УДК 004.942

СОКРАЩЕНИЕ ВРЕМЕНИ АППРОКСИМАЦИИ НАГРУЗКИ ВЫЧИСЛИТЕЛЬНОГО КЛАСТЕРА С ИСПОЛЬЗОВАНИЕМ УПРОЩЕНИЯ ГИПЕР-ГАММА-РАСПРЕДЕЛЕНИЯ

© 2017 В. М. А. Ахмед, С. В. Гаевой, С. А. Фоменков

Волгоградский государственный технический университет (г. Волгоград, Россия)

Важным способом анализа нагрузки вычислительного кластера является моделирование его работы (обслуживания приходящих заданий) с использованием модели входящей нагрузки. Нами уже построен набор таких моделей, но все из них используют аппроксимацию непрерывных случайных величин. Данная работа предлагает упрощенный вариант аппроксимации гипер-гамма-распределением за счет сокращения числа параметров, который работает и с методом моментов, и методом наибольшего правдоподобия. Для проверки качества аппроксимации качества в данной статье используется метод имитационного моделирования с последующим сравнением результатов аппроксимации с результатами оригинала.

Ключевые слова: метод моментов, метод максимального (наибольшего) правдоподобия, нагрузки вычислительных систем, немасштабируемые задачи, имитационное моделирование, стохастическая аппроксимация, гипер-гамма-распределение.

Проблема рационального выполнения параллельных и высокопроизводительных вычислений сейчас достаточно актуальна [10]. В частности стоит вопрос оптимального балансирования нагрузки и подбора оптимальной производительности [9] вычислительных кластеров (ВК). Одним из возможных путей решения этой задачи является моделирование работы ВК – в том числе и имитационное [13]. Последнее требует построения математической модели ВК и поступающей на нее (т. е. входной) нагрузки [18].

В данной работе мы рассмотрим один из аспектов моделирования работы ВК – упрощение аппроксимации стохастической непрерывной величины. Ранее нами уже была построена модель [3, 5] рассматриваемого далее ВК и предложены способы моделирования для нее входящей нагрузки [1, 7], поэтому рассмотрим этот вопрос кратко.

Введем следующие определения из [9]. Длиной задания назовем время его выполнения. Площадью задания назовем произведение длины на ширину. Очевидно, что площадь – это сложность задания. Она же представляет собой суммарное машинное время обслуживания. Отметим, что авторы существующих публикаций по теме данной статьи используют и иную терминологию.

Реальный ВК построен из вычислительных машин, которые обслуживают поступающие задания [9, 13]. В данной работе используется представление такого ВК в виде обслуживающего блока с единой неприоритетной, не ограниченной по размеру очередью. Это обеспечивает обслуживание всех входящих заданий.

Каждое задание может исполняться параллельно на нескольких машинах (каналах обслуживания). Количество вычислительных машин, на которых исполняется задание, называется его шириной. Будем считать, что ширина задания определяется в момент его создания, что является довольно частым допущением [12, 15-17].

При завершении исполнения очередного задания очередь просматривается вся от начала и до конца для выбора (извлечения) заданий на выполнение. При этом извлекаемых из нее заданий может быть несколько – ведь ушедшее из очереди (т. е. исполненное) задание может освободить не только один,

Ахмед Весам Мохаммед Абдо – Волгоградский государственный технический университет, аспирант кафедры «САПРиПК», wesamalsofi@gmail.com.
Гаевой Сергей Владимирович – Волгоградский государственный технический университет, к. т. н., ст. преподаватель кафедры «ЭВМиС», gaevserge@mail.ru.
Фоменков Сергей Адексеевич – Волгоградский государственный технический университет д. т. н., проф., проф. кафедры «САПРиПК», saf@vstu.ru.

но и несколько каналов. Задание, для выполнения которого имеется достаточное количество свободных машин, ставится на исполнение. Затем просмотр очереди продолжается со следующего за ним задания.

Длиной очереди назовем количество, находящихся в ней заданий. Шириной (сложностью, площадью) очереди – сумму ширин (сложностей, площадей) входящих в нее заданий. Аналогично определим понятия длины, ширины и площади всей вычислительной системы, т. е. ВК.

