УДК 519.872

А.И. Мартышкин

РАЗРАБОТКА И ИССЛЕДОВАНИЕ РАЗОМКНУТЫХ МОДЕЛЕЙ ПОДСИСТЕМЫ «ПРОЦЕССОР-ПАМЯТЬ» МНОГОПРОЦЕССОРНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ АРХИТЕКТУР UMA И NUMA

На моделях исследованы варианты реализации подсистемы «процессор-память» многопроцессорных систем с архитектурой UMA и NUMA. Математические модели рассматривались с использованием аппарата теории массового обслуживания. Проведен анализ влияния различных факторов на рассматриваемые модели. В ходе исследований был проведен сравнительный анализ систем с архитектурой UMA и NUMA. Были выявлены их достоинства и недостатки. В заключении сделаны соответствующие выводы.

Ключевые слова: *многопроцессорная система, архитектура, модель, производительность, система массового обслуживания.*

Введение. Повышение производительности вычислительных систем непосредственно связано с увеличением быстродействия и емкости памяти. Несмотря на стремительный рост пропускной способности оперативной памяти, наблюдающийся в последние годы, разрыв "*CPU-Memory*" не сокращается, а наоборот – увеличивается.

Целью работы являются исследование архитектуры подсистемы «процессор-память» современных вычислительных систем, расчет и сравнение полученных характеристик, а также вывод о влиянии конфликтов из-за доступа к общим ресурсам на общую производительность системы в целом.

Объектами исследования данной работы являются подсистема «процессор-память» многопроцессорных вычислительных систем, существующие разновидности архитектуры построения данной подсистемы.

Для исследования и анализа вычислительных систем всё чаще применяют элементы теории массового обслуживания, а именно системы и сети массового обслуживания (СМО и СеМО). Любой блок, любой модуль системы можно представить в виде системы массового обслуживания, а совокупность устройств вычислительной системы представляется сетью массового обслуживания и довольно легко поддается исследованию при минимальных финансовых затратах – не требуется строить реальную систему, достаточно её модели [1].

Теоретическая часть. Данная статья носит исследовательский характер. Для решения поставленной задачи была проанализирована литература, описывающая исследования в данной предметной области с целью поиска нерешенных проблем. Были проанализированы литературные источники [2, 3]. Однако ряд вопросов, связанных с исследованием подсистемы «процессор-память», не нашел должного отражения.

Характеристики открытой сети массового обслуживания. Открытые сети массового обслуживания довольно хорошо изучены и подробно описаны в источниках [4-7]. В данной работе приведем часть выражений и формул, которые используются в созданном программном комплексе [8], необходимом для исследования разработанных моделей.

Если заданы параметры сети, то можно определить следующие характеристики каждой СМО и сети в целом [4, 7, 9]: среднюю длину очереди заявок в i-й СМО – l_i и в сети – L; среднее число заявок, пребывающих в i-й СМО – m_i и в сети – M; среднее время ожидания обслуживания заявки i-й СМО – ω_i и в сети – W; среднее время пребывания заявки в i-й СМО – u_i и в сети – U.

Среднее время ожидания заявки в очереди для одноканальной СМО равно частному от деления средней длины очереди l_i на интенсивность λ_i входящего в i-й СМО потока [4, 5, 7]:

$$\omega_i = \frac{l_i}{\lambda_i} = \frac{\upsilon_i \cdot \rho_i}{1-\rho_i}. \qquad (1)$$

Для многоканальной СМО

$$\omega_i = \frac{l_i}{\lambda_i} = \frac{\upsilon_i \cdot \beta_i^{k_i}}{k_i! k_i \left(1 - \beta_i / k_i\right)^2} p_{0i} \cdot \qquad (2)$$

Среднее время пребывания заявки в системе определяется средней задержкой её в очереди и временем обслуживания в *i*-й СМО.

