным числом серверов. Эти результаты в основном опираются на классические результаты для односерверных систем, которые также приводятся для сравнения.

Система G/G/s FIFO

Пусть в систему обслуживания в моменты времени $\{t_n, n \geqslant 1\}$, образующие процесс восстановления, поступают заявки с независимыми, одинаково распределенными временами обслуживания S_n на одном из s идентичных серверов (процессоров). Обозначим функцию распределения времени между приходами заявок $T_n := t_{n+1} - t_n$ через $A(x) = P(T \leqslant x)$, соответственно $B(x) = P(S \leqslant x)$ — распределение времени обслуживания заявки (индекс опущен, когда рассматривается типичный представитель последовательности). Заявка ожидает время $D_n \geqslant 0$ в очереди, формируемой в порядке поступления (FIFO), и поступает на обслуживание в наименее загруженный сервер. Вектор загрузки W_n, состоящий из оставшегося времени обслуживания на каждом процессоре в момент прихода заявки n, удовлетворяет рекурсии Кифера-Вольфовица [Scheller-Wolf, Sigman, 1997a]):

$$W_{n+1} = R \begin{pmatrix} W_n(1) + S_n - T_n \\ W_n(2) - T_n \\ \dots \\ W_n(s) - T_n \end{pmatrix}^+, \qquad (3)$$

где оператор $R(\cdot)$ располагает компоненты в возрастающем порядке, $W_n(1) \leqslant \ldots \leqslant W_n(s)$, а $(x)^+ = \max(0, x)$. Таким образом, время

ожидания n-й заявки $D_n := W_n(1)$. При условии $\rho := ES/ET < s$ имеет место слабая сходимость вектора загрузки к стационарному значению, $W_n \Longrightarrow W := (W(1), \ldots, W(s))$.

Обозначим через $U_n = W_n(2) - W_n(1)$ время, оставшееся на втором наименее загруженном сервере в тот момент, когда заявка n поступит на обслуживание. Тогда первый из этих двух серверов освободится через время $P_n = \min(U_n, S_n)$, и время ожидания D_n удовлетворяет модифицированной рекурсии Линдли [Scheller-Wolf, Sigman, 1997a]:

$$D_{n+1} = (D_n + P_n - T_n)^+, \quad n \geqslant 1. \qquad (4)$$

Заметим, что при $s = 1$ формула (4) определяет классическую рекурсию Линдли

$$D_{n+1} = (D_n + S_n - T_n)^+, \quad n \geqslant 1.$$

Условия конечности моментов компонент вектора загрузки

Следующее необходимое и достаточное условие конечности моментов времени ожидания в односерверной системе G/G/1 получено в работе [Kiefer, Wolfowitz, 1956]: при условии $ET < \infty$,

$$ES^{\alpha+1} < \infty \quad \Leftrightarrow \quad ED^{\alpha} < \infty. \qquad (5)$$

Обобщение результата (5) для вектора загрузки W в s-серверной системе было получено в работе [Scheller-Wolf, Vesilo, 2011]. Так, при условии устойчивости $\rho = ES/ET < s$, для всех компонент вектора с индексами $i \leqslant \lceil \rho \rceil$ (наименьшее целое большее ρ) имеют место *одни и те же* достаточные условия конечности моментов порядка $\alpha \geqslant 1$:

$$E\left(S^{1+\alpha/(s-\lfloor \rho \rfloor)}\right) < \infty \quad \Rightarrow \quad E\left[W(i)\right]^{\alpha} < \infty.$$

В то же время для компонент с индексами $i > \lceil \rho \rceil$ моментные свойства зависят от индекса компоненты $W(i)$:

$$E\left(S^{1+\alpha/(s-i)}\right) < \infty \quad \Rightarrow \quad E\left[W(i)\right]^{\alpha} < \infty.$$

Частным случаем этого результата является достаточное условие конечности момента порядка α стационарного времени ожидания $D = W(1)$ в системе G/G/s FIFO, полученное ранее в работе [Scheller-Wolf, Vesilo, 2006]:

$$E\left(S^{1+\alpha/(s-\lfloor \rho \rfloor)}\right) < \infty \;\Rightarrow\; E(D^{\alpha}) < \infty. \qquad (6)$$

При дополнительных ограничениях на распределение B, приведенные выше условия являются также необходимыми [Scheller-Wolf, Vesilo, 2011].

Таким образом, имеет место зависимость моментов компонент вектора W как от индекса i, так и от коэффициента загрузки ρ. Фактически имеет место *взаимопомощь* среди первых $\lceil \rho \rceil$ наименее загруженных серверов, моментные свойства которых лучше. Это объясняется *запасом мощности*, состоящим из $s - \lceil \rho \rceil$ серверов, которые могут быть удалены из системы (или заняты обслуживанием очень длинных заявок) без потери устойчивости оставшейся части системы [Scheller-Wolf, Vesilo, 2011].

