

УДК 681.142.38:681.323

ОБ ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ
БУФЕРА КОМАНД ПРИ РАБОТЕ ОДНОРОДНОЙ ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЫ
В СИНХРОННО-ПРОГРАММНОМ РЕЖИМЕ

Ю.К. Дмитриев

Одним из основных свойств однородных вычислительных систем (ОВС) является их гибкость, проявляющаяся, в частности, в возможности реализации в них режимов, связанных с потоками команд и данных [1]. Как правило, в ОВС реализуется режим множественного потока команд с множественным потоком данных, называемый здесь А-режимом (автономным режимом). На ОВС возможна простая аппаратная реализация другого режима: одиночного потока команд с множественным потоком данных (С-режима). Этот режим используется, например, в системе ПЦЛАС-IV.

При работе в С-режиме специально выделенные элементарные машины (С-ЭМ) системы выполняют над своими операндами одну и ту же программу, которая выдается покомандно из одной С-ЭМ. С-ЭМ, которая хранит, выдает и выполняет программу, называется ведущей (В-ЭМ), а С-ЭМ, которые только исполняют программу, называются ведомыми (В-ЭМ).

Основные достоинства и недостатки С-режима описаны в [2,3]. В [2] показано, что одним из основных факторов, влияющих на эффективность использования этого режима, является зависимость времени выполнения команд от значений операндов. Установлено, что величина потерь времени зависит от длины программы и числа выполняющих ее машин. Показано, что для любой программы существует такое число машин, при котором математическое ожидание потерь времени меньше любой наперед заданной сколь угодно малой величины ε .

Величина потерь времени из-за разных затрат времени на выполнение одинаковых команд в разных машинах системы зависит от по-

пользуемого метода синхронизации. В [2] подробно исследован метод покомандной синхронизации С-ЭМ, при котором машина, выдающая программу, и машины, выполняющие ее, работают покомандно и поочередно, останавливаясь после завершения своей части текущего взаимодействия. Выдача очередной команды осуществляется лишь после выполнения текущей. Анализ покомандной синхронизации дает оценку потерь времени для С-режима сверху.

В данной работе будет использован метод синхронизации С-ЭМ с помощью сверхоперативной памяти, входящей в каждую такую машину.

1. Архитектура системы. При работе в А-режиме темп выдачи команд из оперативной памяти (ОП) В-ЭМ определяется внутренней структурой вычислительного модуля элементарной машины, в частности наличием в нем сверхоперативной памяти команд (буфера команд). Блок-схема вычислительного модуля, имеющего буфер команд, при работе его в А-режиме показана на рис.1, а (П - процессор, БК - буфер команд, СС - схема старшинства запросов в ОП). Здесь выделены пути обращения к ОП за командой. Блок-схема соответствует мультипроцессорной и/или мультипрограммной машине, отдельные устройства которой (ОП, процессоры и др.) работают асинхронно.

С точки зрения загрузки устройств вычислительного модуля, для которых буфер команд является источником информации, емкость буфера команд должна быть такой, чтобы в нем была хотя бы одна команда при заданных колебаниях значений скорости загрузки f_I и скорости разгрузки f_O . При этом буфер команд должен проектироваться как асинхронный блок, загрузка в который возможна лишь, если буфер команд неполон, а разгрузка - если буфер команд не пуст. В соответствии со сказанным, сигнал "Запрос команды", который может интерпретироваться как "Готовность буфера к приему", является функцией от загруженности последнего. Этот сигнал существует до тех пор, пока имеется хотя бы одна нагруженная ячейка буфера. Предполагается, что запрос ОП на выборку команд схемно отличается от запроса ОП на выборку операндов и т.п. Это связано с особенностями обработки кодов команд и зависящей от неё загрузки процессоров.