Для одноканальной СМО [4, 5, 7]:

$$u_i = \omega_i + \upsilon_i = \frac{\upsilon_i}{1-\rho_i}; \qquad (3)$$

для многоканальной:

$$u_i = \omega_i + \upsilon_i = \frac{\upsilon_i \cdot \beta_i^{k_i}}{k_i! k_i \left(1 - \beta_i / k_i\right)^2} p_{0i} + \upsilon_i . \qquad (4)$$

На основании полученных характеристик отдельных СМО определяют характеристики сети в целом.

Среднее число заявок, ожидающих обслуживания в сети (т.е. среднее число заявок в очередях сети):

$$L = \sum_{i=1}^{n} l_i . \qquad (5)$$

Среднее число заявок, пребывающих в сети:

$$M = \sum_{i=1}^{n} m_i . \qquad (6)$$

Поскольку каждая заявка может получить обслуживание в *i*-й СМО в среднем α раз, то время ожидания обслуживания и время пребывания её в системе увеличится в α раз. Среднее время ожидания заявки в очередях сети:

$$W = \sum_{i=1}^{n} \alpha_i \omega_i , \qquad (7)$$

а время пребывания:

$$U = \sum_{i=1}^{n} \alpha_i u_i . \qquad (8)$$

Разработка модели подсистемы «процессор-память» архитектуры *UMA*. Структура архитектуры *UMA* (*Unified Memory Access*) и граф передач представлены на рисунке 1, а, б.

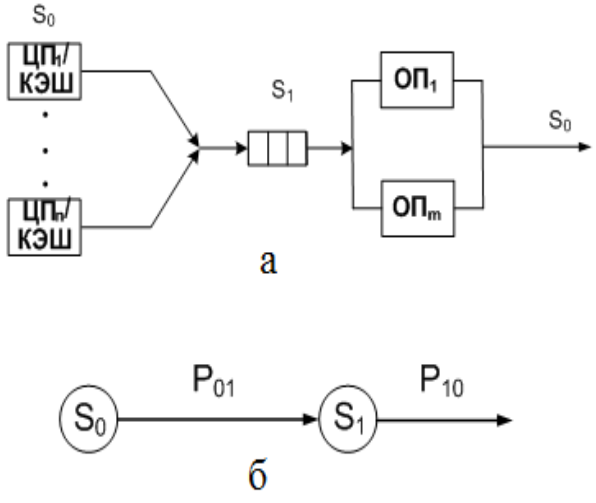


Рисунок 1 – Структура (а) и граф передач (б) *UMA*-системы

На рисунке: $ЦП_n$/КЭШ – *n* процессорных узлов; $ОП_m$ – *m* модулей памяти; S_0 – источник заявок, представляющий *n* процессорных узлов; S_1 – *m* модулей памяти в виде многоканального прибора обслуживания.

Считается, что приложения формируют простейшие потоки запросов, а времена обслуживания подчиняются экспоненциальному закону. Это распределение позволит получить результаты заведомо хуже реальных значений, что, в свою очередь, позволит сделать оценку полученных результатов сверху.

Считается, что память имеет единое адресное пространство, причём контроллеры памяти содержат в своем составе буферные регистры для хранения данных, записываемых в память или читаемых из памяти, имеющие объем, достаточный для того, чтобы заявки не получали отказ в обслуживании.

Определение исходных данных для моделирования. Для исследования данной архитектуры возьмем процессор семейства *Intel Pentium* IV. Объём кэша *L*2 процессоров серии *Pentium* 4 500 – 1 Мб. Процессоры *Intel Pentium* 4 серии 500 имеют частоту шины 200 МГц (длительность такта Т=5нс), обозначаемой с учетом технологии Quad Pumped Bus как 800 МГц. Такая шина за такт может передать 4 готовых к передаче 64-разрядных слова. Но если готово к передаче только одно слово, то и оно будет передано за 1 такт.