Частный случай результата (6) при малой загрузке $\rho < 1$ в системе G/G/s был получен в работе [Scheller-Wolf, 2000]:

$$E\left(S^{1+1/s}\right) < \infty \quad \Rightarrow \quad ED < \infty.$$

В пределе при $s \to \infty$ отсюда следует классическое условие $ES < \infty$, влекущее конечность среднего стационарного времени пребывания заявки в системе G/G/∞. В работе [Scheller-Wolf, Sigman, 1997b] приведена также верхняя граница для стационарной средней задержки в системе G/G/s.

На практике формула (6) позволяет обосновать выбор между s *медленными* серверами (работающими со скоростью $1/s$) и одним *быстрым* (работающим со скоростью 1) в случае, когда распределение времени обслуживания имеет тяжелый хвост. Действительно, при числе серверов $s > 1/(1-\rho)$, одном и том же входном потоке и временах обслуживания $\tilde{S}$ в s-серверной системе, пропорционально увеличенных в соответствии со скоростями серверов (т. е. $E\tilde{S} = s \cdot ES$), в s-серверной системе задержки будут иметь конечный момент более высокого порядка, чем в системе $G/G/1$. Для доказательства переформулируем соотношение (6), обозначив $\beta = 1 + \alpha/(s - \lfloor \rho \rfloor)$. Получим

$$E(\tilde{S}^{\beta}) < \infty \Rightarrow E\left(D^{(s-\lfloor \tilde{\rho} \rfloor)(\beta-1)}\right) < \infty. \qquad (7)$$

Для односерверной системы имеем

$$ES^{\beta} < \infty \;\text{ влечет }\; ED^{\beta-1} < \infty.$$

Коэффициент загрузки в s-серверной системе $\tilde{\rho} = E\tilde{S}/ET = sES/ET < s$, т. е. система устойчива. С учетом выбора $s > 1/(1-\rho)$ получаем $s - s\rho > 1$, и поэтому $s - \lfloor \tilde{\rho} \rfloor = s - \lfloor s\rho \rfloor = \lfloor s - s\rho \rfloor \geqslant 1$.

Таким образом, если времена обслуживания имеют бесконечный момент порядка β (например, если они распределены по закону Парето (2) с параметром $\beta > 0$), то целесообраз-

нее использовать несколько медленных серверов, нежели один быстрый [Scheller-Wolf, Vesilo, 2006].

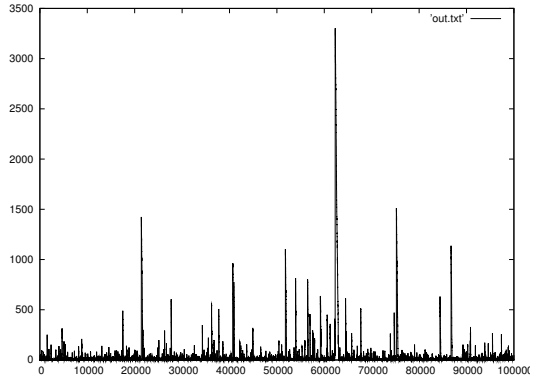

Рис. 1. Задержки в системе M/G/1, $\rho = 0,3$

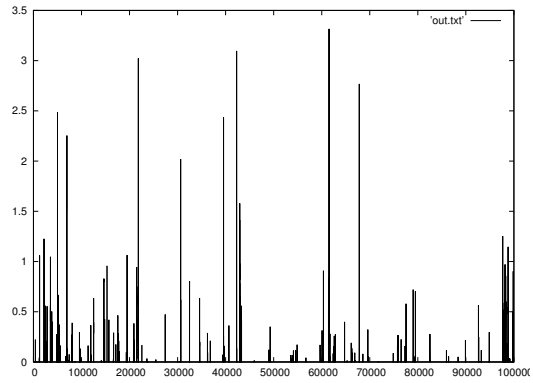

Рис. 2. Задержки в системе M/G/8, $\rho = 2,4$

Рассмотрим численный пример. Пусть система с процессором частотой 3 ГГц обслуживает процессы, времена выполнения которых имеют тяжелый хвост, $\overline{B}(x) = x^{-1,5}$. Пусть загрузка системы равна $\rho = 0,3$. Согласно формуле (5), стационарное время ожидания в системе имеет конечный момент только лишь порядка $0,5$, $ED^{0,5} < \infty$, т. е. бесконечное среднее. Заменим исходную систему на двухпроцессорную, где каждый процессор имеет частоту 1,5 ГГц. Общая загрузка системы возрастет до $0,6$. В то же время согласно (7), $ED^{2\cdot 0,5} = ED < \infty$ (поскольку $\lfloor\tilde{\rho}\rfloor = 0$), и средняя задержка конечна. Если же заменить систему на восьмипроцессорную, то $\lfloor\tilde{\rho}\rfloor = 2$. Поэтому $ED^3 < \infty$ и задержка имеет конечный третий момент. Иллюстрацией к численному примеру служат рисунки 1, 2, где приведены величины задержек в односерверной и восьмисерверной системах, соответственно, для выборки 100 000 заявок (задержки указаны по оси ординат).