Буфер команд тем лучше справляется с задачей сглаживания колебаний f_I и f_O и с задачей сокращения числа обращений к ОП, чем больше его емкость E . Однако увеличение емкости буфера увеличи-

вае стоимость вычислительного модуля. В связи с этим задача оп-
ределения емкости буфера команд ставится как задача расчета числа
ячеек, которое обеспечивает при заданных колебаниях f_I и f_0 уста-
новленное значение загрузки процессора. Такая задача решена для
случая одной изолированно работающей машины. Целью нашей работы
будет определение влияния на загрузку процессоров машины ОВС ко-
личества машин, выполняющих одну и ту же программу. Оценка будем
давать для программ, не содержащих ветвлений.

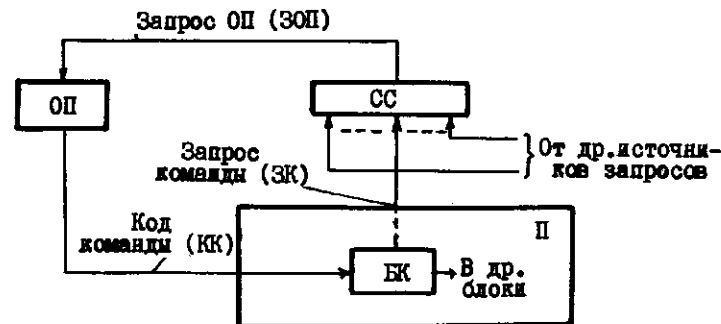
На рис. 1,б (СВЗК - схема выработки запроса команды, ММС -
межмашинные связи, В-ЭМ - ведущая ЭМ, В-ЭМ - ведомая ЭМ) приведе-
на блок-схема ОВС, в ЭМ которой используется буфер команд при ра-
боте в С-режиме. Как и на рис. 1,а, выделены только пути обращения
к ОП за кодом команды. Линии задержки, показанные на рис. 1,б пунк-
тиром, определяют задержки распространения сигналов по межмашин-
ным связям. В данном рассмотрении эти задержки считаются пренеб-
режимо малыми. Схема выработки запросов команды определяет зави-
симость выдачи заявки в ОП В-ЭМ на очередную команду от выработки
соответствующего сигнала в каждой С-машине. Таким образом, схема
выработки запросов команды в логическом отношении представляет со-
бой многоходовую схему И.

Рассмотрим на примере использование буфера команд заданной
емкости Е при выполнении программы в С-режиме. Значение Е найдено
из условия установленной загрузки ЭМ в А-режиме. Положим, для про-
стоты, что программа из $z = 11$ команд выполняется на двух ЭМ, име-
ющих $E = 2$. ЭМ, имеющую наибольшее значение времени выполнения дан-
ной программы в А-режиме, назовем определяющей машиной. Выполнение
программы начинается после полного заполнения буфера команд. Для
упрощения считаем, что обращение к ОП за очередной командой пол-
ностью совмещено в В-ЭМ с выполнением текущей. Это не нарушает об-
щности и вместе с тем упрощает качественное рассмотрение процесса.

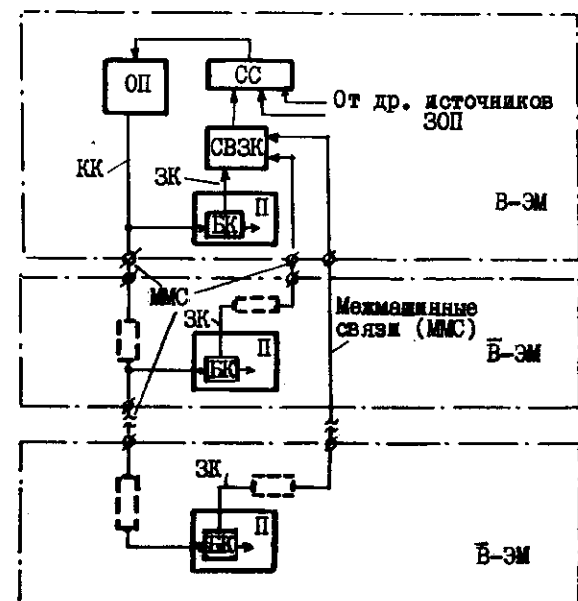
Распределение времени выполнения команд рассматриваемой про-
граммы задано для А-режима на рис. 2,а^{*)}. На рис. 2,б построена за-
висимость времени T_x работы ЭМ от числа x выполненных команд:

$$T_x = \sum_{i=1}^x t_i, \text{ где } x=1, 2, \dots, z, \text{ а } t_i - \text{ время выполнения } i\text{-й команды. Определяющей машине на этом графике соответствует}$$