Частота синхронизации процессора взята равной *Intel Pentium* 4 500 – 2800 МГц. Предполагается, что обращение в память будет производиться каждый такт. В КЭШ память данных согласно статистике попадает не менее 99 % запросов. Поэтому, исходя из частоты процессора, равной 2,8 ГГц, получаем, что частота запросов в оперативную память составит $\lambda = 2{,}8 \cdot 1\% = 0{,}028$ заявки/нс.

При исследовании влияния производительности процессора на работу подсистемы «процессор-память» рассмотрены процессоры *Intel Pentium* 4 500 – 571 с частотой синхронизации 2,8-3,8 ГГц.

В расчетах используем модули памяти *DDR* с частотой 200 МГц. По причине стандарта *DDR* шина маркируется как 400 МГц. Это означает, что за один такт длиной 5 нс шина памяти может передать 2 готовых 64-битных слова [10].

Тайминги памяти *CL–RCD–RP* согласно данным микросхемы равны 15 нс каждый (3-3-3).

Наилучший случай происходит, когда обращение происходит к открытой строке модуля памяти. При этом необходимо подать только

сигнал *CAS*, и модуль памяти подготовит данные через интервал *CL* (15 нс). По статистике такая ситуация происходит в 55 % случаев.

Хуже случай, когда данное находится в другой строке. При этом подаются сигналы *RAS* (активация строки) и через время *RCD–CAS* (активация столбца). Таким образом, модулю памяти потребуется время, равное *CL+RCD* (30 нс), для подготовки данных. Этот вариант получается с вероятностью 40 %.

Наихудший случай, когда данные находятся на неактивной странице. В этом случае текущая страница должна быть закрыта. На подготовку данных модулю памяти требуется время, равное *CL+ RCD+RPT* (45 нс). На этот вариант приходится 5 % случаев [10].

Найдем среднее время работы модуля ОП: $V_{ОП} = 0{,}55 \cdot 15 + 0{,}4 \cdot 30 + 0{,}05 \cdot 45 = 22{,}5$ (нс).

Рассчитаем, сколько требуется времени шине для передачи контроллеру 32-битного адреса. Разрядность шины – 64 бита. То есть 32-битный адрес будет передан за 1 такт: $1 \cdot 5 = 5$ (нс).

Для передачи 64-битного слова от памяти через контроллер к процессору требуется 1 такт шины «память-контроллер» и 1 такт шины «контроллер-процессор»: $1 \cdot 5 + 1 \cdot 5 = 10$ (нс).

Просуммировав полученные значения, получим среднее время обращения к памяти: 22,5 + 5 + 10 = 37,5 (нс).

Итак, получено время обращения к памяти одного процессора.

Анализ влияния эффективности кэш-памяти на реальную пропускную способность подсистемы «процессор-память». Исходные данные: число обслуживающих каналов (модулей ОП) в СМО К = 12; число источников нагрузки (процессоров) ЦП = 4; время обслуживания заявок одним каналом (модулем ОП) ν=37,5 нс.

Интенсивность потока запросов при моделировании изменялась следующим образом: вероятность кэш-промаха, % · (частота синхронизации процессора). Диапазон вероятности кэш-промаха – 0,5-3 %. Результаты моделирования приведены на рисунке 2.

При попадании в кэш 99 % запросов средняя длина очереди l практически равна 0, время отклика памяти состоит из времени доступа к памяти *u*=37,507 нс.

При потоке заявок 0,070 запроса/нс (величина кэш-промаха составляет 2,5 %) длина очереди *l*=3,974 заявки, время ожидания заявки в очереди (ω=14,098 нс) и время пребывания заявки в СМО (время отклика памяти) *u*=51,598 нс), т.е. возрастает на 28 %.