Двухсерверная система: случай тяжелого хвоста

Классический результат (5) имеет интуитивное объяснение. Определяющим параметром для времени ожидания в системе является оставшееся время обслуживания S_I заявки n в момент прихода заявки $n+1$. Известно, что стационарное незавершенное время обслуживания имеет распределение

$$\overline{B}_I(x) = \frac{1}{ES}\int_x^\infty \overline{B}(y)\,dy,\ x \geqslant 0.$$

При этом условие $ES_I^{k-1} < \infty$ выполнено (при некотором $k > 1$), если $ES^k < \infty$, т. е. моментные свойства S_I хуже, чем у S.

В работе [Foss, Korshunov, 2006] приведены асимптотические результаты для хвоста распределения стационарной задержки D в двухсерверной системе $G/G/2$ при условии, что незавершенное время обслуживания имеет субэкспоненциальное распределение (1). Так, в случае *максимальной устойчивости*, т. е. при $\rho < 1$,

$$C_1 \leqslant \liminf_{x\to\infty} \frac{P(D>x)}{(\overline{B}_I(x))^2}$$

$$\leqslant \limsup_{x\to\infty} \frac{P(D>x)}{(\overline{B}_I(x))^2} \leqslant C_2.$$

В частности, в случае правильно меняющейся функции $\overline{B}(x)$ имеет место асимптотика

$$P(D>x) \sim C(\overline{B}_I(x))^2,\ x\to\infty,$$

где постоянная C также зависит от ES, ET [Foss, Korshunov, 2006]. В случае *минимальной устойчивости*, $1 < \rho < 2$, для правильно меняющейся функции распределения B имеет место асимптотика

$$P(D>x) \sim C\overline{B}_I(\frac{ES}{ES-ET}x),\ x\to\infty.$$

Заметим, что постоянные C_1, C_2, C выше зависит от параметров ES, ET.

Модель вычислительного кластера

Вычислительный кластер – это система, состоящая из нескольких узлов, объединенных быстрой коммуникационной сетью и предоставляющая пользователю вычислительные ресурсы. Для построения модели кластера сделаем следующие предположения:

1. поток заявок представляет собой процесс восстановления (времена между приходами независимы и одинаково распределены);

2. разделяемым ресурсом является количество процессоров и время вычисления; времена вычисления заявок и требуемое каждой заявке число процессоров независимы;

3. заявки поступают в буфер неограниченной емкости и обслуживаются в порядке прихода (FIFO);

4. заявка поступает на обслуживание только в момент освобождения требуемого ей числа процессоров.

Теоретический анализ

В дополнение к обозначениям предыдущей секции, пусть N_n — требуемое число процессоров для заявки n, на каждом из которых она будет выполняться в течение S_n единиц времени. Предполагается, что (целочисленные) случайные величины $\{N_n\}$ независимы и одинаково распределены. Очевидно, что $1 \leqslant N_n \leqslant s$. Тогда вектор (упорядоченных по возрастанию) времен ожидания на серверах удовлетворяет следующей модифицированной рекурсии Кифера-Вольфовица

$$W_{n+1} = R \begin{pmatrix} W_n(N_n) + S_n - T_n \\ \dots \\ W_n(N_n) + S_n - T_n \\ W_n(N_n + 1) - T_n \\ \dots \\ W_n(s) - T_n \end{pmatrix}^+ . \qquad (8)$$

Действительно, n-я заявка дожидается в буфере (в очереди) освобождения всех N_n требуемых ей процессоров. Поэтому загруженность всех серверов с номерами $[1, N_n]$ становится равной $W_n(N_n)$, поскольку, по крайней мере, часть этих серверов будет простаивать до освобождения наиболее загруженного из них, и в это время не смогут принять другие заявки. Это объясняет равенство первых N_n строк в формуле (8). Отметим, что на практике такая неэффективная дисциплина используется редко. Более эффективным является алгоритм Backfill, позволяющий заполнять промежутки простоя отдельных процессоров заявками без существенного нарушения приоритетов и очередности.

Заметим, что в худшем случае при $N_n = s$ все компоненты вектора загрузки станут равными, поэтому при условии $W_n(s) + S_n < T_n$ заявка с номером $n + 1$ начнет обслуживание в полностью пустой системе (все процессоры будут освобождены с уходом заявки n). Поскольку входной поток является процессом восстановления, то приходы таких заявок являются *моментами регенерации* системы. Это открывает возможность применения регенеративного моделирования для оценивания стационарных характеристик системы (см., напр., [Morozov, Rumyantsev, 2010]). Отметим, что суммарное время до освобождения всех необходимых n-й заявке процессоров (в момент ее прихода) равно

$$N_n \cdot (W_n(N_n) + S_n - T_n)^+ + \sum_{k=N_n+1}^{s} (W_n(k) - T_n)^+.$$

Эта величина является ключевой для определения условий стационарности рассматриваемой системы. Хотя такой анализ пока провести не удалось, проведенные эксперименты позволяют предположить, что если значения N_n распределены равномерно в интервале $[1, s]$, то условие стационарности имеет вид $\rho = ES/ET < 1$, являясь существенно более ограничительным чем хорошо известное условие $\rho < s$ для классической системы $G/G/s$. Это предположение подкрепляется в частном случае, когда каждая заявка требует ровно s процессоров. Тогда исходная система переходит в классическую систему $G/G/1$ с условием стационарности $\rho < 1$.