Для генерации случайных нагрузок были предложены [1] различные методики. В основе их лежат идеи разделения входного потока на несколько потоков по ширине заданий и/или искажения временной шкалы для моделирования нестационарной интенсивности приходящих заданий.

При использовании любой из этих моделей возникает необходимость аппроксимации следующих видов случайных величин [12, 15-17]: интервалов времени между моментами прихода заданий; ширин заданий; длин или площадей заданий.

Ширина является дискретной случайной величиной и представлена конечным числом значений. Поэтому ее можно аппроксимировать просто в виде массива вероятностей для набора ширин.

Согласно [17], в логах доминируют именно задания, ширина которых является степенью двойки – даже тогда, когда к этому нет технических предпосылок. Такие работы, как [12], выделяют еще и другие доминирующие (но слабее, чем степени двойки) ширины заданий, например, кратные десяти, квадраты натуральных чисел. В других работах [11] наоборот пытаются уйти от этой тенденции.

Для аппроксимации интервалов времени между моментами приходами заданий, длины или площади необходимо использовать непрерывную случайную величину. Эту аппроксимацию можно осуществить как с помощью метода моментов (ММ), так и метода наибольшего правдоподобия (МНП).

Согласно сторонним [12, 15-17] и нашим работам [1-2, 4, 7], аппроксимируемые параметры являются случайными величинами с больш***и***м коэффициентом вариации, поэтому гипер-распределения подходят для целей аппроксимации. Предыдущие наши работы [1-2, 4, 7] показали, что одни из самых точных аппроксимаций дает МНП для гиперэкспоненциального и гипер-гамма-распределений, но наши аппроксимации МНП требуют значительных вычислительных ресурсов.

В работе [15] показано, что возможно использование ММ на гиперэрланговом распределении. Для этого авторам пришлось искусственным путем сократить число определяемых параметров, а параметр порядка просто подбирался.

В работе [7] нами было предложено упрощение, которое позволяет получить более грубое, но и более быстрое решение для гиперэкспоненциального распределения с использованием ММ. В данной работе ставится **цель** – провести аналогичное упрощение для гипергамма-распределения.

В своей работе мы проводим моделирование нагрузок ВК систем (генерацию) и моделирование обслуживания этих нагрузок. В качестве исходного материала мы используем реальные нагрузки ВК, взятые из логов [16]. Для проверки качества аппроксимаций мы сравниваем результаты моделирования сгенерированной нагрузки с результатами моделирования исходной нагрузки (детерминированная имитационная модель [3, 5]).

Введем следующие обозначения: $E(X)$ – математическое ожидание величины X, $VAR(X)$ – ее дисперсия, $stDev(X)$ – среднее квадратичное отклонение, $\operatorname{cov}(X)=\frac{stDev(X)}{E(X)}$ – коэффициент вариации. Если мы имеем дело с оценкой момента, то будем писать над ней горизонтальную черту.

Оценки начальных и второго центрального моментов имеют вид

$$\overline{E(X^k)}=\frac{1}{N}\sum_{i=1}^{N}X_i^k, \qquad (1)$$

$$\overline{VAR(X)}=\frac{N}{N-1}\left(\overline{E(X^2)}-\overline{E(X)}^2\right), \qquad (2)$$

где N – число наблюдений, X_i – конкретное *i*-ое наблюдение

Также нам потребуются $pdf(x)$ и $cdf(x)$ – дифференциальная и интегральная функции распределения.

Обозначим через $H(n)$ гиперэкспонециальное распределение:

$$pdf(x)=\sum_{i=1}^{n}\alpha_i\lambda_i e^{-\lambda_i x}, \qquad (3)$$

$$cdf(x)=1-\sum_{i=1}^{n}\alpha_i e^{-\lambda_i x}, \qquad (4)$$

где n – количество веток распределения (задается перед аппроксимацией как один из параметров распределения), α_i – вероятности использования веток ($1 \ge \alpha_i \ge 0$, $\sum_{i=1}^{n} \alpha_i = 1$), λ_i – интенсивность составляющего его экспоненциального распределения на ветке № i ($\lambda_i \ge 0$). Стоит также отметить, что $cov(X) \ge 1$.