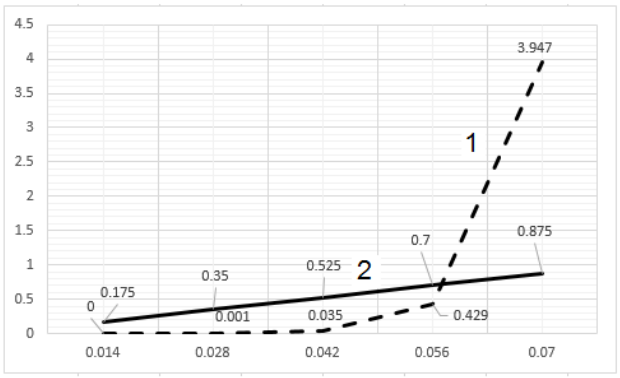


Рисунок 2 – Зависимость коэффициента загрузки (1) и длины очереди перед устройством (2) от вероятности кэш-промаха

При кэш-промахе 3 % система оказывается перегружена (ρ =1,05), так как снижается пропускная способность подсистемы «процессор-память».

При увеличении эффективности кэш-памяти в 2 раза (число попаданий в кэш составляет 99,5 %) в очереди отсутствуют заявки, время ожидания в очереди равно 0.

Анализ влияния числа процессорных узлов на реальную пропускную способность подсистемы «процессор-память». Исходные данные: число обслуживающих каналов (модулей ОП) в СМО К = 8; число источников нагрузки (процессоров) ЦП= 2-8; время обслуживания заявок одним каналом (модулем ОП) ν =37,5 нс; интенсивность потока запросов λ=0,028 запроса/нс. Результаты моделирования приведены на рисунке 3.

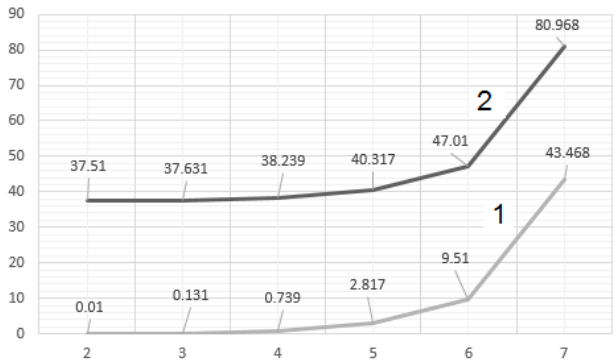


Рисунок 3 – Зависимость времени ожидания в очереди (1) и времени ответа памяти (2) от числа процессорных узлов

При ЦП =2-5 в исследуемой системе длина очереди *l*<1 (от 0,001 до 0,394 заявок), время ожидания в очереди – от 0,01 до 2,817 нс.

При ЦП =6-7 число заявок в очереди достигает 8,52 заявки, время ожидания в очереди увеличивается до 43,468 нс, время ответа памяти равно 80,968, что в 2,2 раза превышает значение при ЦП =2.

При ЦП =8 в системе наблюдается перегрузка ρ=1,05.

Анализ влияния числа модулей памяти на реальную пропускную способность подсистемы «процессор-память». Исходные данные: число обслуживающих каналов (модулей ОП) в СМО К = 4-10; число источников нагрузки (процессоров) М= 4; время обслуживания заявок одним каналом (модулем ОП) ν=37,5 нс; интенсивность потока запросов λ=0,028 запроса/нс. Результаты исследования приведены на рисунке 4.

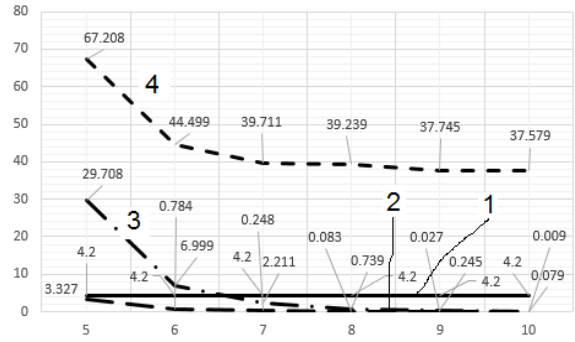


Рисунок 4 – Зависимость среднего числа занятых каналов (1), длины очереди перед устройством (2), времени ожидания в очереди (3) и времени ответа памяти (4) от числа модулей памяти

