УДК 004.75

В. А. Богатырев, д-р техн. наук, проф., e-mall: vladirnir.bogatyrev@gmail.com, **А. В. Богатырев,** аспирант
Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики

Надежность функционирования кластерных систем реального времени с фрагментацией и резервированным обслуживанием запросов

Предложены модели функциональной надежности кластера при обслуживании запросов реального времени с разделением запроса на фрагменты (части) и возможностью их параллельного резервированного выполнения в разных узлах кластера. Проанализировано влияние фрагментации и резервного выполнения копий фрагментов запроса на вероятность выполнения кластером запроса в директивные сроки с учетом возможных сбоев, отказов и ошибок вычислений.

Ключевые слова: *надежность, резервированные вычисления, реальное время, кластер, запросы, система массового обслуживания, очередь, фрагментация*

Введение

Проектирование компьютерных систем ответственного назначения требует обеспечения высокой производительности, отказоустойчивости, живучести, информационной и функциональной надежности и безопасности [1—6].

Надежность компьютерных систем реального времени определяется не только безотказностью и отказоустойчивостью структуры, но и устойчивостью вычислительного процесса, требующей безошибочности и своевременности обслуживания функциональных запросов [7—10].

Устойчивость (надежность) вычислительного процесса к отказам и ошибкам может быть повышена в результате отправки копий поступающих в компьютерную систему запросов на резервированное обслуживание в несколько узлов одновременно [11—14].

Эффективность резервирования запросов во многом зависит от интенсивности потока запросов и выбора дисциплины диспетчеризации. Для систем, представимых многоканальными системами массового обслуживания с общей очередью бесконечной длины, в [11, 12] показано, что резервированное обслуживание запросов при малой загрузке позволяет сократить среднее время ожидания запросов. С учетом этого в [11] предложена дисциплина обслуживания, названная "широковещательное обслуживание с копированием запроса" (Broadcasting with a customer Copying), при которой каждый запрос в момент поступления направляется на резервированное выполнение в свободные узлы (приборы). Если число занятых приборов в момент поступления запроса меньше некоторого порогового значения, то запрос резервируется (копируется) во все свободные узлы, иначе он обслуживается в одном узле, причем запрос считается обслуженным при успешном выполнении хотя бы одной из его копий. Эффективность обслуживания с резервированием запросов в [11, 12] оценивается по критерию снижения среднего времени ожидания запросов.

Для компьютерных систем, функционирующих в реальном времени, эффективность определяется не столько средним временем ожидания запросов, сколько вероятностью безошибочного выполнения запросов в директивные сроки (своевременность безошибочного обслуживания).

Своевременность обслуживания запросов в вычислительных системах реального времени может оцениваться по вероятности времени ожидания запросов в очереди, меньшего предельно допустимого времени t_0 [13, 14] или по вероятности получения результатов к директивному сроку T_0.

Для повышения устойчивости к отказам и ошибкам вычислительного процесса в кластерных компьютерных системах реального времени копии запросов могут распределяться на выполнение в несколько узлов кластера, чем достигается резервирование вычислительного процесса.

В кластерных системах, серверные узлы которых соответствуют одноканальным системами массового обслуживания с локальными очередями (в отличие от многоканальных систем обслуживания с общей очередью [11]), возможен существенный разброс времен ожидания в разных очередях, что потенциально может привести к увеличению вероятности своевременного выполнения резервированных запросов хотя бы в одном узле кластера, несмотря на увеличение нагрузки узлов из-за резервирования запросов.

В работе [13] исследованы варианты дисциплин резервированного обслуживания запросов в кластерных системах реального времени, предусматривающих распределение каждого запроса на выполнение в два или несколько (k) узлов, с выдачей результатов в директивный срок, а в [14] поставлена и решена задача оптимизации резервированного обслуживания запросов реального времени в системах кластерной архитектуры.

Вместе с тем, в работах [13, 14] не проанализированы возможности повышения вероятности свое-

временного и безошибочного резервированного обслуживания запросов реального времени в результате их фрагментации (разделения на части) при распараллеливании выполнения фрагментов разными узлами кластера.

Цель работы — исследование возможностей повышения вероятности своевременного и безошибочного обслуживания запросов реального времени в результате их фрагментации и резервирования.

Для достижения цели ставится задача исследования дисциплин обслуживания, предусматривающих:

- возможность разделения (фрагментации) запроса на части, выполняемые в разных узлах кластера (распараллеливание обслуживания запроса);
- возможность фрагментации запроса с резервированным выполнением копий фрагментов в разных узлах кластера.

Условием успешного выполнения запроса в кластерных системах реального времени для исследуемых дисциплин обслуживания является своевременное безошибочное выполнение запросов, при котором каждый фрагмент запроса должен быть безошибочно выполнен к директивному сроку хотя бы в одном из выделенных для этого узлов.

Ограничения применения исследуемой дисциплины обслуживания могут быть вызваны невозможностью распараллеливания выполнения запросов на разные узлы кластера. К сдерживающим факторам эффективного применения рассматриваемой дисциплины могут быть отнесены дополнительные издержки на диспетчеризацию запросов и невозможность деления задачи (запроса) на равные части (фрагменты), число которых может быть ограничено.

