УДК 004.75

В. А. Богатырев, д-р техн. наук, проф., e-mail: vladimir.bogatyrev@gmail.com, **А. В. Богатырев,** аспирант,
Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики (Университет ИТМО)

Модель резервированного обслуживания запросов реального времени в компьютерном кластере

Для вычислительных систем кластерной архитектуры, работающих в реальном времени, представляемых группой одноканальных систем массового обслуживания с локальными очередями, предложена модель оценки вероятности своевременного и безошибочного обслуживания запросов с выполнением копий этих запросов в нескольких узлах. Рассмотрены варианты диспетчеризации с уничтожением и без уничтожения резервных копий запросов, ожидающих в локальных очередях узлов сверх допустимого времени.

Определена зависимость вероятности своевременного и безошибочного обслуживания запросов от кратности резервирования копий запросов, направляемых на обслуживание в разные узлы.

Ключевые слова: *модель, надежность, резервированное обслуживание запросов, реальное время, кластер, диспетчеризация*

Введение

К компьютерным системам, решающим ответственные задачи, предъявляют высокие требования по отказоустойчивости, надежности вычислительного процесса, достигаемые в результате резервирования и консолидации ресурсов системы при объединении компьютерных узлов в кластеры [1—6].

Для компьютерных систем реального времени при обеспечении функциональной надежности критичны безошибочность и своевременность обслуживания запросов [1—5]. Под функциональной надежностью понимается надежность (устойчивость) вычислительного процесса по своевременному безошибочному выполнению поступающих в систему запросов в условиях отказов, сбоев, ошибок и деструктивных воздействий [6, 7].

Для структурно избыточных компьютерных систем, включая кластеры, устойчивость (надежность) вычислительного процесса повышается при динамическом распределении запросов [8—15], в том числе при направлении резервных копий запросов, поступающих в систему на обслуживание в несколько узлов (каждая копия выполняется отдельным узлом).

Модели многоканальных систем обслуживания с общей очередью при резервированном выполнении запросов предложены и исследованы в работах [16, 17]. Для организации резервированного обслуживания, названного в работах [16, 17] "широковещательное обслуживание с копированием запроса", запрос направляется в свободные в момент его поступления узлы (приборов), причем в каждом узле обслуживание резервной копии запроса выполняется независимо. В работе [17] рассматривается адаптивное широковещательное (резервированное) обслуживание запросов, при котором, если в момент поступления запроса число занятых приборов меньше некоторого порогового значения, то резервные копии запроса направляются для выполнения во все свободные узлы, иначе запрос обслуживается только в одном из узлов. Резервированный запрос считается успешно выполненным при его обслуживании хотя бы в одном из узлов [16, 17]. Эффективность обслуживания запросов в работах [16, 17] оценивается средним временем ожидания, вместе с тем для систем реального времени более важна своевременность вычислений, определяемая по вероятности выполнения запросов за время, меньшее предельно допустимого значения или по вероятности ожидания запросов меньше предельно допустимого времени t [18].

В кластерных системах, консолидирующих ресурсы нескольких серверов, особенность выполнения запросов (в том числе с их резервированным обслуживанием в разных узлах) заключается в организации очередей в каждом сервере, который, таким образом, соответствует одноканальной системе массового обслуживания [18] с локальной бесконечной очередью. Для резервированного выполне-

ния запроса его резервные копии заносятся в несколько очередей разных серверов.

Резервированное выполнение запросов несколькими узлами, представляемыми одноканальными системами массового обслуживания с локальными очередями, приводит к повышению функциональной надежности вычислительного процесса при отказах и ошибках узлов. Вместе с тем резервирование запросов вызывает возрастание загрузки узлов, а это может привести к увеличению среднего времени ожидания и вероятности недопустимой задержки запросов в очередях. Задача оптимизации в кластерных системах с резервированным выполнением запросов в реальном времени поставлена и решена в работе [19].

В то же время резервированное выполнение запросов несколькими узлами с учетом стохастичности обслуживания потенциально может привести к увеличению вероятности своевременного выполнения запроса хотя бы от одного из узлов. Следует также отметить, что резервированное обслуживание запросов при увеличении загрузки узлов приводит к снижению максимально возможной интенсивности обслуживаемого потока запросов, не вызывающей нарушения стационарности процесса обслуживания в узлах.

1. Направления исследований

Для кластерных систем реального времени при направлении резервных копий запроса в очереди нескольких узлов (серверов) рассмотрим вычислительный процесс, при котором считывание результатов вычислений проводится в момент времени t, отсчитываемый после занесения запроса в очередь узлов с учетом времени обслуживания. Такая организация вычислительного процесса обусловливается, например, особенностью систем реального времени (в частности, управляющих систем), в которых результаты вычислений требуются к определенным моментам времени, и не стоит задача минимизации времени ожидания запросов, главное — их получение к заданному моменту времени.

Особенность рассматриваемой организации вычислительного процесса заключается в необходимости начала обслуживания запроса за время, не превышающее предельно допустимый порог ожидания t, после этого времени (с учетом задержки вычислений) начинается считывание результатов вычислений, которые должны быть получены хотя бы в одном из k узлов, задействованных в резервированных вычислениях. После превышения порога допустимого времени ожидания t резервные копии запроса, еще находящиеся в очередях, теряют свою актуальность и могут быть уничтожены. Уничтожение резервных копий запросов с просроченным временем ожидания позволит исключить непродуктивное выполнение резервных копий запросов, актуальность которых при ожидании в очереди сверх времени t теряется, позволяет снизить загрузку узлов и, соответственно, задержки в их очередях.

Выбор варианта диспетчеризации с уничтожением и без уничтожения резервных копий запросов, ожидающих обслуживания сверх допустимого времени, должен сопровождаться соответствующими расчетами средних задержек и вероятностей своевременного обслуживания резервированных запросов.