Как видно из рисунка, среднее число занятых каналов для данной системы при заданной интенсивности потока задач составляет 4,2, т.е. не превышает 5. Среднее число заявок в системе при К>5 также не превышает 5. Таким образом, оптимальное число модулей памяти 6. Это подтверждается и другими характеристиками, поскольку при К>6 характеристики системы изменяются незначительно и стремятся к 0.

При К=10 число заявок в очереди *l* приближается к 0, как и время ожидания в очереди ω, время ответа памяти *u* состоит из времени доступа к памяти.

Анализ влияния времени обращения к памяти на реальную пропускную способность подсистемы «процессор-память». Поскольку в качестве исходных данных для времени обращения к памяти взято среднее значение, в реальной системе оно может быть больше этого значения. Поэтому важно проанализировать модель с большей величиной времени доступа к памяти с целью исследования предельных значений.

Исходные данные: число обслуживающих каналов (модулей ОП) в СМО К = 8; число источников нагрузки (процессоров) ЦП = 4; время обслуживания заявок одним каналом (модулем ОП) ν=37,5-40 нс; интенсивность потока запросов λ=0,028 запроса/нс. Результаты исследования приведены на рисунке 5.

Из рисунка следует, что критическое состояние в системе наблюдается при значительной задержке ответа памяти (до 77,5 нс). При этом коэффициент загрузки каналов ρ составляет 1,085 – наступает перегрузка. При небольшом увеличении времени доступа к памяти (до 45,5 нс) время ответа памяти увеличивается на 10 нс по сравнению со значением исходной системы, число заявок в очереди не превышает 0,319 заявок. Эти значения некритичны, поэтому незначительно влияют на производительность системы.

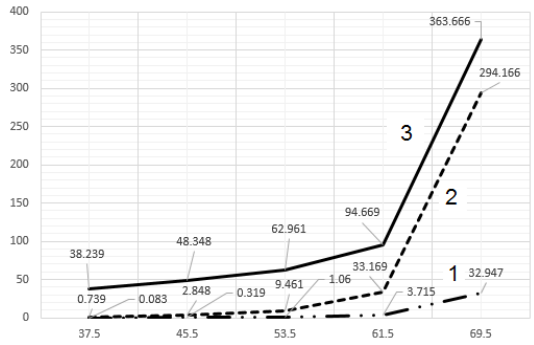
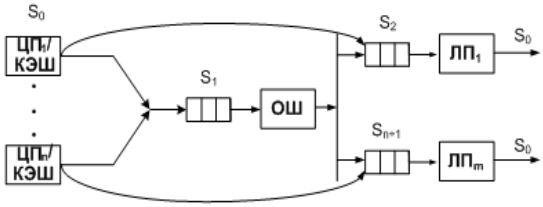
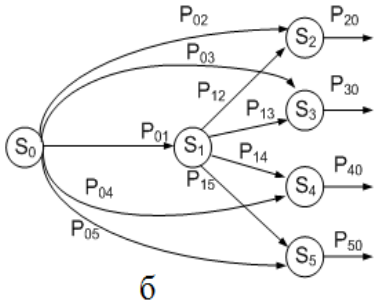


Рисунок 5 – Зависимость длины очереди перед устройством (1), времени ожидания в очереди (2) и времени ответа памяти (3) от времени обращения к памяти

Разработка модели подсистемы «процессор-память» архитектуры NUMA. Структура архитектуры *NUMA* (*Non Unified Memory Access*) и граф передач представлены на рисунке 6, а, б.