Предлагаемая дисциплина может быть потенциально эффективно применена для класса потенциально распараллеливаемых задач, например, для матричных вычислений, для обработки массивов данных, вычислений, связанных с большим числом перебираемых вариантов решений.

Сложность проводимых исследований обусловлена наличием технического противоречия. Действительно, резервированное выполнение запросов несколькими узлами приводит к повышению надежности вычислений при необходимости получения безошибочного результата хотя бы одним узлом, но вызывает возрастание загрузки узлов и, как следствие, увеличение вероятности задержки запросов в очереди сверх предельно допустимого времени t_0. В то же время резервированное выполнение запросов разными узлами, с учетом стохастичности обслуживания, может повысить вероятность выполнения запроса в требуемый срок хотя бы одним из узлов. Разделение запросов на части с их выполнением в разных узлах приводит к ускорению вычислений, но в силу той же стохастичности обслуживания может привести к недопустимой задержке выполнения одной из частей при необходимости безошибочного выполнения в директивный срок всех фрагментов запроса.

Объект и задачи исследования

В качестве объекта исследований рассматривается вычислительный кластер, объединяющий n идентичных компьютерных узлов (серверов), в каждом из которых организуется собственная локальная очередь запросов. Будем считать известными: среднее время выполнения запросов v, интенсивность входного потока запросов Λ, интенсивности отказов λ_0 и ошибок вычислений λ_1 в узлах. Поступающий в кластер запрос может быть распределен на обслуживание в любой компьютерный узел, представляемый системой массового обслуживания типа M/M/1 с бесконечной очередью [15]. Предположим, что запрос может быть разделен на s равных частей, причем выполнение различных частей может осуществляться независимо (параллельно) в разных узлах кластера, при этом время выполнения фрагмента (и всего запроса с учетом распараллеливания) будет равно v/s. Каждый запрос или его фрагмент может копироваться и направляться на резервное выполнение в очередь разных узлов кластера. При поступлении запроса (или его фрагмента) в очередь узла в нем запускается таймер, отсчитывающий директивное (предельно допустимое) время ожидания t_0, при этом если фрагментация запроса не проводится, то $t_0 = T_0 - v$, а если проводится, то $t_0 = T_0 - (v/s)$. Считывание результатов осуществляется после отсчета таймерами директивного времени T_0. Потерями времени на диспетчеризацию запросов, их фрагментацию и дефрагментацию, а также на контроль безошибочности результатов будем пренебрегать.

Постановка задачи исследования дисциплины обслуживания с фрагментацией и резервированием обслуживания запросов

Для узлов кластера, представимых системами массового обслуживания типа M/M/1 с бесконечной очередью, вероятность r того, что время ожидания запросов без фрагментации и резервированного обслуживания меньше предельно допустимого значения t_0, вычисляется [15] по формуле

$$r_0 = 1 - \Lambda v \mathbf{e}^{-t_0(v^{-1} - \Lambda)}. \qquad (1)$$

В случае объединения n компьютеров в кластер при балансировке их нагрузки формула (1) оценки вероятности своевременности обслуживания запросов может быть модифицирована [13, 14]:

$$r_1 = 1 - \frac{\Lambda v}{n} \mathbf{e}^{-t_0\left(\frac{1}{v} - \frac{\Lambda}{n}\right)}. \qquad (2)$$

Формула (2) позволяет вычислить вероятность своевременности обслуживания запросов, когда каждый запрос направляется на выполнение в один из компьютерных узлов, а резервирование выполнения копий запроса не происходит.

С учетом увеличения загрузки узлов при резервировании вычислений в k узлах вероятность непревышения времени ожидания запросов предела t_0 в некотором (конкретном) узле вычисляется как

$$r_2 = 1 - \frac{\Lambda v k}{n}\mathbf{e}^{-t_0\left(\frac{1}{v}-\frac{\Lambda k}{n}\right)},$$

где $\rho = \Lambda v k/n$ — загрузка узла.

При независимости процессов обслуживания в разных узлах вероятность того, что хотя бы в одном из k узлов, выполняющих запрос, его задержка в очереди меньше предельно допустимой величины t_0, определяется выражением

$$R_2 = 1 - (1 - r_2)^k = 1 - \left[\frac{\Lambda v k}{n}\mathbf{e}^{-t_0\left(\frac{1}{v}-\frac{\Lambda k}{n}\right)}\right]^k,$$

отсюда можно найти среднее время ожидания

$$w = \int_0^\infty \left(\left[\frac{\Lambda v k}{n}\mathbf{e}^{-t_0\left(\frac{1}{v}-\frac{\Lambda k}{n}\right)}\right]^k\right) dt.$$

Результаты исследований [11—14] не затрагивают фрагментацию запросов с возможностью резервного выполнения фрагментов разными узлами кластера.

Цель работы — исследование возможностей повышения эффективности кластерных систем реального времени в результате фрагментации запросов и резервного выполнения копий фрагментов в различных узлах кластера, когда требуется безошибочное обслуживание запросов к директивным срокам (с учетом предельно допустимого времени ожидания).

В статье ставится задача построения моделей кластера реального времени и исследование эффективности резервированного выполнения фрагментированных запросов несколькими узлами при требовании выдачи результатов в директивные сроки.