Модели массового обслуживания кластеров реального времени с резервированием вычислительного процесса без уничтожения резервных копий запросов, ожидающих сверх допустимого времени, предложены в работе [18]. При построении модели обслуживания с k-кратным резервированием выполнения запросов в работе [18] предполагается увеличение интенсивности потока обслуженных запросов в k раз относительно исходного входного потока запросов. Применение моделей резервированного обслуживания в системах с уничтожением резервных копий запросов, ожидающих в очереди сверх допустимого времени, как показано в работе [18], приводит к нижней оценке вероятности своевременного получения результатов. Однако оценка погрешности такого нижнего приближения в этой работе не приведена.

Цель исследований — повышение вероятности своевременного и безошибочного обслуживания в результате направления резервных копий запросов на обслуживание в несколько узлов кластера.

Для достижения поставленной цели ставится задача разработать модели, отражающие влияние уничтожения просроченных резервных копий запросов (для которых ожидание превышает допустимое время) на вероятность своевременного и безошибочного обслуживания запросов в кластере.

2. Объект и задачи исследования

В качестве объекта исследований рассматривается вычислительный кластер, объединяющий n идентичных компьютерных узлов (серверов), в каждом из которых организуется собственная очередь запросов, таким образом, каждый узел кластера соответствует одноканальной системе массового обслуживания [20] с локальной очередью.

Будем считать известными среднее время выполнения узлом запроса v, а также интенсивности входного потока Λ, потока отказов λ_0 и ошибок вычислений λ_1 узлов.

Поступающий в кластер запрос может быть распределен на обслуживание в любой компьютерный узел, для повышения надежности вычислительного процесса резервные копии запроса могут быть распределены на обслуживание в k узлов.

Для успешного обслуживания запроса необходимо, чтобы к моменту времени t было начато обслуживание его резервной копии хотя бы в одном из k узлов, задействованных в резервированных вычислениях. Резервные копии запроса, находя-

щиеся в очередях узлов к моменту времени t, теряют свою актуальность (не отвечают условиям своевременности) и могут уничтожаться. Для организации уничтожения просроченных копий резервных запросов в момент их поступления в очередь узла запускается таймер, отсчитывающий предельное время t нахождения запроса в очереди. Если к моменту срабатывания таймера запрос не начал обслуживаться (находится в очереди), то он уничтожается. В результате уничтожения просроченных резервных копий запросов удается снизить загруженность узлов, но при этом несколько усложняется диспетчеризация запросов.

В работе ставится задача построения моделей и сравнения эффективности кластеров при организации в узлах очередей с уничтожением и без уничтожения резервных копий запросов, ожидающих в очередях дольше предельно допустимого времени t.

Разрабатываемые модели направлены на оценку вероятности непревышения допустимого времени ожидания хотя бы в одном из узлов кластера, выделенных для обслуживания резервных копий запросов.

Для решения поставленной задачи сначала исследуем влияние предлагаемой организации резервированного обслуживания запросов на своевременность результатов без учета ошибок вычислений, а затем проанализируем дополнительное влияние возможности возникновения ошибок на своевременность резервированного обслуживания запросов.

3. Модели резервированного обслуживания запросов

При построении модели обслуживания будем пренебрегать потерями на диспетчеризацию, в том числе на уничтожение резервных копий запросов, ожидание которых в очередях превышает лимит времени t. Такое приближение допустимо при несущественном влиянии на замедление вычислительного процесса в сервере процесса уничтожения просроченных запросов в очереди, проводимых при диспетчеризации по таймеру без реализации межмашинного обмена между серверами после завершения выполнения копии запроса одним из них.

Резервированное обслуживание запроса считается успешно выполненным, если результаты требуемых вычислений получены к моменту времени t хотя бы в одном из k узлов, принимающих резервные копии запроса к обслуживанию.

Как показано в работе [18], резервированное обслуживание запросов без уничтожения просроченных в очереди резервных копий запросов сверх допустимого времени ожидания t приводит к увеличению интенсивности запросов в k раз.

В предположении независимости вычислительных процессов в разных узлах, представляемых системами массового обслуживания (СМО) типа M/M/1 [20] с бесконечными локальными очередями, вероятность того, что хотя бы в одном из k узлов, принимающих в очередь резервированный запрос, его ожидание меньше предельно допустимой задержки t, в соответствии с [18] вычислим как

$$R(t) = 1 - (1 - r)^k = \\ = 1 - \left(\frac{\Lambda \nu k}{n}\exp\left(-t\left(\nu^{-1} - \frac{\Lambda \nu k}{n}\right)\right)\right)^k, \qquad (1)$$

где r — вероятность непревышения времени ожидания в некотором узле установленного предела t, $r = 1 - (\nu\Lambda k/n)\exp(-t(\nu^{-1} - \Lambda k/n))$, а среднее время ожидания запросов вычисляется как

$$w = \int_0^\infty (1 - R(t))dt = \int_0^\infty \left(\frac{\Lambda \nu k}{n}\exp\left(-t\left(\nu^{-1} - \frac{\Lambda \nu k}{n}\right)\right)^k\right) dt. \qquad (2)$$

Формула (2) дает верхнюю (пессимистическую) оценку среднего времени ожидания, так как не учитывает возможность уменьшения загрузки в результате удаления из очередей резервных копий запросов, выполнение которых при ожидании сверх времени t становится неактуальным.

Для кластерной системы без резервирования запросов, узлы которой представляются СМО типа M/M/1 с бесконечной очередью, вероятность того, что время ожидания запросов в узле меньше предельно допустимого времени t, вычисляется как

$$r = 1 - \frac{\Lambda}{n}\nu\exp\left(-t_0\left(\nu^{-1} - \frac{\Lambda}{n}\right)\right), \qquad (3)$$

а среднее время ожидания запросов [20] как $w = (\Lambda\nu^2/n)/[1 - (\Lambda\nu/n)]$.