а



б

Рисунок 6 – Структура (а) и граф передач (б) *NUMA*-системы

Здесь: $ЦП_n$/КЭШ – *n* процессорных узлов; ОШ – общая шина; $ЛП_m$ – *m* модулей локальной распределенной памяти; S_0 – источник заявок, представляющий 4 процессорных узла; S1 – шина; S2–S5 – 4 модуля локальной распределенной памяти.

Определение исходных данных для архитектуры NUMA. Данная архитектура также рассматривается на примере семейства *Intel Pentium* IV 500. Объём кэша *L*2 процессоров серии *Pentium* 4 500 – 1 Мб.

Поэтому частота запросов в оперативную память остается неизменной, а именно λ=0,028(з/нс). При исследовании влияния производительности процессора на работу подсистемы «процессор-память» рассмотрены процессоры *Intel Pentium* 4 521 – 571 с частотой синхронизации 2,8-3,8 ГГц.

В расчетах используем модули памяти *DDR* с частотой 200 МГц. Относительно модуля оперативной памяти справедливы расчеты, произведенные выше. $V_{оп}$ = 22,5 (нс)

Поскольку в модели данной архитектуры шина «контроллер-процессор» является отдельным элементом, то среднее время обращения к памяти будет состоять из среднего времени работы модуля ОП и 1 такта шины «память-контроллер» для передачи слова от памяти. Получаем: 22,5+5 =27,5 (нс). При обращении процессора к своему локальному модулю памяти среднее время обращения к памяти составит 27,5 нс.

При обращении процессора к локальному модулю памяти другого процессорного узла дополнительно потребуется 2 такта шины «контроллер-процессор»: для передачи контроллеру 32-битного адреса и для передачи 64-битного слова от памяти через контроллер к процессору. 5x2=10 (нс). Таким образом, задержка шины составит 10 нс.

При этом считаем, что 90 % обращений процессорных узлов происходит к своему локальному модулю памяти, а 10 % обращений – к локальному модулю памяти другого процессорного узла.

Анализ влияния эффективности кэш-памяти на реальную пропускную способность подсистемы «процессор-память». Исходные данные: число модулей ОП в СМО К = 4; число источников нагрузки (процессоров) ЦП = 4; время обслуживания заявок одним каналом (модулем ОП) ν=27,5 нс; время задержки шины – 10 нс.

Интенсивность потока запросов при моделировании изменялась следующим образом: вероятность кэш-промаха, % · (частота синхронизации процессора). Диапазон вероятности кэш-промаха – 0,5-3 %. Результаты исследования приведены на рисунке 7.

При попадании в кэш 99 % запросов (поток заявок в память от процессора 0,028 запроса/нс): среднее число заявок в системе m=13,517 заявок; средняя длина очереди l=10,325 заявок; время ожидания заявки в очереди ω=92,191 нс; время пребывания заявки в СМО (время отклика памяти) u=120,691 нс.

При потоке заявок 0,042 запроса/нс (величина кэш-промаха составляет 1,5 %) в системе наблюдается перегрузка – ρ=1,155.

При увеличении эффективности кэш-памяти в 2 раза (число попаданий в кэш составляет 99,5 %) время отклика памяти составляет 45,774 нс, что по сравнению с исходным вариантом лучше примерно на 38 %. При этом длина очереди l<1.

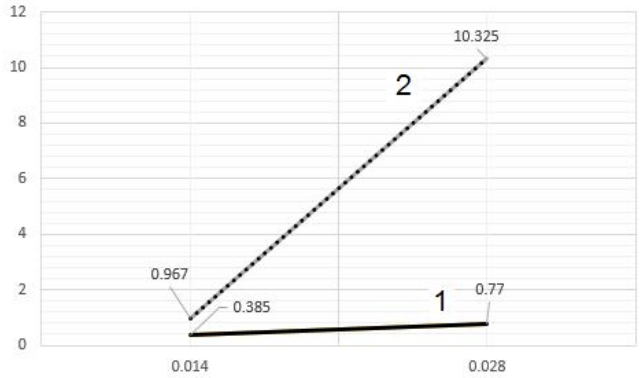


Рисунок 7 – Зависимость коэффициента загрузки (1) и длины очереди перед устройством (2)от вероятности кэш-промаха