Проводимые исследования предполагают анализ:

- своевременности резервированного обслуживания запросов с возможностью фрагментацией в предположении отсутствия отказов, сбоев и ошибок во время пребывания запроса (идеализированный случай для анализа кластера как системы массового обслуживания);
- надежности вычислительного процесса при требовании своевременности и безошибочности выдачи результатов резервированного обслуживания запросов с их возможной фрагментацией в условиях отказов, сбоев и ошибок выполнения запросов.

Показателем эффективности рассматриваемых кластерных систем является вероятность безошибочного обслуживания резервированного запроса реального времени с выполнением каждого выделяемого в нем фрагмента хотя бы в одном из узлов кластера за время, меньшее предельно допустимого времени ожидания обслуживания.

Вероятность своевременности резервированного обслуживания фрагментированных запросов

Рассмотрим оценку вероятности своевременности резервированного вычислительного процесса в предположении безотказности узлов и идеальности контроля, когда к моменту выполнения запроса без дополнительных задержек вырабатывается информация о правильности (достоверности) вычислений. Издержками на диспетчеризацию, фрагментацию и дефрагментацию запросов будем пренебрегать. Будем считать, что все запросы могут при фрагментации быть разделены на s равных частей.

При разделении каждого запроса на s равных частей (фрагментов), независимо (возможно одновременно) выполняемых разными узлами кластера, интенсивность фрагментированных запросов относительно исходного потока запросов Λ увеличивается в s раз. Вместе с тем, в результате сокращения среднего времени выполнения фрагментированных запросов в s раз в итоге средняя нагрузка узлов остается равной $\rho = \Lambda v/n$.

Таким образом, при фрагментации запросов без их резервированного выполнения разными узлами вероятность непревышения времени ожидания запросов предела t_0 в некотором (одном, конкретном) узле определяется выражением

$$r_3 = 1 - \frac{\Delta v}{n}\exp\left(-t_0\left(\frac{s}{v}-\frac{s\Lambda}{n}\right)\right),$$

а при резервированном независимом выполнении каждого фрагмента запроса в k узлах кластера (что увеличивает загрузку узла в k раз) вероятность непревышения времени ожидания запросов предела t_0 в некотором (конкретном) узле вычисляется как

$$r_{23} = 1 - \frac{\Delta v k}{n}\exp\left(-t_0\left(\frac{s}{v}-\frac{s\Lambda k}{n}\right)\right).$$

При фрагментации запроса на s частей время ожидания всех s фрагментов не должно превосходить директивный срок ожидания t_0, поэтому вероятность своевременности обслуживания фрагментированных запросов без их резервированного обслуживания в кластере вычисляется по формуле

$$R_3 = r_3^s = \left[1 - \frac{\Lambda v}{n}\exp\left(-t_0\left(\frac{s}{v}-\frac{s\Lambda}{n}\right)\right)\right]^s.$$

При k-кратном резервировании фрагментированного запроса, т. е. с независимым вычислением каждого из s фрагментов в k узлах кластера, вероятность своевременности обслуживания фрагментированного запроса при условии, что время ожидания каждого фрагмента не превосходит предельное время ожидания t_0 хотя бы в одном из k узлов, определяется выражением

$$R_{23} = [1 - (1 - r_{23})^k]^s =$$

$$= \left\{1 - \left[\frac{\Lambda k v}{n}\exp\left(-t_0\left(\frac{s}{v}-\frac{ks\Lambda}{n}\right)\right)\right]^k\right\}^s.$$

Среднее время ожидания запросов для рассматриваемых случаев можно вычислить следующим образом:

$$w_3 = \int_0^{\infty}\left(1-\left[1-\frac{\Lambda v}{n}\exp\left(-t_0\left(\frac{s}{v}-\frac{s\Lambda}{n}\right)\right)\right]^s\right)dt;$$

$$w_{23} = \int_0^{\infty}\left(1-\left\{1-\left[\frac{\Lambda k v}{n}\exp\left(-t_0\left(\frac{s}{v}-\frac{ks\Lambda}{n}\right)\right)\right]^k\right\}^s\right)dt.$$

С учетом того, что при обслуживании запросов с фрагментацией и без нее допустимое время ожидания определяется соответственно как $t_0 = T_0 - v$ и $t_0 = T_0 - v/s$, вероятности своевременного обслуживания запросов для исследуемых вариантов организации вычислительного процесса оценим как

$$R_1 = r_1 = 1 - (\Lambda/n)v\exp(-(T_0 - v)(v^{-1} - \Lambda/n);$$

$$R_2 = 1 - (1 - r_2)^k =$$

$$= 1 - \left[\frac{\Lambda k v}{n}\exp\left\{-(T_0 - v)\left(\frac{1}{v}-\frac{\Lambda k}{n}\right)\right\}\right]^k; \quad (3)$$

$$R_3 = r_3^s = \left[1-\frac{\Lambda v}{n}\exp\left(-\left(T_0-\frac{v}{s}\right)\left(\frac{s}{v}-\frac{s\Lambda}{n}\right)\right)\right]^s; \quad (4)$$

$$R_{23} = [1 - (1 - r_{23})^k]^s =$$

$$= \left\{1-\left[\frac{\Lambda k v}{n}\exp\left(-\left(T_0-\frac{v}{s}\right)\left(\frac{s}{v}-\frac{ks\Lambda}{n}\right)\right)\right]^k\right\}^s. \quad (5)$$

Последняя формула является обобщением формул для оценки вероятности допустимого времени ожидания R_1—R_3 с учетом значений k и s.