Рассмотрим вариант управления очередями, при котором реализуется уничтожение резервных копий запросов, находящихся в очередях сверх времени t. В результате такого уничтожения просроченных резервных копий интенсивность обслуживаемых запросов (загрузка узлов) вследствие k-кратного резервирования увеличивается не в k раз, как для варианта без уничтожения просроченных копий запросов, а в $I \leqslant k$ раз относительно обслуживания без резервирования ($I \geqslant 1$).

Коэффициент I определяется как математическое ожидание числа узлов, принимающих запрос к резервированному обслуживанию, для которых время ожидания запросов меньше допустимого значения t (в этих узлах копии ожидающих запросов не уничтожаются).

Значение предельно допустимого времени ожидания запроса в очереди t обусловлено требованиями реального времени и задается как константа в зависимости от особенностей прикладного процесса. Очевидно, что величина I зависит от интенсивности запросов и от ограничения допустимого времени ожидания t. Действительно, если время t велико (нет жестких требований, вызванных реальным временем обслуживания запросов), то вероятность

своевременного выполнения запросов в узлах повышается, в результате повышается и значение I. С уменьшением предельно допустимого времени ожидания t (ужесточение требований реального времени) вероятность своевременного вычисления в каждом узле и, соответственно, математическое ожидание числа узлов, своевременно выполнивших запрос за время t, уменьшается.

Таким образом, с уменьшением предельно допустимого времени ожидания t потенциально должна расти эффективность от уничтожения просроченных запросов в очередях.

Составим уравнение для вычисления коэффициента I увеличения интенсивности потока обслуженных запросов с учетом уничтожения просроченных копий запросов в очередях как

$$I = kr = k\left(1 - \frac{\Lambda\nu I}{n}\exp\left(-t\left(\nu^{-1} - \frac{\Lambda I}{n}\right)\right)\right),$$

где kr — математическое ожидание числа узлов, принимающих копии запроса к резервированному обслуживанию, при котором время ожидания в очереди меньше предельно допустимого значения t, а r — вероятность того, что ожидание в очереди узла меньше предельно допустимого значения t:

$$r = 1 - \frac{\Lambda\nu I}{n}\exp\left(-t\left(\nu^{-1} - \frac{\Lambda I}{n}\right)\right), \qquad (4)$$

при этом загрузка узла с учетом уничтожения в очередях просроченных запросов (в предположении сбалансированности загрузки всех n узлов) $\rho = \Lambda\nu I/n$.

Уравнение легко решается, например, в системе компьютерной математики MathCAD-15 с использованием встроенной функции *root*:

$$s := root\left(-I + \left[1 - \frac{\Lambda\nu I}{n}\exp\left(-t\left(\nu^{-1} - \frac{\Lambda I}{n}\right)\right)\right], I\right),$$

где s принимает значение искомой величины I.

После определения коэффициента увеличения загрузки I при k-кратном резервировании запросов в очередях узлов кластера вероятность непревышения допустимой задержки ожидания t хотя бы одним из k узлов, принимающих запрос в очереди, находим как

$$R(t) = 1 - \left(\left(\frac{\Lambda\nu I}{n}\right)\exp\left(-t\left(\nu^{-1} - \frac{\Lambda\nu I}{n}\right)\right)\right)^k. \qquad (5)$$

Среднее время ожидания резервированных запросов определим по следующей формуле:

$$w = \int_0^\infty (1 - R(t))dt = \int_0^\infty \left(\frac{\Lambda\nu I}{n}\exp\left(-t\left(\nu^{-1} - \frac{\Lambda\nu I}{n}\right)\right)^k\right) dt. \qquad (6)$$

В кластере, содержащем n одинаковых компьютерных узлов (серверов), вероятность своевременности, безошибочности и надежности выполнения запроса в некотором узле кластера определим как $b = rp$, где r — вероятность того, что время ожидания запросов в очереди некоторого узла меньше предельно допустимого значения t; $p = \exp(-\lambda(t + \nu))$ — вероятность того, что за период $t + \nu$ до считывания результатов, включающий время вычислений, ожидания запроса в очереди и его результата в выходном буфере, отказы и ошибки вычислений в рассматриваемом узле не возникают, при этом $\lambda = \lambda_0 + \lambda_1$ — суммарная интенсивность сбоев, отказов и ошибок узла.

При диспетчеризации очередей без уничтожения резервных копий запросов, ожидающих сверх допустимого времени t, вероятность r вычисляется по формуле (3), а с уничтожением копий — по формуле (4).

Вероятность своевременного получения безошибочных результатов хотя бы от одного из k узлов кластера, задействованных в резервированном выполнении запроса, определим как $P = 1 - (1 - b)^k$, где $b = rp$.

4. Расчет вероятности своевременного обслуживания резервированных запросов

Приведем расчет вероятности своевременного обслуживания запросов для вариантов с уничтожением и без уничтожения резервных копий запросов, ожидающих в очередях сверх установленного предельно допустимого срока t.

При расчетах будем считать, что $n = 20$ шт., $\nu = 0{,}1$ с. Расчет проведем для идеального случая безошибочности вычислений и безотказности узлов.