Начальные моменты

$$E(X^k) = k! \sum_{i=1}^{n} \frac{\alpha_i}{\lambda_i^k}. \tag{5}$$

Из условия $\sum_{i=1}^{n} \alpha_i = 1$ следует, что одна из α_i задается по остаточному принципу. Поэтому число параметров этого распределения равно $2n - 1$.

Использовать ММ можно лишь для распределений, у которых не более четырех параметров, так как на практике определение моментов выше четвертого порядка затруднено из-за большой доли случайности [7-8, 15].

Число параметров гиперэкспоненциального распределения равняется $2n - 1$, поэтому для распределения с двумя ветками ММ применим. Обозначим решение из [7] как H_μ. В силу того, что аппроксимация МНП применима для любого количества веток, обозначим ее так же, как и самое распределение, то есть $H(n)$.

Обозначим $H\Gamma(n)$ гипер-гамма-распределение:

$$pdf(x) = \sum_{i=1}^{n} \alpha_i \lambda_i \frac{(\lambda_i x)^{\nu_i - 1}}{\Gamma(\nu_i)} e^{-\lambda_i x}, \tag{6}$$

$$cdf(x) = \sum_{i=1}^{n} \alpha_i P(\nu_i, \lambda_i x), \tag{7}$$

где n – количество веток распределения (задается перед аппроксимацией как один из параметров распределения), α_i – вероятности использования веток ($1 \ge \alpha_i \ge 0$, $\sum_{i=1}^{n} \alpha_i = 1$), λ_i – интенсивность ветки № i ($\lambda_i \ge 0$, но здесь найти физический смысл в параметре сложнее). Стоит также отметить, что здесь уже $cov(X) \in (0;\infty)$, что позволяет учесть большее количество случаев.

Начальные моменты $H\Gamma(n)$ имеют вид

$$E(X^k) = \sum_{i=1}^{n} \frac{\alpha_i \Gamma(\nu_i + k)}{\lambda_i^k \Gamma(\nu_i)}. \tag{8}$$

Очевидно, что приняв $\nu_i = 1$ во всех слагаемых, получаем моменты $H(n)$.

Число параметров гипер-гамма-распределения равняется $3n - 1$, поэтому даже для распределения с двумя ветками ММ не применим. Рассматривать вариант с одной веткой нет смысла, так как это обычно будет экспоненциальное или гамма-распределение, которые уже рассмотрены в [2, 4]. Значит, необходимо упрощение. Распределение $H\Gamma(2)$ имеет пять параметров. Это много для ММ – будем считать, что $\nu_1 = \nu_2 = \nu$, тогда распределение имеет четыре параметра и моменты принимают вид

$$E(X^k) = \frac{\Gamma(\nu + k)}{\Gamma(\nu)} \sum_{i=1}^{2} \frac{\alpha_i}{\lambda_i^k} \tag{9}$$

или

$$E(X^k) \frac{k!\,\Gamma(\nu)}{\Gamma(\nu + k)} = k! \sum_{i=1}^{2} \frac{\alpha_i}{\lambda_i^k}. \tag{10}$$

Иными словами, мы перешли от случайной величины X $H\Gamma(2)$ к случайной величине Y $H(2)$ с моментами

$$E(Y^k) = E(X^k) \frac{k!\,\Gamma(\nu)}{\Gamma(\nu + k)}. \tag{11}$$

Аппроксимация величины Y уже произведена нами в работе [7]. Перейдем в этом решении от дисперсии (второго центрального момента) ко второму начальному моменту:

$$E(Y) = \overline{E(Y)}, \tag{12}$$

$$E(Y^2) = \overline{E(Y^2)}, \tag{13}$$

$$cov^2(Y) = \frac{E(Y^2)}{E^2(Y)} - 1, \tag{14}$$

$$\beta = \sqrt{\frac{cov^2(Y) - 1}{2}}, \tag{15}$$

$$\gamma = \frac{\overline{E(Y^3)}}{6E^3(Y)\beta^3} - \frac{1 + 3\beta^2}{\beta^3}, \tag{16}$$