Вероятность своевременности и безошибочности обслуживания запросов

Определим вероятности своевременного и безошибочного выполнения запросов в кластере для рассматриваемых дисциплин обслуживания с резервированием фрагментированных запросов в предположении, что отказ или ошибка выполнения запроса могут возникать от момента его помещения в очередь узла до момента считывания результата (этот интервал времени составляет T_0).

Вероятность своевременности и безошибочности за время T_0 нахождения запроса (без резервирования и фрагментации) в узле кластера определяется выражением [13]:

$$B = r_1\exp(-\lambda T_0) =$$

$$= [1 - (\Lambda/n)v\exp(-(T_0 - v)(v^{-1} - \Lambda/n))]\exp(-\lambda T_0),$$

где r_1 — вероятность своевременности обслуживания запроса в узле кластера, вычисляемая по формуле (2), а $\exp(-\lambda T_0)$ — вероятность безошибочности и безотказности узла за время нахождения в нем запроса T_0, при этом $\lambda = \lambda_0 + \lambda_1$ — суммарная интенсивность отказов узлов и ошибок их вычислений.

Вероятность своевременного и безошибочного обслуживания запросов в кластере при резервировании запросов без фрагментации вычисляется по формуле

$$P_2 = 1 - (1 - B_2)^k,$$

где B_2 — вероятность своевременного и безошибочного выполнения запроса в одном узле кластера, определяемая выражением

$$B_2 = r_2\exp(-\lambda T_0) =$$

$$= \left[1-\frac{\Lambda k v}{n}\exp\left(-(T_0 - v)\left(\frac{1}{v}-\frac{\Lambda k}{n}\right)\right)\right]\exp(-\lambda T_0). \quad (6)$$

Вероятность своевременного и безошибочного обслуживания запросов в кластере с фрагментацией запросов без их резервирования вычисляется как

$$B_3 = (r_3\exp(-\lambda T_0))^s =$$

$$= \left[\left\{1-\frac{\Lambda v}{n}\exp\left(-\left(T_0-\frac{v}{s}\right)\left(\frac{s}{v}-\frac{s\Lambda}{n}\right)\right)\right\}\exp(-\lambda T_0)\right]^s. \quad (7)$$

Вероятность своевременного и безошибочного обслуживания запросов в кластере при резервировании и фрагментации запросов в предположении, что все запросы могут быть разделены на s равных частей, выполняемых в разных узлах, при независимом резервном обслуживании в k узлах каждого из s фрагментов, вычисляется по формуле

$$P_2 = [1 - (1 - B_{23})^k]^s,$$

где B_{23} — вероятность своевременного и безошибочного выполнения фрагмента запроса в одном узле кластера, определяемая выражением

$$B_{23} = \left\{1-\frac{k\Lambda v}{n}\exp\left(-\left(T_0-\frac{v}{s}\right)\left(\frac{s}{v}-\frac{ks\Lambda}{n}\right)\right)\right\}\exp(-\lambda T_0). \quad (8)$$

Вероятность своевременности и безошибочности обслуживания запросов с различной возможностью фрагментации

Рассмотренные выше модели соответствуют идеализированному случаю, когда все запросы допускают возможность фрагментации на одинаковое число частей, выполняемых разными узлами кластера.

Рассмотрим теперь случай неоднородности запросов по возможному числу частей, выделяемых при их фрагментации, и по кратности их резервированного обслуживания в кластере.

Допустим, что в общем случае имеется n типов запросов, которые допускают разбиение соответственно на $i = 1, 2, 3, ..., n$ фрагментов, независимо выполняемых в разных узлах кластера (запрос i-го типа допускает разбиение на i независимо выполняемых частей). Будем считать заданными кратности резервирования соответствующих типов запросов (g_1, g_2, g_3, ..., g_n) и известными вероятности поступления соответствующих типов запросов (b_1, b_2, b_3, ..., b_n) $\left(\sum_{i=1}^{n} b_i\right) = 1$, причем некоторые из вероятностей b_i могут быть нулевыми.

Математическое ожидание числа выделяемых фрагментов

$$S = \sum_{i=1}^{n} b_i i,$$

а средняя кратность резервированного выполнения запросов

$$K = \sum_{i=1}^{n} b_i k_i.$$

Интенсивность запросов, поступающих в некоторый узел, и интенсивность их обслуживания при фрагментации определим как $S\Lambda/n$ и S/v, а при резервировании запросов без фрагментации эти интенсивности равны $K\Lambda/n$ и $1/v$, если при этом также реализуется фрагментация, то рассматриваемые интенсивности определяются как $KS\Lambda/n$ и S/v.