Зависимость увеличения загрузки узлов I от интенсивности запросов для варианта с уничтожением резервных копий запросов, находящихся в очередях сверх времени t, представлена на рис. 1. На рис. 1 при кратности резервирования запросов $k = 3$

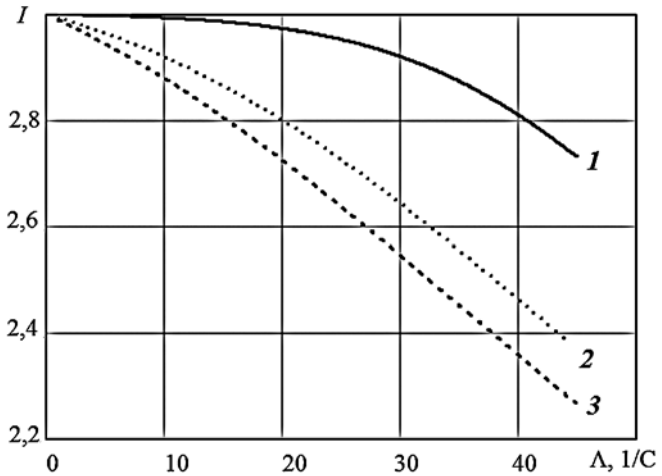

Рис. 1. Зависимость увеличения загрузки узлов I при уничтожении просроченных в очередях запросов

кривые 1, 2, 3 соответствуют предельно допустимому времени ожидания запросов $t = 0{,}5$; 0,2; 0,15 с.

Из представленных зависимостей видна эффективность снижения загрузки узлов в результате уничтожения просроченных резервных копий запросов. Анализируемая эффективность растет с увеличением интенсивности запросов и с уменьшением предельно допустимого времени ожидания t.

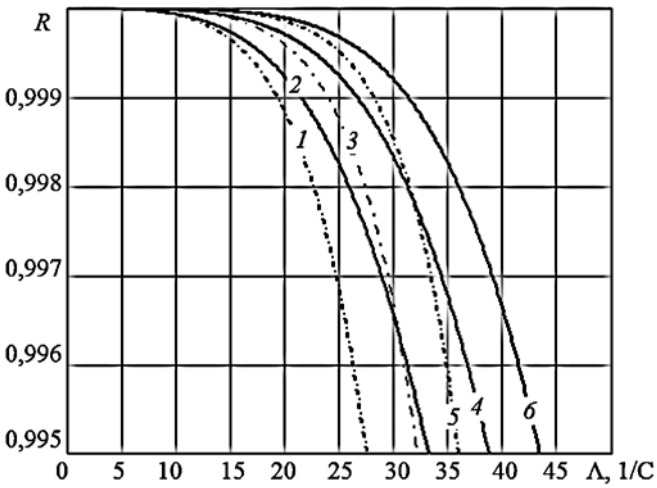

Рис. 2. Зависимость вероятности непревышения допустимой задержки ожидания от интенсивности входного потока

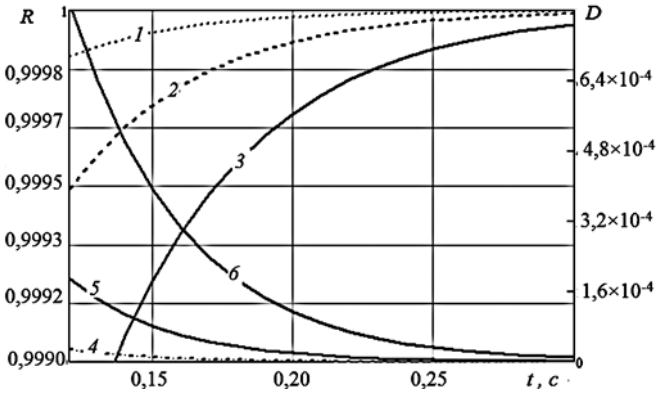

Рис. 3. Зависимость вероятности непревышения допустимой задержки ожидания от значения этой задержки

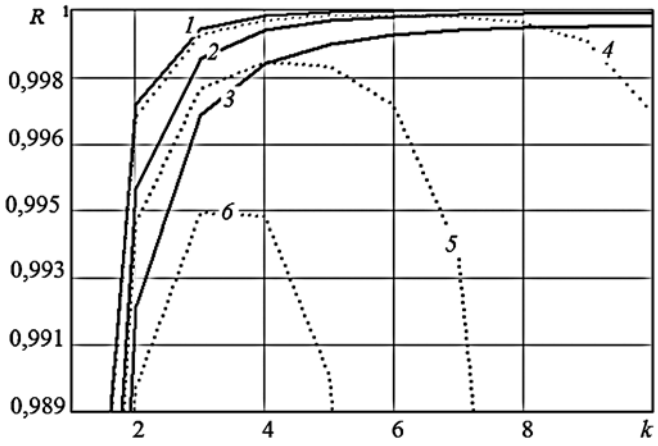

Рис. 4. Вероятность непревышения допустимой задержки ожидания для диспетчеризации с уничтожением и без уничтожения просроченных в очередях запросов

Зависимость вероятности непревышения допустимой задержки ожидания t резервированных вычислений хотя бы одним из k узлов, выполняющих запрос, от интенсивности входного потока Λ представлена на рис. 2, а от значения допустимой задержки t — на рис. 3. Расчеты выполнены при кратности резервирования вычислений $k = 3$. На рис. 2 кривые *1* и *2* соответствуют предельно допустимой задержке $t = 0{,}15$ с в случае управления очередями при уничтожении просроченных к моменту времени t резервных копий запросов и без такого уничтожения. Кривые *3* и *4* соответствуют вариантам диспетчеризации с уничтожением и без уничтожения просроченных копий запросов при $t = 0{,}2$ с, а кривые *5* и *6* — при $t = 0{,}25$ с.

При организации управления очередями с уничтожением просроченных резервных копий запросов на рис. 3 кривыми *1—3* представлена зависимость вероятности своевременного обслуживания запросов R, вычисляемой по формуле (5), от значения допустимой задержки t при интенсивности запросов $\Lambda = 10$; 15; 20 1/с соответственно. На рис. 3 кривые *4—6* соответствуют разнице D вероятностей своевременности резервированного обслуживания запросов R с уничтожением и без уничтожения просроченных к моменту времени t резервных копий запросов (вычисляемых соответственно по формулам (5) и (1)).