$$\alpha_1 = \max\left(\frac{1}{2}\left(1 + \frac{\gamma}{\sqrt{\gamma^2 + 4}}\right); \frac{\beta^2}{1 + \beta^2}\right), \tag{17}$$

$$\alpha_2 = 1 - \alpha_1, \tag{18}$$

$$\lambda_1 = \left(E(Y)\left(1-\sqrt{\frac{\alpha_2}{\alpha_1}\beta}\right)\right)^{-1}, \quad (19)$$

$$\lambda_2 = \left(E(Y)\left(1+\sqrt{\frac{\alpha_1}{\alpha_2}\beta}\right)\right)^{-1} \quad (20)$$

(для случая $cov\ (Y)=1$ получаем $\lambda_1 = \lambda_2 = (E(Y))^{-1}$ и α_1 можно выбрать любой, то есть получаем обычное экспоненциальное распределение).

Отметим, что такой переход к начальному моменту сокращает значение коэффициента вариации $cov(Y)$ в $\sqrt{\frac{N}{N-1}}$ раз, где N – число наблюдений (см. формулы (1) и (2)), но с точки зрения ММ мы можем придерживаться любого из этих вариантов (выбирать любые моменты, центральные или начальные). Учитывая, что размер выборки обычно бывает от нескольких сотен до нескольких сотен тысяч, эта разница почти незаметна. Поэтому для упрощения формул мы оперируем исключительно оценками начальных моментов.

Таким образом, зная ν, мы сможем по первым трем моментам определить оставшиеся три параметра распределения. Теперь, варьируя ν, мы сможем приблизить значения третьего и четвертого моментов оригинальной величины X. В силу особенностей гиперэкспоненциального распределения получить равенство третьего момента получается не всегда [7], поэтому его тоже надо учитывать.

Также надо учесть, что гиперэкспоненциальное распределение может аппроксимировать лишь распределения, у которых $cov(Y)\geq 1$, поэтому должно выполняться условие

$$VAR(Y) \geq E^2(Y), \quad (21)$$

то есть

$$E(Y^2) - E^2(Y) \geq E^2(Y) \quad (22)$$

или

$$E(Y^2) \geq 2E^2(Y). \quad (23)$$

Переходя к изначальной величине, получаем

$$E(X^2)\frac{2!\,\Gamma(\nu)}{\Gamma(\nu+2)} \geq 2E^2(X)\left(\frac{1!\,\Gamma(\nu)}{\Gamma(\nu+1)}\right)^2 \quad (24)$$

Преобразуем в

$$\frac{E(X^2)}{\nu(\nu+1)} \geq \frac{E^2(X)}{\nu^2}, \quad (25)$$

затем

$$\nu E(X^2) \geq (\nu+1)E^2(X), \quad (26)$$

а потом

$$\nu\left(E(X^2) - E^2(X)\right) \geq E^2(X), \quad (27)$$

то есть

$$\nu \geq \frac{E^2(X)}{VAR(X)}. \quad (28)$$

Таким образом,

$$\nu \geq \frac{1}{cov^2(X)}. \quad (29)$$

Для поиска ν воспользуемся методом Хука-Дживса. В качестве стартового значения возьмем $\nu = \frac{1}{\overline{cov(X)}^2}$, а в качестве минимизируемой функции

$$f(\nu) = \sum_{i=3}^{4}\left(\frac{E(X^i) - \overline{E(X^i)}}{\overline{E(X^i)}}\right)^2. \quad (30)$$

Заметим, что

$$\overline{cov(X)}^2 = \frac{\overline{E(X^2)}}{\overline{E(X)}^2} - 1, \quad (31)$$

так как мы используем только начальные моменты.

Обозначим такую аппроксимацию по аналогии $H\Gamma_\mu$. Также по аналогии будем использовать $H\Gamma(n)$ как обозначение аппроксимации МНП. Очевидно, что имеет смысл опробовать упрощение $\nu_1 = \nu_2 = \nu$ и на МНП-аппроксимации $H\Gamma(2)$, что сократит время аппроксимации. Обозначим этот способ как $H\Gamma_\lambda$.