С учетом того, что при обслуживании запросов с фрагментацией допустимое время ожидания для запросов i-го типа $t_0 = T_0 - (v/s_i)$, а без фрагментации — $t_0 = T_0 - v$, на основе модификации формул (3)—(5) вычислим вероятности своевременного обслуживания запроса i-го типа для вариантов:

а) с резервированием выполнения запросов без их фрагментации

$$A_{2i} = 1 - \left[\frac{\Lambda v K}{n}\exp\left\{-(T_0 - v)\left(\frac{1}{v} - \frac{\Lambda K}{n}\right)\right\}\right]^{k_i};$$

б) с фрагментацией без резервирования

$$A_{3i} = \left[1 - \frac{\Lambda v}{n}\exp\left(-\left(T_0 - \frac{v}{s_i}\right)\left(\frac{S}{v} - \frac{S\Lambda}{n}\right)\right)\right]^{s_i};$$

в) с резервированием и фрагментацией выполнения запросов

$$A_{23i} = \left\{1 - \left[\frac{\Lambda K v}{n}\exp\left(-\left(T_0 - \frac{v}{s_i}\right)\left(\frac{S}{v} - \frac{KS\Lambda}{n}\right)\right)\right]^{k_i}\right\}^{s_i}.$$

На основе модификации формул (6)—(8) вычислим вероятности своевременного и безошибочного обслуживания запроса i-го типа для вариантов:

а) с резервированным выполнением запросов без их фрагментации

$$P_{2i} = 1 - (1 - D_2)^{k_i},$$

где

$$D_2 = \left[1 - \frac{\Lambda v K}{n}\exp\left(-(T_0 - v)\left(\frac{1}{v} - \frac{\Lambda K}{n}\right)\right)\right]\exp(-\lambda T_0);$$

б) с фрагментацией запросов без их резервирования

$$P_{3i} = \left[\left\{1 - \frac{\Lambda v}{n}\exp\left(-\left(T_0 - \frac{v}{S}\right)\left(\frac{S}{v} - \frac{S\Lambda}{n}\right)\right)\right\}\exp(-\lambda T_0)\right]^{s_i};$$

в) при резервировании и фрагментации запросов

$$P_{23i} = [1 - (1 - D_{23})^{k_i}]^{s_i},$$

где вероятность своевременного и безошибочного выполнения фрагмента запроса в одном узле кластера

$$D_{23} = \left\{1 - \frac{K\Lambda v}{n}\exp\left(-\left(T_0 - \frac{v}{S}\right)\left(\frac{S}{v} - \frac{KS\Lambda}{n}\right)\right)\right\}\exp(-\lambda T_0).$$

Примеры расчета вероятности своевременного и безошибочного обслуживания запросов

Расчеты проведем для случая, когда все запросы могут быть разделены на s равных частей, выполняемых в разных узлах.

Зависимость вероятности своевременного обслуживания при фрагментации запроса (без резервирования) от интенсивности потока запросов Λ представлена на рис. 1. Запрос считается успешно выполненным в кластере при условии, что время ожидания всех s фрагментов не превосходит директивный срок ожидания $t_0 = T_0 - (v/s)$. Расчет проведен при $v = 1$ с и числе узлов кластера $n = 5$.

Кривые *1—4* соответствуют разбиению запроса на $s = 1, 2, 3, 4$ частей при директивном сроке пребывания запросов $T_0 = 1{,}5$ с, а кривые *5—8* — разбиению на $s = 1, 2, 3, 4$ при директивном сроке $T_0 = 3$ с.

Из представленных графиков видно существование границы интенсивности входного потока, ниже которой разбиение запроса на части с их параллельным обслуживанием в разных узлах эффективно (приводит к увеличению вероятности обслуживания запроса в директивные сроки), а выше — нет. Причем при увеличении числа фрагментов, на

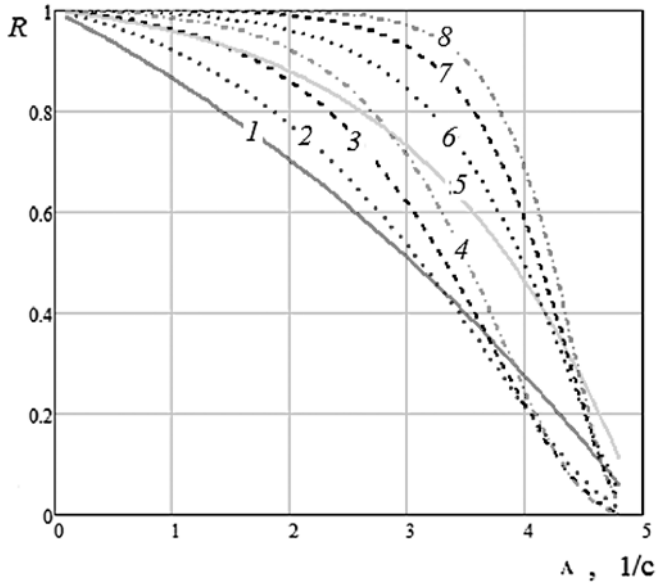

Рис. 1. Зависимость вероятности своевременного обслуживания при фрагментации запросов без их резервирования

которые разбивается запрос, вероятность своевременного обслуживания запросов при нагрузке ниже граничной увеличивается. Приведенные расчеты подтверждают, что чем менее жесткие директивные сроки выполнения запросов, тем больше вероятность их своевременного обслуживания.

Зависимость вероятности своевременного обслуживания запросов от интенсивности их входного потока Λ при резервировании копий фрагментированных запросов представлена на рис. 2. Запрос считается успешно выполненным в кластере, если время ожидания хотя бы одной из k резервных копий, выполняемых в разных узлах кластера, всех s фрагментов не превосходит директивный срок ожидания $t_0 = T_0 - (v/s)$. Расчет проведен при $v = 1$ с, директивном сроке обслуживания $T_0 = 1{,}5$ с и числе узлов кластера $n = 15$.