Из графиков видна высокая эффективность управления очередями с уничтожением просроченных резервных копий запросов. Причем эффективность уничтожения запросов в очередях после потери их актуальности возрастает при уменьшении допустимого предела времени t ожидания. При этом с ростом интенсивности потока запросов эффективность уничтожения резервных копий запросов, ожидающих в очереди сверх допустимого времени, повышается.

На рис. 4 приведены результаты расчетов вероятностей непревышения допустимого времени ожидания t в зависимости от кратности резервирования запросов k. Расчеты выполнены для вариантов диспетчеризации с уничтожением и без уничтожения просроченных в очередях резервных копий запросов. Расчеты проведены для $v = 0{,}1$ с, $n = 20$ шт., $t = 0{,}12$ с. Кривые *1*, *2*, *3* соответствуют $\Lambda = 15$, 20, 25 1/с в случае уничтожения просроченных резервных копий запросов в очередях, а кривые *4*, *5*, *6* — без их уничтожения. Результаты расчетов показывают влияние уничтожения просроченных в очередях резервных копий запросов реального времени на повышение эффективности их обслуживания.

На рис. 5 приведены графики вероятностей непревышения допустимого времени ожидания t для кластеров без резервирования и с резервированием вычислений в зависимости от кратности их резер-

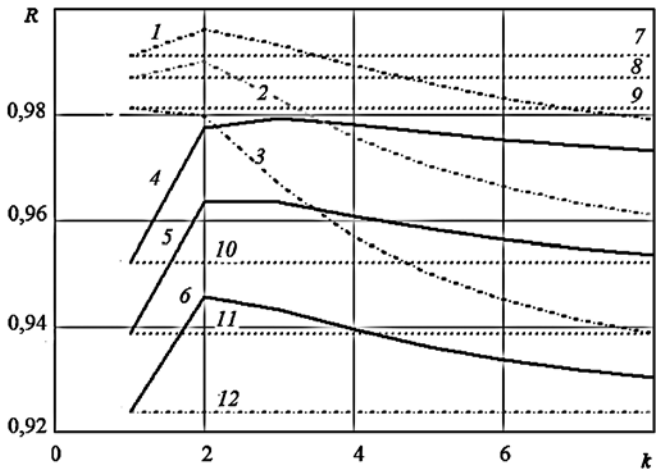

Рис. 5. Вероятности не превышения допустимого времени ожидания *t* для кластеров без резервирования и с резервированием вычислений

вирования *k*. Расчеты выполнены с учетом уничтожения просроченных в очереди резервных копий запросов для различных значений интенсивности запросов Λ и допустимого времени ожидания в очереди *t*.

При интенсивности входного потока Λ = 60, 70, 80 1/с кривые *1—3* соответствуют резервированным вычислениям для *t* = 0,25 с, а кривые *4—6* — для *t* = 0,5 с. Кривые *7—9* на рис. 5 соответствуют нерезервированному выполнению запросов для *t* = 0,25 с, а кривые *10—12* — для *t* = 0,5 с.

Приведенные зависимости показывают существование области эффективности резервирования вычислений (обслуживания с резервированием запросов).

Таким образом, для систем реального масштаба времени, предусматривающих уничтожение просроченных в очередях резервных копий запросов, существует область эффективности резервированного обслуживания запросов. При этом оптимальная кратность резервирования и целесообразность резервирования запросов зависит от их интенсивности и от ограничения допустимого времени ожидания *t*. Таким образом, возникает необходимость оптимизации резервированных вычислений в кластерных системах реального времени.

5. Расчет среднего времени ожидания при резервировании вычислений

В предыдущем разделе показано, что резервированное выполнение запросов в нескольких узлах кластера позволяет существенно повысить вероятность своевременности вычислений, когда требуется завершение вычислений к установленному сроку хотя бы в одном из *k* узлов, выполняющих резервные копии запросов. Проанализируем теперь влияние резервирования выполнения запросов на среднее время их ожидания в очередях с уничтожением и без уничтожения резервных копий запросов, ожидающих в очередях сверх установленного предельно допустимого срока.

Для идеального случая безошибочности вычислений и безотказности узлов расчет среднего времени ожидания в очередях при резервированном обслуживании запросов *w* проведем по формулам (6), (2) соответственно, когда просроченные резервные копии запросов в очередях уничтожаются и не уничтожаются. При расчетах примем число узлов в кластере *n* = 20 шт., а среднее время выполнения запроса ν = 0,1 с.

Зависимости среднего времени ожидания запросов *w* (рис. 6) от интенсивности входного потока Λ для диспетчеризации очередей с уничтожением и без уничтожения просроченных резервных копий запросов при кратности резервирования вычислений *k* = 4, 3 и 2 представлены соответственно парами кривых *1*, *2*; *3*, *4* и *5*, *6*. Среднему времени ожидания в системах без резервирования обслужи-

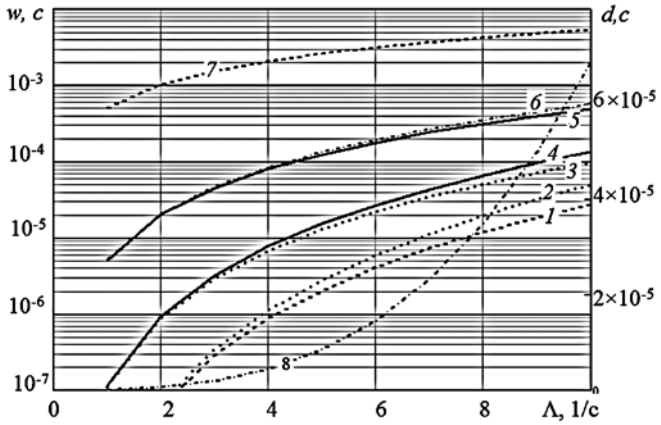

Рис. 6. Зависимости среднего времени ожидания запросов *w* от интенсивности входного потока Λ