Для какого бы распределения не использовался МНП, функция правдоподобия имеет вид:

$$L = \prod_{j=1}^{N} pdf(X_j), \quad (32)$$

где N – число наблюдений, X_j – конкретное наблюдение № j.

В соответствии с методом необходимо произвести максимизацию этой функции. Аналитическое решение в общем случае может быть затруднительно. Функция может принимать значения очень близкие к нулю, и это может привести к серьезным проблемам из-за округления чисел в ЭВМ, поэтому максимизацию функции надо заменить максимизацией её логарифма lnL (или минимизацией $-lnL$). В качестве метода численной оптимизации возьмем метод Хука-Дживса.

В [16] предоставлены логи работы реальных вычислительных систем, где указываются времена прихода заданий, длины и ширины. Будет использован лог UniLu-Gaia-2014-2.swf [14], который принадлежит кластеру шириной 2004 на момент фиксации лога.

В таблице 1 представлена производительность методов аппроксимации. В сравнение с [7] мы несколько оптимизировали алгоритм аппроксимации, поэтому время аппроксимации сократилось. Так же, как и в [7], наша аппроксимация может сопровождаться анализом. Можно заметить, что для H_μ и $H(2)$ анализ по-прежнему занимает примерно 2,5 мин. Для $H(3)$ он получается как бы короче, так как отчасти выполняется еще в процессе аппроксимации на простаивающих ядрах процессора. В остальных случаях анализ сопряжен с расчетом частичных гамма-функций, время которых зависит от значений аргументов, поэтому его результат менее предсказуем.

Таблица 1

Скорости выполнения различных видов аппроксимации

Анализ	Время выполнения					
Не проводится	16 сек.	1 мин. 44 сек	6 мин. 53 сек	18 сек.	2 мин. 22 сек.	4 мин. 58 сек.
Проводится	2 мин. 47 сек.	4 мин. 14 сек.	8 мин. 42 сек.	3 мин. 24 сек.	5 мин. 12 сек.	7 мин. 22 сек.

Для оценки качества аппроксимации было проведено стохастическое имитационное моделирование предложенных моделей. С этой целью было усовершенствовано средство, описанное в [3, 5]. В качестве допустимой погрешности имитационного моделирования была взята величина 5 %. В соответствии с Центральной предельной теоремой [8] это требует более 40 испытаний на проверку каждого из вариантов.

Каждая модель нагрузки состоит из двух частей: модели времен приходов заданий и модели обслуживания (определение длины, ширины и площади). Обсуждение этих моделей, как уже говорилось, тема отдельной статьи [2, 4]. Они рассматриваются в нашей публикации [1].

В качестве эталонного результата моделирования возьмем результат детерминированного моделирования исходной нагрузки ВК.

Результаты проверки моделей представлены в таблице 2.

Чтобы выбрать лучший вариант используется критерий отклонения

$$Dev = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\frac{\overline{P_i} - P_i}{P_i}\right)^2}, \qquad (33)$$

где M – число параметров, P_i – эталонное значение параметра, $\overline{P_i}$ – значение, полученное по стохастической модели.

В таблице 2 указаны два средних времени ожидания. Из-за того, что в очередь попадают не все задания, а лишь их определенный процент, то среднее время ожидания можно рассчитывать двумя способами: для всех заданий, учитывая нулевое время ожидания не попавших в очередь (без штриха) или только для попавших в очередь (со штрихом). В данном случае, если задание попадает в очередь, оно попадает туда надолго (примерно 30-40 минут, хотя в среднем на каждое задание приходится лишь 1-1,5 минуты ожидания).

Очевидно, что бессмысленно говорить об отклонении в 1 % при погрешности каждой из величин 5 %, поэтому $H\Gamma_\lambda$ и $H\Gamma(2)$ можно считать эквивалентными, но с учетом примерно в два раза более быстрой аппроксимации более удачной является $H\Gamma_\lambda$.

$H\Gamma_\mu$ оказалась не только значительно лучше H_μ при практически одинаковой скорости аппроксимации, но и сопоставима с $H(2)$ и $H(3)$, время работы которых значительно больше.