*) На рис. 2,а представлены значения времени выполнения команд для
однопроцессорных ЭМ без совмещения операций. В общем случае эти
значения есть промежутки времени между последовательными обра-
щениями к буферу и находятся в интервале $\left[\frac{1}{f_{\text{min}}} = \frac{1}{f_{\text{max}}} \right]$.



а)



б)

Рис. 1

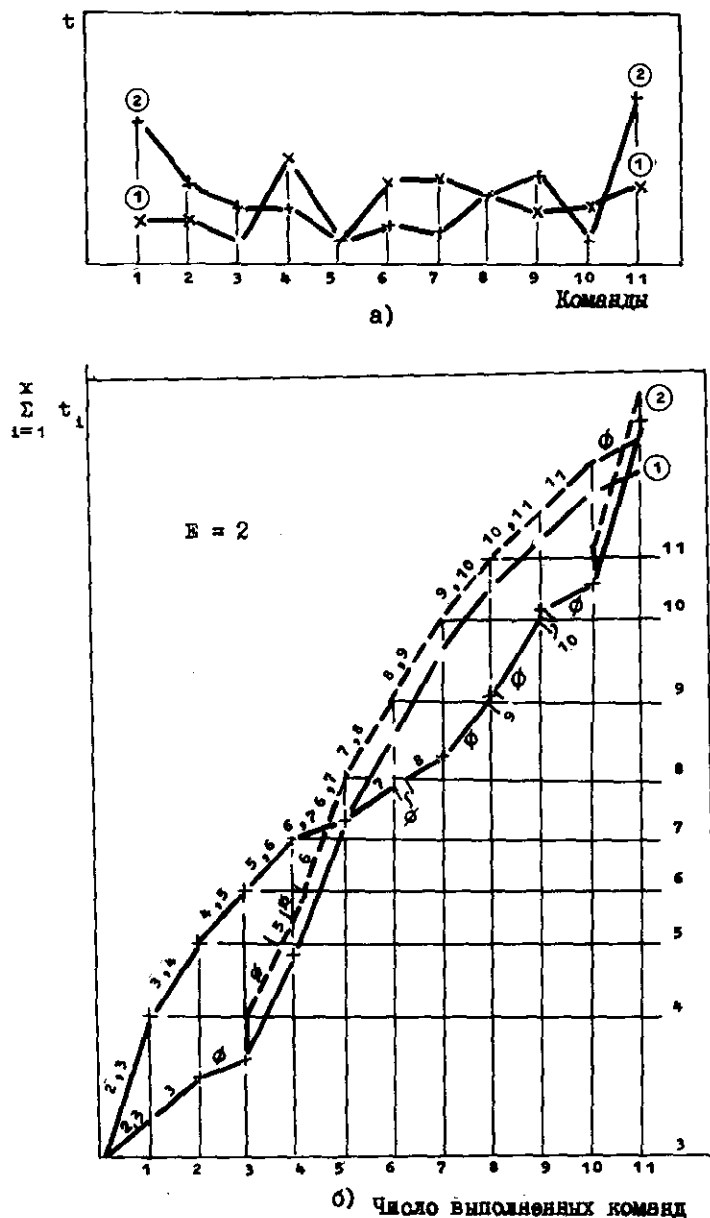


Рис. 2.

наибольшее значение ординаты при $x = 2$. Горизонталь с номером x на рис. 2, б отмечает момент поступления кода x -й команды из В-ЭМ на входы буферов команд ЭМ ОВС в С-режиме. Номера команд, находящихся в буфере команд, для каждой ЭМ выписаны над соответствующими участками графика.