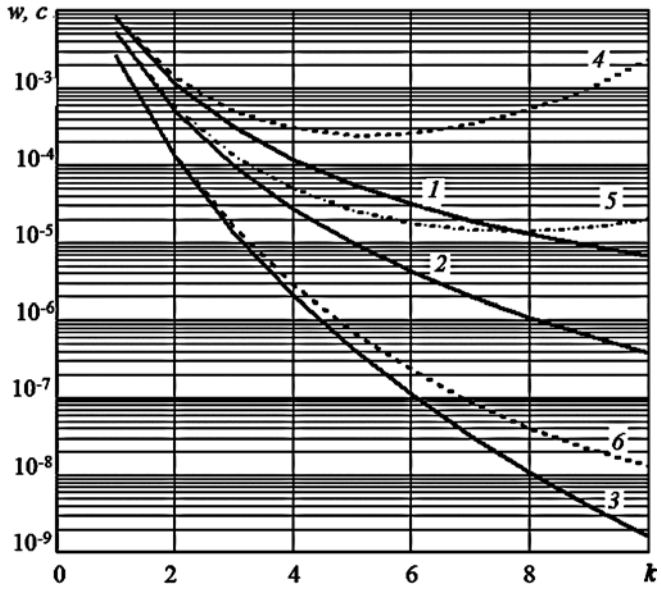

Рис. 7. Зависимости среднего времени ожидания запросов *w* от кратности резервирования вычислений *k*

вания запросов соответствует кривая *7*. Кривая *8* уточняет разницу *d* среднего времени ожидания для диспетчеризации без уничтожения и с уничтожением просроченных резервных копий запросов в очередях при кратности резервирования вычислений $k = 2$ (уточнение вызвано близостью кривых *5*, *6*).

Зависимости среднего времени ожидания запросов *w* от кратности резервирования вычислений представлены на рис. 7 кривыми *1—3* для интенсивности входного потока $\Lambda = 15, 10, 5$ 1/с в случае диспетчеризации с уничтожением просроченных копий запросов, а кривыми *4—6* — без уничтожения копий.

Представленные на рис. 6 и 7 зависимости показывают эффективность снижения среднего времени ожидания в результате резервированного выполнения запросов несколькими (*k*) узлами кластера, причем эта эффективность повышается при уничтожении в очередях резервных копий запросов, ожидающих сверх допустимого лимита времени *t*.

Заключение

Для вычислительной системы кластерной архитектуры предложены модели обслуживания при диспетчеризации очередей с уничтожением и без уничтожения резервных копий запросов, срок нахождения которых в очередях превысил допустимое ожидание.

Проанализировано влияние кратности резервирования запросов на вероятность их своевременного и безошибочного обслуживания с уничтожением и без уничтожения просроченных в очередях резервных копий запросов.

Показано значительное повышение вероятности своевременного обслуживания в результате резервирования запросов. Сделан вывод, что диспетчеризация очередей с уничтожением просроченных в очереди резервных копий запросов позволяет существенно повысить эффективность обслуживания запросов.

Доказано существование области эффективного резервирования запросов.

Установлено существование оптимальной кратности резервирования запросов, при котором достигается максимум вероятности своевременного выполнения запросов, причем оптимальная кратность резервирования зависит от интенсивности запросов и их допустимой задержки в очереди.

Проведенные исследования эффективности предлагаемой организации резервированного обслуживания запросов ограничены аналитическим моделированием при представлении процессов обслуживания запросов СМО типа M/M/1. Подтверждение эффективности предлагаемых решений в более общих случаях требует дальнейших исследований, в том числе основанных на натурных испытаниях в кластерах, выполняющих реальные прикладные задачи.

Список литературы

1. **Aysan H.** Fault-tolerance strategies and probabilistic guarantees for real-time systems Mälardalen University, Västerås, Sweden, 2012. 190 p.

2. **Kopetz H.** Real-Time Systems: Design Principles for Distributed Embedded Applications. Springer, 2011. 396 p.

3. **Koren I.** Fault tolerant systems. San Francisco: Morgan Kaufrnann publications. 2009. 378 p.

4. **Черкесов Г. Н.** Живучесть и отказобезопасность ответственных технических систем // Проектирование и технология электронных средств. 2015. № 1. С. 15—24.

5. **Алиев Т. И., Муравьева-Витковская Л. А.** Приоритетные стратегии управления трафиком в мультисервисных компьютерных сетях // Известия вузов. Приборостроение. 2011. Т. 54, № 6. С. 44—48.

6. **Щеглов А. Ю., Щеглов К. А.** Возможности методов резервирования для повышения уровня интегрированной информационно-эксплуатационной безопасности современных информационных систем // Информационные технологии. Том 21, № 7. 2015. С. 521—527.

7. **Гатчин Ю. А., Жаринов И. О., Коробейников А. Г.** Математические модели оценки инфраструктуры системы защиты информации на предприятии // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 2 (78). С. 92—95.

8. **Богатырев В. А.** Надежность вариантов размещения функциональных ресурсов в однородных вычислительных сетях // Электронное моделирование. 1997. № 3. С. 21—29.

9. **Богатырев В. А.** К распределению функциональных ресурсов в отказоустойчивых многомашинных вычислительных системах // Приборы и системы. Управление, контроль, диагностика. 2001. № 12. С. 1—5.