Таким образом, были предложены два новых способа аппроксимации непрерывных величин: $H\Gamma_\mu$ и $H\Gamma_\lambda$, которые дали довольно хорошие результаты при более низких временных затратах. Даже $H\Gamma_\lambda$ в сравнении с $H\Gamma(2)$ дала двукратный выигрыш во времени при том же качестве. Полученные методы планируется использовать при моделировании работы ВК кафедры «ЭВМиС», но и они могут быть обобщены и для аппроксимации иных вычислительных систем.

Все модели [1-5, 7] и рассматриваемые здесь реализованы в виде программного продукта «SWFJParser.JDSBrocker», зарегистрированного в госреестре [6].

Таблица 2

Результаты проверки моделей

	~&HΓλ/$ HΓλ	~&HΓ(2)/H Γ(2)^	~&H(3)/H(3)^	&~HΓμ/& HΓμ^	$~H(2)/$H (2)	~&Hμ/ Hμ^	Эталон
Среднее время выполнения, сек	14299	14200	14191	14332	14334	14198	14329
Среднее число выполняемых заданий	96,707	95,97	95,475	97,005	96,553	95,692	93,067
Среднее число занятых каналов	939,86	910,88	955,45	921,58	918,34	959,77	872,26
Среднее время ожидания, сек	76,798	65,846	57,245	81,735	90,422	39,225	72,41
Среднее время ожидания', сек	2272,6	2255,2	1808	1850,2	3013	1428,7	2259,7
Доля попавших в очередь	0,02652	0,02591	0,02823	0,04196	0,02379	0,02415	0,03204
Средняя длина очереди	0,52236	0,44629	0,38729	0,55681	0,6103	0,26619	0,4703
Средняя ширина очереди	18,413	18,943	11,001	20,381	18,714	11,499	15,31
Среднее время пребывания в системе, сек	14376	14265	14248	14413	14425	14238	14402
Средняя длина системы	97,229	96,416	95,862	97,562	97,163	95,958	93,537
Средняя ширина системы	958,27	929,82	966,45	941,96	937,05	971,27	887,57
Отклонение	0,09639	0,10001	0,1436	0,16498	0,18738	0,24807	0

ЛИТЕРАТУРА

1 Аппроксимация потока заданий на примере вычислительного кластера UniLu-Gaia / С. В. Гаевой, Весам М. А. Ахмед, Д. В. Быков, С. А. Фоменков // Известия ВолгГТУ. Сер. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. – Волгоград, 2017. – № 8 (203). – С. 96-102.

2 Гаевой С.В. Аппроксимация времени выполнения заданий на примере вычислительного кластера LPC EGEE 2004 / С. В. Гаевой, Ф. А. Х. Аль-Хадша, С. А.Фоменков // Известия ВолгГТУ. Серия «Актуальные проблемы управления, вычислительной техники и информатики в технических системах». Вып. 2: межвуз. сб. науч. ст. / ВолгГТУ. – Волгоград, 2014. – № 12 (139). – С. 135-141.

3 Гаевой С. В. Детерминированная имитационная модель кластеров грид-системы, обслуживающих задания / С. В. Гаевой, Ф. А. Х. Аль-Хадша, В. С. Лукьянов // Вестник компьютерных и информационных технологий. – 2014. – № 6. – С. 39-43.

4 Гаевой С. В. Моделирование работы вычислительного кластера на примере LANL CM5 [Электронный ресурс] / С. В. Гаевой, Ф. А. Х. Аль-Хадша // SCI-ARTICLE.RU: электронный периодический научный журнал. – 2013. – № 3 (ноябрь). – С. 304-313. - (http://sci-article.ru/stat.php?i=modelirovanie_rabotY_vYchislitelnogo_klastera_na_primere_LANL_CM5).

5 Детерминированная имитационная модель кластеров грид-системы для сравнения эффективности использования эвристик распределения заданий / С. В. Гаевой, Аль-

Ф. А. Х. Хадша, С. А. Фоменков, В. С. Лукьянов // Прикаспийский журнал: управление и высокие технологии. – 2014. – № 2. – С. 148-157.