Кривые *1—3* соответствуют разбиению запроса на $s = 1, 2, 3$ частей без резервирования их выполнения, кривые *4—6* — разбиению запроса на $s = 1, 2, 3$ частей при дублировании их выполнения в разных узлах ($k = 2$), а кривые *7—9* — разбиению запроса на $s = 1, 2, 3$ частей при кратности резервирования вычислений $k = 3$.

Результаты расчетов показывают возможность увеличения вероятности своевременного выполнения запросов при разбиении запросов на части (фрагменты) с резервированным параллельным выполнением фрагментов запроса в разных узлах кластера. При этом при низкой загрузке эффективность обслуживания запросов по вероятности своевременного обслуживания повышается при увеличении числа фрагментов и кратности резервирования их выполнения в разных узлах кластера. По мере увеличения интенсивности входного потока запросов (загрузки узлов) для увеличения вероятности своевременного безошибочного выполнения запросов число частей (фрагментация) разбиения запроса для параллельного обслуживания и кратность резервирования при их выполнении

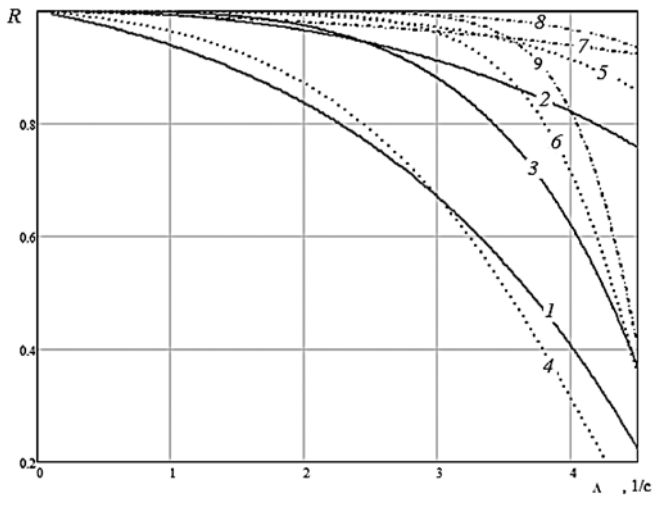

Рис. 2. Зависимость вероятности своевременного обслуживания при резервировании и фрагментации запросов

должна уменьшаться. При этом выбор оптимального числа частей разбиения запроса и кратность их резервирования может определяться на основе рассмотренных моделей.

Предложенный подход позволяет реализовать адаптивное обслуживание запросов с выбором числа выделяемых фрагментов и кратности их резервирования в зависимости от интенсивности входного потока и загруженности узлов кластера.

Для решения целесообразности применения предлагаемых дисциплин обслуживания в конкретных прикладных компьютерных системах необходим дополнительный комплекс работ, связанных с учетом конкретных особенностей реализуемых систем и прикладных процессов реального времени.

Достоверность принятия решений о целесообразности рассматриваемых дисциплин обслуживания может быть повышена при построении моделей с учетом:

- взаимозависимости запросов и возможностей их распараллеливания (в том числе на неравные части, распределение случайной величины времени, которое может отличаться от экспоненциального);
- влияния на эффективность фрагментации процессов распараллеливания, диспетчеризации и распределения фрагментированных запросов между узлами кластера;
- синхронизации вычислении при возможной взаимной зависимости результатов выполнения резервированных фрагментов.

Решение о целесообразности применения предлагаемых дисциплин обслуживания в конкретных условиях реализации кластеров, помимо использования предлагаемых аналитических моделей, может потребовать их уточнение, использование имитационного моделирования или натурных испытаний.

Заключение

Для кластерных систем реального времени, консолидирующих ресурсы серверов, с организацией локальных очередей в каждом из них предложена оценка вероятности того, что время ожидания запросов при их фрагментации и резервированном параллельном выполнении фрагментов в разных узлах кластера меньше предельно допустимой величины t_0.

Проанализировано влияние фрагментации и кратности резервирования вычислений на вероятность своевременного выполнения запросов в условиях возможных отказов и ошибок вычислений.

Показана возможность увеличения вероятности своевременного выполнения запросов при разбиении запросов на части (фрагменты) с резервированным параллельным выполнением фрагментов запроса в разных узлах кластера. При этом при низкой загрузке вероятность своевременного обслуживания повышается при увеличении числа фрагментов и

кратности резервирования их выполнения в разных узлах кластера.

Показано существование границы интенсивности входного потока, ниже которой разбиение запроса на части с их параллельным выполнением разными узлами приводит к увеличению вероятности обслуживания запроса в директивные сроки, причем эффективность фрагментации возрастает при увеличении кратности резервирования при выполнении фрагментов.

Предложенные модели могут быть использованы при оценке функциональной надежности и выборе организации распараллеливания и резервирования вычислительного процесса реального времени для вычислительных и управляющих систем кластерной архитектуры.