10. **Богатырев В. А., Богатырев С. В.** Объединение резервированных серверов в кластеры высоконадежной компьютерной системы // Информационные технологии. 2009. № 6. С. 41—47.

11. **Богатырев В. А.** Мультипроцессорные системы с динамическим перераспределением запросов через общую магистраль // Известия вузов. Приборостроение. 1985. № 3. С. 33—38.

12. **Богатырев В. А., Богатырев А. В., Богатырев С. В.** Перераспределение запросов между вычислительными кластерами при их деградации // Известия вузов. Приборостроение. 2014. Т. 57, № 9. С. 54—58.

13. **Богатырев В. А., Богатырев А. В., Богатырев С. В.** Оценка надежности выполнения кластерами запросов реального времени // Известия вузов. Приборостроение. 2014. Т. 57, № 4. С. 46—48.

14. **Богатырев В. А., Богатырев А. В., Богатырев С. В.** Оценка своевременности выполнения критических запросов в двухуровневых кластерах // Научно-технический вестник информационных технологий, механики и оптики. 2014. № 2 (90). С. 177—179.

15. **Богатырев В. А., Богатырев А. В., Богатырев С. В.** Оптимизация перераспределения нагрузки в кластерах при изменяющейся активности источников запросов // Известия вузов. Приборостроение. 2014. Т. 57, № 4. С. 41—45.

16. **Lee M. H., Dudin A. N., Klimenok V. I.** The SM/V/N queueing system with broadcasting service // Math. Probl. in Engineer. 2006. V. 2006. Article ID 98171. 18 p.

17. **Дудин А. Н., Сунь Б.** Многолинейная ненадежная система с управляемым широковещательным обслуживанием // Автоматика и телемеханика. 2009. Т. 70, № 12. С 147—160.

18. **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.

19. **Богатырев В. А., Богатырев А. В.** Оптимизация резервированного распределения запросов в кластерных системах реального времени // Информационные технологии. 2015. Том 21, № 7. С. 495—502.

20. **Вишневский В. М.** Теоретические основы проектирования компьютерных сетей. М.: Техносфера, 2003. 512 с.

V. A. Bogatyrev, Professor, **A. V. Bogatyrev,** Graduate Student
Saint-Petersburg National Research University of Information Technologies,
Mechanics and Optics (University ITMO) Kronverkskiy Prospekt 49, 197101 St Petersburg, Russia,
e-mail: vladimir.bogatyrev@gmail.com

The Model of Redundant Service Requests Real-Time in a Computer Cluster

Considered computing cluster architecture, operating in real time. The aim of the research is increasing the probability of timely and error-free service. This goal is achieved due to the fact that the system provides the direction of backup service requests to multiple cluster nodes with the possibility of destruction after exceeding a certain threshold expectations. Consideration of options for dispatching with destruction and without destruction of backup copies of requests waiting in the local queues of nodes in excess of the allowable time. The dependence of the probability of timely and accurate service requests from the multiplicity of backup copies of requests to the service at different nodes. Shown a significant increase in the probability of timely services as a result of reservation requests. It is shown that the scheduling queue with the destruction of expired queued backup requests can significantly improve the efficiency of service requests. The existence of effective reservation requests depending on scheduling the rate and permissible delay in queue.

Keywords: *model, reliability, a redundant service requests, real-time cluster scheduling*

References

1. **Aysan H.** *Fault-tolerance strategies and probabilistic guarantees for realtime systems*, Västerås, Sweden, Märdalen University, 2012, 190 p.

2. **Kopetz H.** *Real-Time Systems: Design Princi ples for Distributed Embedded*, Applications Springer, 201l, 396 p.

3. **Koren I.** *Fault tolerant systems,* San Francisco, Morgan Kaufmann publications, 2009, 378 p.

4. **Cherkesov G. N.** Zhivuchest' i otkazobezopasnost' otvetstvennyh tehnicheskih sistem, *Proektirovanie i tehnologija jelektronnyh sredstv*, 2015, no. 1, pp. 15—24 (Cherkesov G. N. The durability and fail-safety critical technical systems, Engineering and technology of electronic means 2015, no. 1, pp. 15—24).

5. **Aliev T. I., Murav'eva-Vitkovskaja L. A.** Prioritetnye strategii upravlenija trafikom v mul'tiservisnyh komp'juternyh setjah, *Izvestija Vuzov. Priborostroenie*, 2011, vol. 54, no. 6, pp. 44—48. (Aliev T. I., Muravyova-Vitkovskaya L. A. Priority strategies traffic management in multi-service computer networks, Izvestiya Vysshikh Uchebnykh Zavedeniy, Priborostroenie, 2011, vol. 54, no. 6, pp. 44—48).

6. **Shheglov A. Ju., Shheglov K. A.** Vozmozhnosti metodov rezervirovanija dlja povyshenija urovnja integrirovannoj informacionno-ekspluatacionnoj bezopasnosti sovremennyh informacionnyh sistem, *Informacionnye tehnologii*, 2015, no. 7, vol. 21, pp. 521—527 (Shcheglov A. Yu., Shcheglov K. A., Possible methods of redundancy to increase the level of integrated information and operational security of modern information systems, Information technology, 2015, no. 7, vol. 21, pp. 521—527).

7. **Gatchin Ju. A., Zharinov I. O., Korobejnikov A. G.** Matematicheskie modeli ocenki infrastruktury sistemy zashhity informacii na predprijatii, *Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki*, 2012, no. 2 (78), pp. 92—95 (Hatchin J. A., Zharinov I. O., Korobeinikov A. G. Mathematical models for estimating system infrastructure informatio n protection at the enterprise, Scientific and technical Bulletin of information technologies, mechanics and optic, 2012, no. 2 (78), pp. 92—95).