Как можно видеть, различие скоростей разгрузки буферов команд в С-машинах может привести к увеличению времени решения задачи в С-режиме сравнительно с временем её решения в А-режиме. После выполнения третьей команды ЭМ-1 простаивает в ожидании поступления кода четвертой команды, вследствие чего график для нее смещается в положение, указанное пунктиром. Аналогичным образом ЭМ-2 простаивает в ожидании кода одиннадцатой команды.

Проведенное рассмотрение показывает, что время выполнения программы в С-режиме при синхронизации с помощью буфера команд по-прежнему зависит от емкости этого буфера. Видно, что величина потеря С-режима зависит не только от емкости буфера команд, но и от числа взаимодействующих машин. Легко видеть также, что время выполнения программы в А-режиме l машинами также зависит от l : $T_A = \max_{1 \leq j \leq l} T_j$, где T_j - время выполнения программы в j -й ЭМ.

На эффективность использования С-режима, кроме емкости буфера команд, влияет также время обращения к ОП за командой в ведущей машине. Эффективность работы ОВС в С-режиме может быть повышена, если запросам на ОП за командами в С-режиме будет выделен собственный, более высокий, чем для А-режима, приоритет. Изучим указанные факторы методом статистического моделирования.

2. Модель системы. ЭМ системы описывается как двухузловая двухфазная система массового обслуживания без потерь. Узел генерации заявок (оперативная память) выдает заявки в случайные моменты времени, подчиняющиеся закону распределения,

$$F_1(t) = \begin{cases} 0, & \text{если } e = E, \\ P(\tau_1 > t), & \text{если } e < E. \end{cases}$$

Здесь e - текущее значение числа заявок в накопителе заявок (буфер команд), E - заданная емкость накопителя заявок, $P(\tau_1 > t)$ - вероятность того, что если заявка поступила в систему в данный момент, то очередная заявка не поступит за время t . Узел обслуживания (процессор) имеет закон распределения времени обслуживания

$$F_2(t) = \begin{cases} \text{не определено, если } e = 0, \\ P(\tau_e > t), & \text{если } e \leq E. \end{cases}$$

В качестве параметров обслуживания для ЭМ используются среднее время выполнения команды t_0 и коэффициент загрузки процессора $A = t_n / t_0$, где t_n — среднее время выполнения команды процессором.

Величины t_0 и A имеют вид:

$$t_0 = f_1(E), \quad A = f_2(E).$$

Описанная модель массового обслуживания M используется в качестве основной при построении модели ОВС в А- и С-режимах (M_A и M_C соответственно). Модель M_A состоит из ℓ экземпляров модели M , работающих независимо. Параметрами обслуживания такой системы являются величины:

$$t_A = \max_{1 \leq j \leq \ell} (t_j), \quad A_A = \frac{1}{\ell} \sum_{j=1}^{\ell} (t_{nj} / t_A).$$

Модель M_C состоит из узла генерации, связанного параллельно с ℓ узлами обслуживания через соответствующие последним накопители заявок емкостью E . Закон распределения времени между заявками от узла генерации следующий:

$$F_{1c}(t) = \begin{cases} 0, & \text{если } (\max e_j) = E, \\ P(\tau_1 > t), & \text{если } (\max e_j) < E. \end{cases}$$

Работа каждого из узлов обслуживания определяется состоянием его собственного накопителя e_j ($j = 1, 2, \dots, \ell$). Параметрами обслуживания M_C являются величины t_c и A_c , определяемые аналогично t_A и A_A .

Исследуются зависимости указанных параметров обслуживания от величин E , ℓ и от приоритетности R запроса к узлу генерации от накопителя заявок. Исследования проводятся в условиях достижения системой стационарного режима. В качестве условия стационарности используется неравенство

$$\max_{1 \leq j \leq \ell} \{T_{cj}(z) / T_A(z)\} - \min_{1 \leq j \leq \ell} \{T_{cj}(z) / T_A(z)\} \leq \xi,$$

где $T_{cj}(z)$ — время выполнения программы из z команд на j -й ЭМ в С-режиме, $T_A(z)$ — время выполнения той же программы в А-режиме на системе из ℓ ЭМ, ξ — заданная малая величина.