6 Свид. о гос. регистрации программы для ЭВМ № 2017619355 от 24 августа 2017 г. РФ, МПК (нет). Средство аппроксимации и имитационного моделирования вычислительных нагрузок (SWFJParser.JDSBrocker) / С. В. Гаевой, В. М. А. Ахмед, С. А. Фоменков; ВолгГТУ. - 2017.

7 Сокращение времени аппроксимации логов вычислительного кластера с использованием методов моментов на гиперэкспоненциальном распределении / С. В. Гаевой, В. М. А.Ахмед, Д. В. Быков, С. А. Фоменков// Прикаспийский журнал: управление и высокие технологии. – 2017. – № 1. – С. 94-105.

8 Фоменков С. А. Математическое моделирование системных объектов: учеб. пособ. (гриф). Доп. УМО вузов по университетскому политехн. образованию / С. А. Фоменков, В. А. Камаев, Ю. А.Орлова; ВолгГТУ. – Волгоград, 2014. - 335 с.

9 Эвристики распределения задач для брокера ресурсов Grid [Электронный ресурс] / А. И. Аветисян [и др.]. – [2009]. – (http://www.citforum.ru/nets/digest/grid/index.shtml).

10 Clúster de balanceo de carga y alta disponibilidad para servicios web y mail [Электронный ресурс] / M. M. Sinisterra [et al.]. – (https://dialnet.unirioja.es/descarga/articulo/4364562.pdf)

11 Downey A. B. A Parallel Workload Model and Its Implications for Processor Allocation [Электронный ресурс] / A. B. Downey // The Rachel and Selim Benin School of Computer Science and Engineering. – [2013]. – (http://allendowney.com/research/allocation/)

12 The Feitelson 1996 Model [Электронный ресурс] // The Rachel and Selim Benin School of Computer Science and Engineering. – [2013]. – (http://www.cs.huji.ac.il/labs/parallel/workload/m_feitelson96/)

13 GridMe: Grid modeling environment [Электронный ресурс] // Google code. – [2014]. – (https://code.google.com/p/gridme/)

14 HPC @ Uni.lu [Электронный ресурс]. – [2017]. – (https://hpc.uni.lu/systems/gaia/)

15 The Jann et al 1997 Model [Электронный ресурс] // The Rachel and Selim Benin School of Computer Science and Engineering. – [2013]. – (http://www.cs.huji.ac.il/labs/parallel/workload/m_jann97/)

16 Logs of Real Parallel Workloads from Production Systems [Электронный ресурс] // The Rachel and Selim Benin School of Computer Science and Engineering. – [2013]. – (http://www.cs.huji.ac.il/labs/parallel/workload/logs.html).

17 Lublin, U. The Workload on Parallel Supercomputers: Modeling the Characteristics of Rigid Jobs [Электронный ресурс] / U. Lublin, D. G. Feitelson // The Rachel and Selim Benin School of Computer Science and Engineering. – [2013]. – (http://www.cs.huji.ac.il/~feit/papers/Rigid01TR.pdf).

18 Modeling of Workload in MPPs / J. Jann [et al.] // Job Scheduling Strategies for Parallel Processing: Lect. Notes Comput. Sci. / ed. by D.G. Feitelson, L. Rudolph. – Springer-Verlag, 1997. – vol. 1291. – P. 95-116.

REDUCING THE APPROXIMATION TIME OF CLUSTER WORKLOAD BY USING SIMPLIFIED HYPERGAMMA DISTRIBUTION

© 2017 W. M. A. Ahmed, S. V. Gaevoy, S. A. Fomenkov

Volgograd State Technical University (Volgograd, Russia)

An important method to analyze parallel workloads is modeling execution of those systems by using parallel workload models. We have already proposed many models, but all these models use a continuous variable approximation. In this paper a simplified method of Hypergamma distribution approximation is proposed. It reduces the number of the approximated distribution’s parameters and then uses Method of Moemnts or Maximum Likelihood Method. To validate the quality of the results described in this paper we use the simulation of this approximation and compare the results with the original workload in this paper.

Key words: computing cluster, method of moments, maximum likelihood method, cumulative distribution function, parallel workloads, workload approximation, rigid jobs, job length, simulation, stochastic approximation, Hypergamma distribution, Hyperexpoential distribution.