Анализ влияния числа процессорных узлов и модулей памяти на реальную пропускную способность подсистемы «процессор-память». Исходные данные: число модулей ОП в СМО К = 4-7; число источников нагрузки (процессоров) ЦП = 4-7; время обслуживания заявок одним каналом (модулем ОП) ν=27,5 нс; время задержки шины – 10 нс; интенсивность потока запросов λ =0,028 запроса/нс. Результаты исследования приведены на рисунке 8.

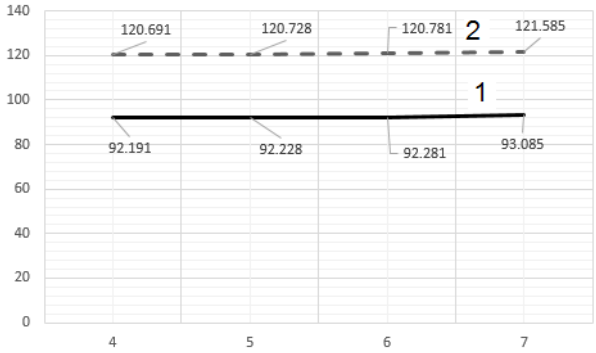


Рисунок 8 – Зависимость времени ожидания в очереди (1) и времени ответа памяти (2) от числа процессорных узлов и модулей памяти

Несмотря на одинаковое время обращения к памяти, длина очереди, число заявок в системе, время ожидания в очереди и время ответа памяти увеличиваются при увеличении числа процессорных узлов и модулей памяти. Это происходит в связи с увеличением вероятности обращения нескольких процессорных узлов к удаленному модулю памяти.

В целом увеличение характеристик незначительно (время ответа памяти увеличивается примерно на 1 нс). Поскольку 90 % обращений к

памяти процессорный узел производит к «своему» локальному модулю.

Анализ влияния времени обращения к памяти на реальную пропускную способность подсистемы «процессор-память». Поскольку в качестве исходных данных для времени обращения к памяти взято среднее значение, в реальной системе оно может быть больше этого значения. Поэтому важно проанализировать модель с большей величиной времени доступа к памяти с целью исследования предельных значений.

Исходные данные: число модулей ОП в СМО K = 4; число источников нагрузки (процессоров) ЦП = 4; время обслуживания заявок одним каналом (модулем ОП) ν =27,5-37,5 нс; время задержки шины – 10 нс; интенсивность потока запросов λ=0,028 запроса/нс. Результаты исследования приведены на рисунке 9.

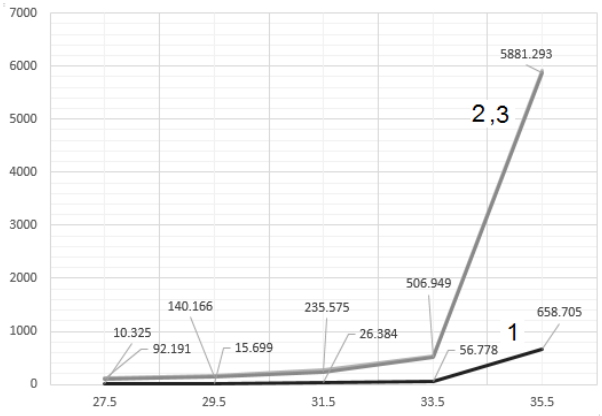


Рисунок 9 – Зависимость длины очереди перед устройством (1), времени ожидания в очереди (2) и времени ответа памяти (3) от времени обращения к памяти

При незначительном увеличении времени доступа к памяти (до 29,5-31,5 нс) время отклика памяти возрастает в 1,4-2,2 раза соответственно. С дальнейшим ростом значения доступа к памяти при 35,5 нс задержка перед ответом памяти достигает 5917,793 нс, что в 49 раз больше исходного значения. Перегрузка системы происходит при 37,5 нс доступа к памяти.