Теорема 1. *В рассматриваемой модели при $\rho = ES/ET < 1$ условие $ES^{\alpha+1} < \infty$ является достаточным для конечности момента порядка α стационарного времени ожидания в системе, т. е. $ED^\alpha < \infty$.*

Доказательство. Отметим, что задержка в системе $D_n := W_n(1)$ ограничена снизу значением задержки $D_n^{(low)} := W_n^{(low)}(1)$ для соответствующей системы $\mathrm{G/G}/s$ с идентичными входным потоком с интервалами T_n и временами обслуживания S_n. Действительно, из рекурсии (8) следует, что, в предположении индукции $D_n \geqslant D_n^{(low)}$

$$\begin{aligned} D_{n+1} = W_{n+1}(1) &= (W_n(N_n) + S_n - T_n)^+ \\ &\geqslant (W_n(1) + S_n - T_n)^+ = (D_n + S_n - T_n)^+ \\ &\geqslant (D_n^{(low)} + S_n - T_n) \geqslant (D_n^{(low)} + P_n - T_n)^+ \\ &= D_{n+1}^{(low)}. \end{aligned}$$

Одновременно задержка D_n ограничена сверху задержкой $D_n^{(up)}$ для *доминирующей*

системы, где число требуемых каждой заявке процессоров $N_n = s, n \geqslant 1$. Действительно,

$$D_{n+1} = W_{n+1}(1) = (W_n(N_n) + S_n - T_n)^+ \leqslant (W_n(s) + S_n - T_n)^+ = (D_n^{(up)} + S_n - T_n)^+ = D_{n+1}^{(up)}.$$

Доминирующая система по сути представляет систему G/G/1, с теми же интервалами между приходами T_n и временами обслуживания S_n, что и в исходной системе. Таким образом, для любых реализаций $\{T_n, S_n\}$

$$D_n^{(low)} \leqslant D_n \leqslant D_n^{(up)}. \quad (9)$$

Переформулируя (9) на уровне моментов порядка α и устремляя $n \to \infty$, получим

$$E(D^{(low)})^\alpha \leqslant ED^\alpha \leqslant E(D^{(up)})^\alpha.$$

Однако у *нижней* и *верхней* систем обслуживания, при условии $ES/ET < 1$, одно и то же достаточное условие конечности моментов времени ожидания, именно, $ES^{\alpha+1} < \infty$. Следовательно, это условие является достаточным для конечности момента ED^α в рассматриваемой модели. □

Численный эксперимент

Моделирование рассматриваемых систем обслуживания проводилось на вычислительном кластере ЦКП КарНЦ РАН [ЦВОД ЦКП КарНЦ РАН, 2010], причем исходные данные описывают функционирование самого кластера. Именно данные для анализа получены из log-файлов системы управления заданиями Cleo, которая эксплуатировалась на кластере ЦКП КарНЦ РАН в период с 03.06.2009 г. по 04.02.2011 г.

Опишем подробнее структуру кластера и условия эксперимента. Кластер содержит 80 вычислительных ядер, сгруппированных по 8 ядер на каждый вычислительный узел. В процессе вычислений каждому пользователю разрешено занимать все узлы, но не менее одного (узел занимается пользователем целиком, даже если для вычислений используется лишь одно ядро этого узла). Кроме того, имеется ряд административных ограничений для пользователей кластера. В частности, введен запрет на постановку более чем 3 задач в очередь одновременно, введено ограничение на максимальное время вычисления задачи (3 суток), также есть ряд ограничений для студентов, использующих кластер в учебных целях.

За период эксплуатации системы Cleo на вычислительном кластере произведены расчеты 8292 заданий в однопроцессорном и многопроцессорном режимах. Общее время вычисления составило 28760 часов 50 минут 49 секунд. Среднее время вычисления одной задачи составило $ES \approx 3,4685$ часа или $0,1445$ суток. В то же время, астрономическое время эксплуатации системы составило 611 суток, т. е. среднее время между поступлениями заданий равно $ET \approx 1,7684$ часа или $0,0737$ суток.

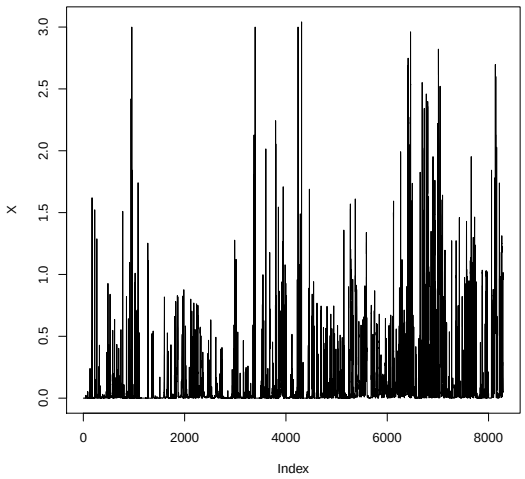

Рис. 3. Времена обслуживания заявок на кластере

Времена обслуживания. На рис. 3 изображены величины времен обслуживания заявок за рассматриваемый период. Из графика эмпирической функции распределения (рис. 4) видно, что большинство значений сконцентрировано в окрестности нуля. Это связано с тем, что пользователи кластера выполняют *отладку* задач, что часто приводит к почти мгновенному завершению задачи (в случае какой-либо ошибки в программе) и появлению (практически) нулевого времени вычисления в лог-файле. Для исключения отладочных запусков учитывались заявки, которые вычислялись более 1 часа, т. е. $0,04$ суток (2767 заявки).