Список литературы

1. **Kopetz H.** Real-Time Systems: Design Principles for Distributed Embedded Applications. Springer. 2011. 396 p.
2. **Черкесов Г. Н.** Живучесть и отказобезопасность ответственных технических систем // Проектирование и технология электронных средств. 2015. № 1. С. 15—24.
3. **Шубинский И. Б.** Функциональная надежность информационных систем: методы анализа // Надежность, 2012. 296 с.
4. **Gurov S. V., Utkin L. V.** An inverse problem of the load-sharing system reliability analysis: Constructing the load function // Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability. 2015. P. 1—10.
5. **Aleksanin S. A., Zharinov I. O., Korobeynikov A. G., Perezyabov O. A., Zharinov O. O.** Evaluation of chromaticity coordinate shifts for visually perceived image in terms of exposure to external illuminance // ARPN Journal of Engineering and Applied Sciences. 2015. Т. 10. № 17.
6. **Aliev T. I., Rebezova M. I., Russ A. A.** Statistical Methods for Monitoring Travel Agencies // Automatic Control and Computer Sciences. 2015. Vol. 49, N. 6. P. 321—327.
7. **Богатырев В. А., Богатырев А. В., Богатырев С. В.** Оценка надежности выполнения кластерами запросов реального времени // Изв. вузов. Приборостроение. 2014. Т. 57. № 4. С. 46—48.
8. **Богатырев В. А., Богатырев А. В., Богатырев С. В.** Перераспределение запросов между вычислительными кластерами при их деградации // Изв. вузов. Приборостроение. 2014. Т. 57, № 9. С. 54—58,.
9. **Богатырев В. А.** Оценка надежности и оптимальное резервирование кластерных компьютерных систем // Приборы и системы. Управление, контроль, диагностика. 2006. № 10. С. 18—21.
10. **Богатырев В. А., Богатырев С. В., Богатырев А. В.** Оптимизация кластера с ограниченной доступностью кластерных групп // Научно-технический вестник ИТМО. 2011. № 1 (71). С. 63—67.
11. **Дудин А. Н., Сунь Б.** Многолинейная ненадежная система с управляемым широковещательным обслуживанием // Автоматика и телемеханика. 2009. Т. 70. № 12. С. 147—160.
12. **Dudin A. N., Sun B.** A multiserver MAP/PH/N system with controlled broadcasting by unreliable servers // Automatic Control and Computer Sciences. 2009. N. 5. P. 32—44.
13. **Bogatyrev V. A., Bogatyrev A. V.** Functional Reliability of a Real-Time Redundant Computational Process in Cluster Architecture Systems // Automatic Control and Computer Sciences. 2015. Vol. 49, N. 1. P. 46—56.
14. **Богатырев В. А., Богатырев А. В.** Оптимизация резервированного распределения запросов в кластерных системах реального времени // Информационные технологии. 2015. Т. 21, № 7. С. 495—502.
15. **Вишневский В. М.** Теоретические основы проектирования компьютерных сетей М.: ТЕХНОСФЕРА, 2003. 512 с.

V. A. Bogatyrev, Professor, e-mail: Vladimir.bogatvrev@gmail.com,
A. V. Bogatyrev, Post Graduate
Saint Petersburg National Research University of Information, Technologies, Mechanics and Optics

The Reliability of the Cluster Real-Time Systems with Fragmentation and Redundant Service Requests

For clustered systems, real-time, consolidating server resources, with the organization of local queues in each of them proposed the assessment of the probability that the waiting time of requests in their fragmentation and redundant parallel execution fragments in different nodes of a cluster is lower than the maximum permissible value.

Purpose — to improve the timeliness of service requests in real time the result of fragmentation and the reservation requests.

To achieve the goal of the research disciplines of service at the requirement of delivery of results within the deadlines, providing:

- *splitting (fragmentation) of the query into parts that are executed in different nodes of a cluster (paralleling service the request);*
- *fragmentation of the query with redundant execution copies of fragments in different nodes of a cluster.*

Analyzed the impact of fragmentation and multiplicity of redundancy calculations on the probability of timely fulfillment of requests in terms of possible failures and errors of calculations.

The possibility of increasing the probability of timely fulfillment of requests when splitting a query into pieces (fragments) with redundant parallel execution of query fragments to different nodes of a cluster. In this case, when the low probability of timely maintenance increases with the increase of number oi fragments and the multiplicity of redundancy they will run on different nodes of a cluster.

The existence of the border intensity input stream, below which the query is broken into pieces with their concurrent execution of different nodes leads to an increase in the probability of servicing the request within the deadlines, and the fragmentation efficiency increases with increasing multiplicity of redundancy when performing fragments.

The proposed model can be used to assess the functional reliability and the choice of the paralleling and redundancy of the computational process for real-time computing and control systems of cluster architecture.