8. **Bogatyrev V. A.** Nadezhnost' variantov razmeshhenija funkcional'nyh resursov v odnorodnyh vychislitel'nyh setjah, *Jelektronnoe modelirovanie*, 1997, no. 3, pp. 21—29 (Bogatyrev V. A. Yhe reliability properties of the functional resources in a homogeneous computing networks, Engineering Simulation, 1997, no. 3, pp. 21—29).

9. **Bogatyrev V. A.** K raspredeleniju funkcional'nyh resursov v otkazoustojchivyh mnogomashinnyh vychislitel'nyh sistemah, *Pribory i sistemy. Upravlenie, kontrol', diagnostika,* 2001, no. 12, pp. 1—5 (Bogatyrev V. A. the distribution of functional resources in a fault-tolerant multicomputer systems, Devices and systems. Management, control, diagnostics, 2001, no. 12, pp. 1—5).

10. **Bogatyrev V. A., Bogatyrev S. V.** Ob#edinenie rezervirovannyh serverov v klastery vysokonadezhnoj komp'juternoj sistemy, *Informacionnye tehnologii*, 2009, no. 6, pp. 41—47. (Bogatyrev V. A., Bogatyrev S. V. Association of the redundant servers in the cluster of highly reliable computer systems, Inforrnation technologies, 2009, no. 6, pp, 41—47.)

11. **Bogatyrev V. A.** Mul'tiprocessornye sistemy s dinamicheskim pereraspredeleniem zaprosov cherez obshhuju magistral, *Izvestiya Vysshikh Uchebnykh Zavedeniy. Priborostroenie,* 1985, no. 3, pp. 33—38 (Bogatyrev V. A. Multiprocessor system with dynamic reallocation requests through a common line, Izvestiya Vysshikh Uchebnykh Zavedeniy. Priborostroenie, 1985, no. 3, pp. 33—38.)

12. **Bogatyrev V. A., Bogatyrev A. V., Bogatyrev S. V.** Pereraspredelenie zaprosov mezhdu vychislitel'nymi klasterami pri ih degradacii, *Izvestija Vuzov Priborostroenie*, 2014, vol. 57, no. 9, pp. 54—58 (Bogatyrev V. A., Bogatyrev V. A., Bogatyrev S. V. the Redistribution of requests between the compute clusters at their degradation, *Izvestiya Vysshikh Uchebnykh Zavedeniy. Priborostroenie*, 2014, vol. 57, no. 9, pp. 54—58.

13. **Bogatyrev V. A., Bogatyrev A. V., Bogatyrev S. V.** Ocenka nadezhnosti vypolnenija klasterami zaprosov real'nogo vremeni, *Izvestija Vuzov Priborostroenie*, 2014, vol. 57, no. 4, pp. 46—48 (Bogatyrev V. A., Bogatyrev A. V., Bogatyrev S. V. the Estimation of the reliability of clusters perform real-time queries, *Izvestiya Vysshikh Uchebnykh Zavedeniy. Priborostroenie*, 2014, vol. 57, no. 4, pp. 46—48).

14. **Bogatyrev V. A., Bogatyrev A. V., Bogatyrev S. V.** Ocenka svoevremennosti vypolnenija kriticheskih zaprosov v dvuhurovnevyh klasterah, *Nauchno-tehnicheskii vestnik informacionnyh tehnologij, mehaniki i optiki*, 2014, no. 2 (90), pp. 177—179, (Bogatyrev V. A., Bogatyrev A. V., Bogatyrev S. V., Assessment of the timeliness of implementation of the critical queries in two-level clusters, *Scientific-technical Bulletin of information technologies, mechanics and optics*, 2014, no. 2 (90), pp. 177—179).

15. **Bogatyrev V. A., Bogatyrev A. V., Bogatyrev S. V.** Optimizacija pereraspredelenija nagruzki v klasterah pri izmenjajushhejsja aktivnosti istochnikov zaprosov, *Izvestija Vuzov Priborostroenie*, 2014, vol. 57, no. 4, pp. 41—45 (Bogatyrev V. A., Bogatyrev A. V., Bogatyrev S. V. Optimization of load redistribution in clusters of variable activity of the source of the query, *Izvestiya Vysshikh Uchebnykh Zavedeniy. Priborostroenie*, 2014, vol. 57, no. 4, pp. 41—45).

16. **Lee M. H., Dudin A. N., Klimenok V. I.** The SM/V/N queueing system with broadcasting service, *Math. Probl. in Engineer*, 2006, vol. 2006, Article ID 98171, 18 p.

17. **Dudin A. N., Sun' B.** Mnogolinejnaja nenadezhnaja sistema s upravljaemym shirokoveshhatel'nym obsluzhivaniem, *Avtomatika i telemehanika*, 2009, vol. 70, no. 12, pp. 147—160. (Dudin A. N., Sun B. unreliable Multi-line system with controllable broadcasting service, Automation and remote control, 2009, vol. 70, no. 12, pp. 147—160).

18. **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.

19. **Bogatyrev V. A., Bogatyrev A. V.** Optimizacija rezervirovannogo raspredelenija zaprosov v klasternyh sistemah real'nogo vremeni, *Informacionnye tehnologii*, 2015, no. 7, vol. 21, pp. 495—502 (Bogatyrev V. A., Bogatyrev A. V., Optimization of the redundant distribution of requests in a clustered real-time systems, Information technology, 2015, vol. 21, no. 7, pp. 495—502).

20. **Vishnevskij V. M.** *Teoreticheskie osnovy proektirovanija komp'juternyh setej,* Moscow, Tehnosfera, 2003, 512 p. (Vishnevsky V. M. Theoretical bases of design of computer networks. Moscow, Technosphere, 2003, 512 p.).