Времена вычисления $\{S_n\}$ часто принимают значения, близкие к максимально допустимому (3 суток). Это, а также достаточно медленный рост эмпирической функции распределения может указывать на присутствие распределения с тяжелым хвостом. Однако ограниченность S (типичное время вычисления) исключает такую возможность. В та-

кой ситуации для моделирования времен обслуживания в компьютерных системах можно использовать усеченное распределение Парето [Harchol-Balter, 1999], имеющее функцию распределения

$$F(x) = \frac{\left(\frac{k}{x}\right)^{\alpha} - 1}{\left(\frac{k}{p}\right)^{\alpha} - 1},\ 0 < k \leqslant x \leqslant p,\ \alpha > 0. \quad (10)$$

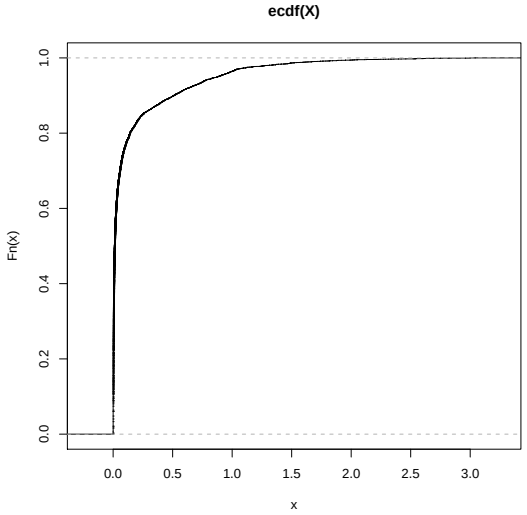

Рис. 4. Эмпирическая ф.р. времени вычисления

Случайная величина с таким распределением принимает значения в интервале $[k,\, p]$ и, кроме того, имеет конечные моменты всех порядков. Однако с убыванием параметра α имеет место экспоненциальный рост значений моментов [Harchol-Balter, 1999], (например, дисперсии при $\alpha < 2$). Это свойство до некоторой степени отражает свойство распределений с тяжелым хвостом. На рис. 5 показаны плотности распределения времен обслуживания (сплошная линия) и времен, имеющих усеченное распределение Парето (10) с параметрами $k = 0,04$, $p = 3,04$, $\alpha = 0,3$ (штриховая линия, выборка из 2767 значений). Для различения графиков в области малых значений S выбран логарифмический масштаб по оси абсцисс. В качестве гипотезы H_0 рассматривалось утверждение об одинаковом распределении двух выборок. Значение критерия Колмогорова-Смирнова для этих двух выборок равно $D = 0,0517$, а соответствующее значение статистического уровня значимости, при котором гипотеза H_0 принимается, не больше $0,001228$. Это значение мало, так как для практических целей обычно используют уровни значимости $\geqslant 0,01$ (чтобы увеличить мощность критерия, т. е. уменьшить вероятность ошибки второго рода).

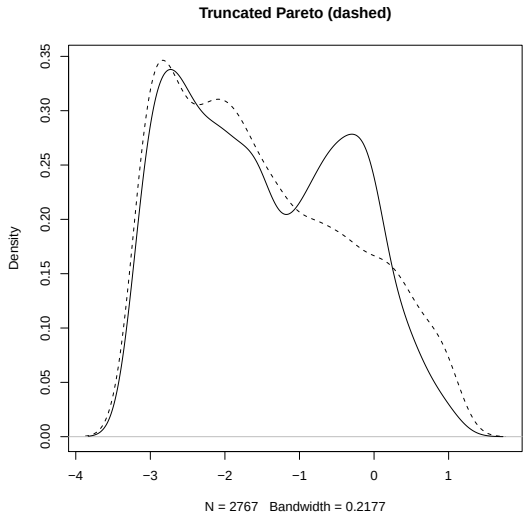

Рис. 5. Плотность распределения времени вычисления (усеченное Парето, $\alpha = 0,3$)

Ввиду широкого применения марковских моделей, естественно было бы также проверить применимость (усеченного) экспоненциального распределения, плотность которого имеет вид

$$f(x) = \frac{\lambda e^{-\lambda x}}{e^{-\lambda k} - e^{-\lambda p}},\ 0 < k \leqslant x \leqslant p, \qquad (11)$$

где постоянная $\lambda > 0$. Однако это распределение не подходит для моделирования времен $\{S_n\}$, как хорошо видно из рис. 6. Критерий Колмогорова-Смирнова дает в этом случае значение уровня значимости менее 10^{-16}. На рис. 6 в логарифмическом масштабе по оси абсцисс показаны плотности исходной выборки (сплошная линия) и выборки из распределения (11) (штриховая линия, параметры $k = 0,04$, $p = 3,04$, $\lambda = 2,25$).