Заключение. По результатам проведенных исследований можно сделать следующие выводы.

В МВС архитектуры *UMA* при функционировании в многозадачном режиме поток заявок непрерывно возрастает, что объясняет большее число обслуженных заявок. При этом латентность памяти данной системы ниже, чем при однозадачном режиме. Это объясняется тем, что ПУ, не ожидая ответа памяти, делают новый запрос. При этом жизнеспособность системы выше, поскольку даже при высоком потоке заявок система не перегружена в отличие от первой, где подсистема памяти не справляется с высокой интенсивностью запросов.

Архитектура *NUMA* отличается тем, что при резком возрастании потока запросов в модули памяти по общей шине происходит резкое увеличение времени ответа памяти, поскольку межпроцессорная шина постоянно занята, это приводит к росту времени ожидания. Хотя при постоянном потоке заявок наблюдается стабильность работы системы.

Разработанные модели могут использоваться при проектировании многопроцессорных вычислительных систем. Данные разработки дают возможность производить оценку характеристик многопроцессорных систем и их подсистем без построения реального макета. За счет этого достигается экономический эффект, поскольку оценку характеристик проектируемых систем и выбор наиболее оптимальных вариантов можно проводить без построения реальной системы.

Работа выполнена при финансовой поддержке стипендии Президента РФ молодым ученым и аспирантам на 2015—2017 гг. (СП-828.2015.5)

Библиографический список

1. *Богуславский Л.Б.* Вероятностные методы и модели управления потоками данных и ресурсами в сетях и многопроцессорных системах: автореф. дис. … докт. техн. наук. – М.: Институт проблем управления, 1995. – 38 с.

2. *Крил П.* NUMA-Q – архитектура для многопроцессорных систем // Computer world. 1996. - № 43 (60) URL: http://www.osp.ru/data/www2/cw/1996/43/32.htm (Дата обращения: 01.10.2015).

3. *Беседин Д.* Non-Uniform Memory Architecture (NUMA). Часть 2: исследование подсистемы памяти четырехпроцессорных платформ AMD Opteron с помощью RightMark Memory Analyzer. 2006. URL: http://www.ixbt.com/cpu/rmma-numa2.shtml (Дата обращения: 01.10.2015).

4. *Майоров С. А.* Основы теории вычислительных систем: учеб. пособие для вузов / С. А. Майоров, Г. И. Новиков, Т. И. Алиев, Э. И. Махарев, Б. Д. Тимченко; под ред. С. А. Майорова. – М.: Высшая школа, 1978. – 409 с.

5. *Клейнрок Л.* Вычислительные системы с очередями. – М.: Мир, 1979. – 600 с.

6. *Вентцель Е.С.* Теория вероятностей: учебник для вузов. — 6-е изд. стер. — М.: Высшая школа, 1999.— 576 с.

7. *Алиев Т. И.* Основы моделирования дискретных систем. – СПб.: СПбГУ ИТМО, 2009. – 363 с.

8. Свидетельство о государственной регистрации программы для ЭВМ № 2013611117. Программный комплекс для расчета вероятностно-временных характеристик стохастических сетей массового обслуживания / А.И. Мартышкин, Р.А. Бикташев.

9. *Бикташев Р.А.* Многопроцессорные системы. Архитектура, топология, анализ производительности: учеб. пособие. – Пенза: Изд-во Пенз. гос. ун-та, 2004. – 107 с.

10. *Пахомов С.* Скоростная память DDR3: стоит ли игра свеч?// Компьютерпрресс. 2008. - №3 URL: http://compress.ru/article.aspx?id=18761 (Дата обращения: 01.10.2015).