Keywords: *reliability, redundant computation, real-time, cluster, requests, Queuing system, Queuing, fragmentation*

References

1. **Kopetz H.** *Real-Time Systems: Design Principles for Distributed Embedded Applications.* Springer, 2011, 396 p.
2. **Cherkesov G. N.** Zhivuchest' i otkazobezopasnost' otvetstvennyh tehnicheskih sistem, *Proektirovanie i tehnologija jelektronnyh sredstv*, 2015, no. 1, pp. 15—24.
3. **Shubinskij I. B.** Funkcional'naja nadezhnost1 informacionnyh sistem: metody analiza. *Nadezhnost'*, 2012. 296 p.
4. **Gurov S. V., Utkin L. V.** An inverse problem of the load-sharing system reliability analysis: Constructing the load function, *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 2015, pp. 1—10.
5. **Aleksanin S. A., Zharinov I. P., Korobeynikov A. G., Perezyabov O. A., Zharinov O. O.** Evaluation of chromaticity coordinate shifts for visually perceived image in terms of exposure to external illuminance, *ARPN Journal of Engineering and Applied Sciences*, 2015, vol. 10, no. 7.
6. **Aliev T. I., Rebezova M. I., Russ A. A.** Statistical Methods for Monitoring Travel Agencies, *Automatic Control and Computer Sciences*, 2015, vol. 49, no. 4, pp. 321—327.
7. **Bogatyrev V. A., Bogatyrev A. V., Bogatyrev S. V.** Ocenka nadezhnosti vypolnenija klasterami zaprosov real'nogo vremeni, *Izv. vyzov. Priborostroenie,* 2014, vol. 57, no. 4, pp. 46—48.
8. **Bogatyrev V. A., Bogatyrev A. V., Bogatyrev S. V.** Pereraspredelenie zaprosov mezhdu vychislitel'nymi klasterami pri ih degradacii, *Izv. vuzov Priborostroenie*, 2014, vol. 57, no. 9, pp. 54—58.
9. **Bogatyrev V. A.** Ocenka nadezhnosti i optimal'noe rezervirovanie klasternyh komp'juternyh sistem, *Pribory i sistemy. Upravlenie, kontrol', diagnostika*, 2006, no. 10, pp. 18—21.
10. **Bogatyrev V. A., Bogatyrev S. V., Bogatyrev A. V.** Optimizacija klastera s ogranichennoj dostupnost'ju klastemyh grupp, *Nauchno-tehnicheskij vestnik ITMO*, 2011, no. 1 (71), pp. 63—67.
11. **Dudin A. N., Sun' B.** Mnogolinejnaja nenadezhnaja sistema s upravljaemym shirokoveshhatel'nym obsluzhivaniem, *Avtomatika i telemehanika*, 2009, vol. 70, no. 12, pp. 147—160.
12. **Dudin A. N., Sun' B.** A multiserver MAP/PH/N system with controlled broadcasting by unreliable servers, *Automatic Control and Computer-Sciences,* 2009, no. 5, pp. 32—44.
13. **Bogatyrev V. A., Bogatyrev A. V.** Functional Reliability of a Real-Time Redundant Computational Process in Cluster Architecture Systems, *Automatic Control and Computer Sciences*, 2015, vol. 49, no. 1, pp. 46—56.
14. **Bogatyrev V. A., Bogatyrev A. V.** Optimizacija rezervirovannogo raspredelenija zaprosov v klastemyh sistemah real'nogo vremeni, *Informacionnye tehnologii*, 2015, vol. 21, no. 7, pp. 495—502.
15. **Vishnevskij V. M.** *Teoreticheskie osnovy proektirovanija komp'jutemyh setej.* Moscow: TEHNOSFERA, 2003. 512 p.

УДК 004.09

М. Ю. Неустроев, аспирант, e-mail: m.neustroev@gmail.com

Государственное бюджетное образовательное учреждение высшего образования Московской области "Технологический университет", г. Королев, Московская область

Анализ показателей эффективности и скорости обслуживания в центрах обработки вызовов

Проводится анализ уже существующих моделей центров обслуживания вызовов с применением IVR-систем, показана необходимость дальнейших исследований и разработки алгоритмов интеллектуального управления вызовами IVR-системой без участия операторов для снижения загруженности.

Ключевые слова: *Центр обслуживания вызовов (ЦОВ), интерактивное голосовое меню (IVR), эффективность ЦОВ, ASA, время ожидания, время обслуживания, загрузка операторов, голосовые сообщения*

Введение

Чтобы выжить в современной конкурентной среде, компаниям необходимо найти способы повышения качества обслуживания клиентов. Современный потребитель услуг знает, чего он хочет. Теперь если предприятие хочет повысить уровень оказываемых услуг и пойти дальше, чем простое выживание, а также добиться стабильного успеха, оно должно модернизировать и внедрять новые сервисы повышения услуг обслуживания своих клиентов. Интеграция Центров осблуживания вызовов (ЦОВ) в текущую информационную структуру помогает повысить производительность и эффективность обслуживания клиентов [1, 2]. Создание собственного ЦОВ — серьезный шаг, требующий значительных затрат. И если для крупной компании такой шаг может быть оправдан, для среднего и особенно малого бизнеса попытка самостоятельно обслуживать входящие и совершать исходящие звонки может стать дорогой ошибкой с потерей не только денег, но и клиентов.

Большинство современных ЦОВ используют одну или более технологий обработки вызовов, таких как автоматическое распределение звонков (ACD), интерактивное голосовое меню (IVR), распознавание речи (VRU). Для качественного и гарантированного обслуживания поступающих звонков каждый из модулей должен быть изучен и правильно запрограммирован. Новые типовые системы распределения вызовов значительно отличаются от систем предыдущего поколения своими алгоритмами и процессами [3].

Показатели эффективности обслуживания в ЦОВ

Для прогнозирования эффективности работы, выполняемой ЦОВ, на передний план выдвигаются