Таким образом, использование распределения Парето статистически более обосновано. Полезно отметить, что распределение с тяжелым хвостом возможно сколь угодно точно аппроксимировать с помощью свертки экспоненциальных распределений, см. [Feldmann, Whitt, 1997]. Рассмотрим более подробно основные элементы рассматриваемой системы.

Входной поток. Рассмотрим времена между приходами заявок. Для устранения влияния этапа ввода кластера в эксплуатацию использовалась выборка, содержащая последние 2000 заявок. Кроме того, для исключения *отладочных* задач анализ был ограничен заявка-

ми, интервалы между приходами которых составляют не менее 50 секунд, т. е., $6{\cdot}10^{-4}$ суток (1632 заявок), см. рис. 7.

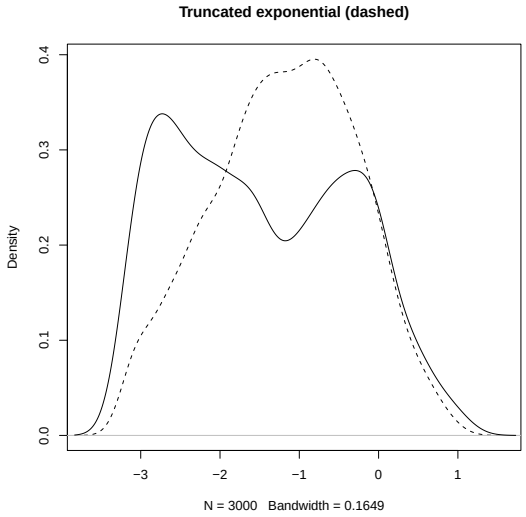

Рис. 6. Плотность распределения времени вычисления (усеченное экспоненциальное, $\lambda = 2,25$)

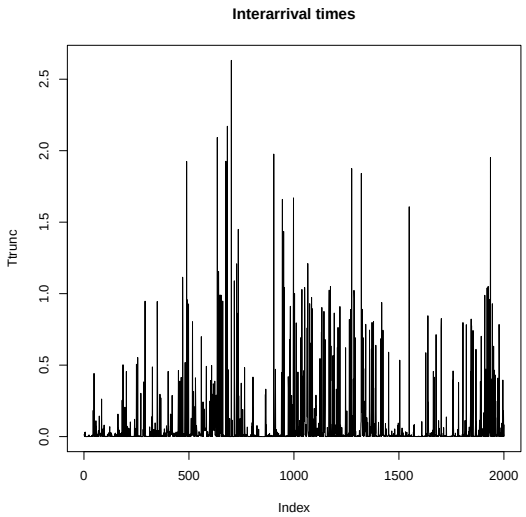

Рис. 7. Выборка интервалов между приходами (более 50 секунд)

Рассматривалась основная гипотеза H_0 о том, что интервалы $\{T_n\}$ между заявками имеют усеченное распределение Парето против альтернативной гипотезы (о том, что эти распределения различны). Рис. 8 иллюстрирует хорошее согласие между эмпирическими данными и усеченным распределением Парето с параметрами $k = 6 \cdot 10^{-4}, p = 2,63$, $\alpha = 0,22$. Графики плотностей построены по выборке из 2000 значений и представлены в логарифмическом масштабе по оси абсцисс. Критерий Колмогорова-Смирнова дает значение $D = 0,0384$ и значение уровня значимости не больше $0,1420$. Это значение достаточно для практических целей, так как позволяет принять гипотезу H_0, например, при уровне значимости $0,05$. (Отметим, что выбор параметров распределения можно производить с учетом методов, предложенных в работе [Aban et al., 2006].)

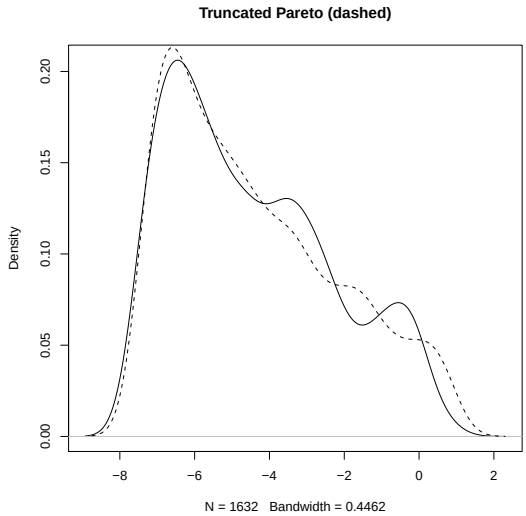

Рис. 8. Плотность распределения времени между приходами (усеченное Парето, $\alpha = 0,22$)

Число процессоров. На рис. 9 представлена гистограмма распределения (типичного) числа процессоров N, требуемых заявкой. Видно, что подавляющее число пользователей кластера ведет расчеты в однопроцессорном режиме, а программы, реализующие параллельные вычисления, используют число процессоров 8, 16, 32 и 64, кратное размеру узла.

С учетом наблюдений в модели принято равномерное распределение числа процессоров N в диапазоне от 8 до 64, кратных 8. Однако в дальнейшем целесообразно использовать иные распределения. В частности, анализ гистограммы показывает, что распределение Зипфа (дискретный аналог распределения Парето), имеющее вид

$$P(N = i) = \frac{i^{-\alpha}}{\sum_{k=1}^{s} k^{-\alpha}},\ \alpha > 0, 1 \leqslant i \leqslant s,$$

может достаточно адекватно описать распределение числа требуемых процессоров.

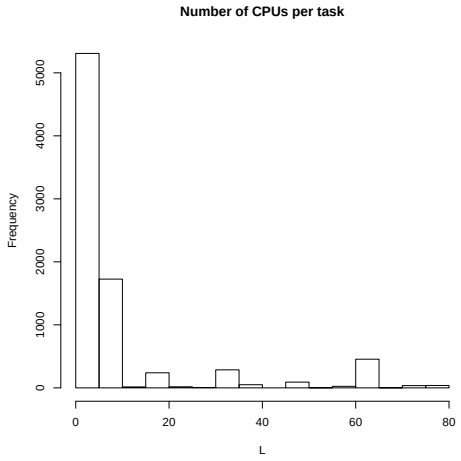

Рис. 9. Гистограмма числа процессоров в заявке

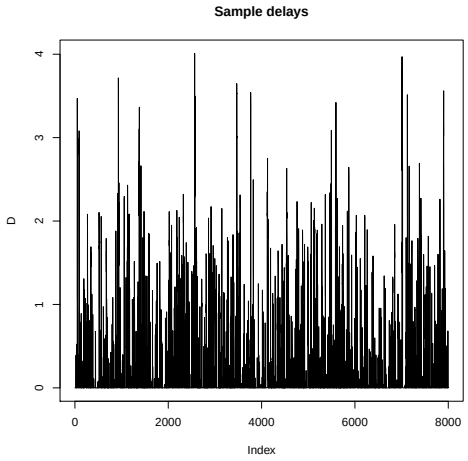

Рис. 10. Задержки в системе, число процессоров $s = 80$

Результаты экспериментов. Численное исследование модели велось на основе рекуррентного соотношения (8) для вектора загрузки системы. На рис. 10 представлен пример траектории задержек в очереди $\{D_n\}$ по выборке из 8000 значений (заявок). Средняя задержка равна $0,21$ суток. Максимальная задержка составила примерно 4 суток. Отметим, что хотя достаточное условие стационарности нарушено, $ES/ET > 1$, но неограниченное нарастание задержки не наблюдается. На гистограмме рис. 11 видно, что большинство задержек малы (меньше $0,5$ суток).

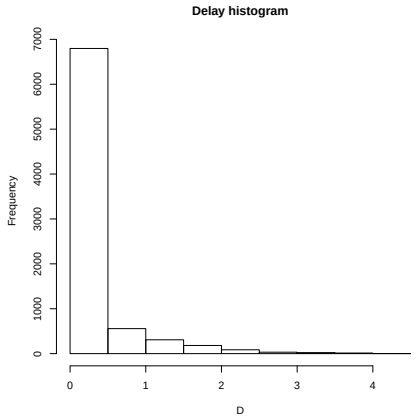

Рис. 11. Гистограмма задержек в системе, число процессоров $s = 80$

Проведенные эксперименты подтверждают, что модель адекватно отражает некоторые основные характеристики системы. Адекватность модели (несмотря на ее очевидную неполноту) позволяет, в частности, оценить эффект, вызываемый увеличением числа процессоров. Хорошо известно, что в классических системах обслуживания, в режимах, близких к полной загрузке, даже небольшое увеличение пропускной способности приводит к значительному сокращению времен ожидания. Предположим, что число узлов кластера увеличилось на два, т. е. имеется $s = 96$ процессоров. На рис. 12 видно, что число нулевых задержек в очереди увеличилось (6607 значений против 5755 в случае 80 процессоров). В этом случае средняя задержка составила приближенно $0,12$ суток, т. е., 57 % от значения для случая $s = 80$. Таким образом, незначительное увеличение числа процессоров позволяет существенно снизить среднюю задержку.

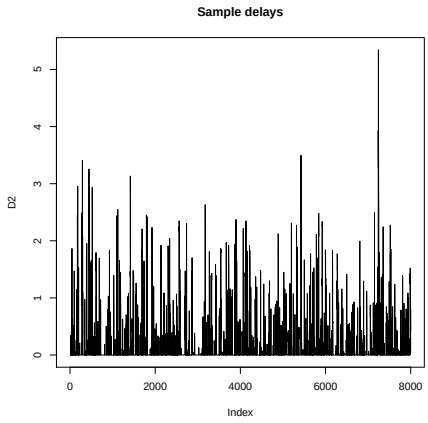

Рис. 12. Задержки в системе, число процессоров $s = 96$

Заключение

В работе рассмотрены некоторые важные результаты для многосерверных систем обслуживания, в том числе, когда время обслуживания имеет тяжелый хвост. Подробно проанализированы моментные свойства процесса задержки в очереди. Предложена модель вычислительного кластера на основе модифицированной рекурсии Кифера-Вольфовица для многосерверной системы G/G/s, для которой получены моментные свойства стационарного времени ожидания в очереди. Полученные результаты могут применяться для оценивания качества обслуживания вычислительных кластеров и Грид. Проведенный на кластере вычислительный эксперимент показал хорошее согласие предложенной модели с реальной работой кластера.

Дальнейшая верификация модели, в том числе проверка условий ее стационарности, может быть проведена с использованием логфайлов других вычислительных кластеров.

Кроме того, для лучшей адаптации модели к реальным кластерам целесообразно рассмотреть модели с конечным буфером для ожидания, с возможностью появления групп заявок, а также с более сложными дисциплинами выбора заявок на обслуживание (например, алгоритм Backfill и некоторые другие алгоритмы планировщиков). Поскольку аналитическое исследование таких моделей представляется трудно реализуемым, основным методом анализа может быть имитационное моделирование.

Литература

Центр высокопроизводительной обработки данных ЦКП КарНЦ РАН. URL: http://cluster.krc.karelia.ru (дата обращения: 13.12.2010 г.).

Aban I., Meerschaert M., Panorska A. Parameter Estimation for the Truncated Pareto Distribution // Journal of the American Statistical Association. 2006. Vol. 101, N. 473. P. 270–277.

Borst S., Boxma O., Núñez-Queija R. Heavy Tails:The Effect of the Service Discipline // Proceedings of Performance TOOLS 2002. 2002. P. 1–30.

Crovella M., Bestavos A. Self-similarity in World Wide Web traffic: evidence and possible causes. // IEEE/ACM Transactions on Networking. 2002. Vol. 5, N. 6. P. 835–845.

Feldmann A., Whitt W. Fitting Mixtures of Exponentials to Long-Tail Distributions to Analyze Network Performance Models // INFOCOM'97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE. 1997. P. 1096–1104.

Foss S., Korshunov D. Heavy Tails in Multi-Server Queue // Queueing Systems. 2006. Vol. 52. P. 31–48.

Harchol-Balter M. The Effect of Heavy-Tailed Job Size Distributions on Computer System Design // Proceedings of ASA-IMS Conference on Applications of Heavy Tailed Distributions in Economics, Engineering and Statistics, Washington, DC. June 1999.

Huang T., Sigman K. Steady-state Asymptotics for Tandem, Split-Match and Other Feedforward Queues with Heavy-Tailed Service // Queueing Systems. 1999. Vol. 33. P. 233–259.

Kiefer J., Wolfowitz J. On the characteristics of the general queueing process, with applications to random walks // Ann. Math. Statist. 1956. Vol. 27. P. 147–161.

Keilson J., Seidmann A. M/G/∞ with Batch Arrivals // Operations Research Letters October 1988. Vol. 7, N. 5. P. 219–222.

Liu L. Batch arrival infinite server queues with constant service times // Proceedings of the First Symposium on Queueing Theory in China September 1993. P. 47–53.

Morozov E., Pagano M., Rumyantsev A. Heavy-tailed Distributions with Applications to Broadband Communication Systems // Proceedings of AMICT'2007. 2008. Vol. 9. P. 157–174.

Morozov E., Rumyantsev A. Moment properties of queueing systems and networks // Proceedings of ICUMT'2010 (Ультрасовременные телекоммуникации и системы управления). 2010.

Scheller-Wolf A. Further delay moment results for FIFO multiserver queues // Queueing Systems. 2000. Vol. 34. P. 387–400.

Scheller-Wolf A., Sigman K. Delay Moments for FIFO GI/GI/s queues // Queueing Systems. 1997a. Vol. 25. P. 77–95.

Scheller-Wolf A., Sigman K. New bounds for expected delay in FIFO GI/GI/c queues // Queueing Systems. 1997b. Vol. 26. P. 169–186.

Scheller-Wolf A., Vesilo R. Sink or Swim Together: Necessary and Sufficient Conditions for Finite Moments of Workload Components in FIFO Multiserver Queues // Queueing Systems. January 2011. Vol. 67, N. 1. P. 47–61. March 2008.

Scheller-Wolf A., Vesilo R. Structural interpretation and derivation of necessary and sufficient conditions for delay moments in FIFO multiserver queues // Queueing Systems. 2006. Vol. 54. P. 221–232.

Wolff R. On Finite Delay-Moment Conditions in Queues // Operations Research. September-October 1991. Vol. 39, N. 5. P. 771–775.

СВЕДЕНИЯ ОБ АВТОРАХ:

Морозов Евсей Викторович
ведущий научный сотрудник, д. ф.-м. н., профессор
Институт прикладных математических исследований КарНЦ РАН
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: ar0@krc.karelia.ru
тел.: (8142) 763370

Morozov, Evsey
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Science
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: ar0@krc.karelia.ru
tel.: (8142) 763370

Румянцев Александр Сергеевич
аспирант
Институт прикладных математических исследований КарНЦ РАН
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: ar0@krc.karelia.ru
тел.: (8142) 763370

Rumyantsev, Alexandr
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Science
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: ar0@krc.karelia.ru
tel.: (8142) 763370