Изучение моделей M_A и M_C проводилось методом статистического моделирования для случая простейших потоков $F_1(t)$ и $F_2(t)$ с интенсивностями λ и μ соответственно. Приводимые ниже результаты соответствуют отношению $\lambda/\mu = 10:1$, что обеспечивает приемлемые для практики значения коэффициентов загрузки процессоров ЭМ в пределах 0,65–0,95. Ограничения на машинное время обусловили диапазон изменений независимых параметров при моделировании в следующих пределах: $E \leq 20$, $\ell \leq 20$, $R \leq 4$. Значение $\xi \leq 0,1$. Программы моделирования написаны на языке ФОРТРАН-ЦЕРН для ЭВМ БЭСМ-6 и имеют объем 500 операторов.

3. Обсуждение результатов. Оценка функционирования машины вне системной работы осуществляется по значению среднего коэффициента загрузки процессоров

$$A_{Aop} = \frac{1}{N \cdot \ell} \sum_{s=1}^N \sum_{j=1}^{\ell} k_{najs}, \quad (1)$$

где N — число проведенных испытаний системы, а k_{najs} — средний коэффициент загрузки процессора j -й ЭМ в s -м испытании. Значение k_{najs} определяется соотношением

$$k_{najs} = \sum_{i=1}^z t_{nji} / \sum_{i=1}^z t_{aji}.$$

Здесь t_{nji} — время выполнения i -й по счету команды процессором j -й ЭМ; t_{aji} — полное время выполнения i -й по счету команды j -м процессором: $t_{aji} = t_{nji} + t_{pmji}$; t_{pmji} — время обращения к оперативной памяти за i -й по счету командой j -й машины.

Оценка влияния системы машин на значение коэффициента загрузки процессора требует учета времени T_A выполнения заданной программы всеми машинами. Она производится по формуле

$$A_A = \frac{1}{N \cdot \ell} \sum_{s=1}^N \sum_{j=1}^{\ell} k_{najs}^*, \quad (2)$$

где

$$k_{najs}^* = \sum_{i=1}^z t_{nji} / T_A,$$

$$T_A = \max_{1 \leq j \leq \ell} \left\{ \sum_{i=1}^z t_{aji} \right\}.$$

Оценка функционирования системы в С-режиме производится по значениям средних коэффициентов $\bar{K}_{\text{СР}}$ и $\bar{K}_{\text{С}}$, определяемых аналогично $\bar{K}_{\text{АСР}}$ и $\bar{K}_{\text{А}}$. Время выполнения программы в С-режиме для j -й ЭМ ($T_{\text{С}j}$) отличается от общего времени $T_{\text{С}}$ выполнения этой программы системой из ℓ машин не более чем на величину полной разгрузки накопитель заявок емкостью E . При достаточно больших τ можно пренебречь этой разностью, вследствие чего для оценки С-режима определялось только значение $\bar{K}_{\text{С}}$.

Характер изменения указанных величин показан на графиках рис.3-6.

На рис.3 показан характер изменения коэффициентов загрузки процессора в А-режиме для разных значений E , ℓ и R . Большому значению R соответствует меньший приоритетный уровень запроса на команду. Из данных графиков можно видеть, что коэффициент загрузки процессора $\bar{K}_{\text{АСР}}$, служащий при достаточно высоких значениях R и E надежным показателем эффективности использования процессора одиночной машины, утрачивает свое значение уже при $\ell = 3$. Вместе с тем видно, что уже при $\ell = 10$ коэффициент $\bar{K}_{\text{А}}$ достигает постоянного значения (прямая линия $K_{\text{А}}$, пунктирная — $K_{\text{АСР}}$).

Характер зависимости $\bar{K}_{\text{С}}$ от значений E , ℓ и R аналогичен полученному для А-режима, хотя и дает несколько более низкие значения коэффициентов загрузки. На рис.4 приведены графики, иллюстрирующие относительное падение коэффициента загрузки процессора при переходе от А- к С-режиму. Из этих графиков следует, что при емкости буфера команд порядка 10 ячеек относительное падение коэффициента загрузки при переходе на С-режим составляет около 10% общего времени выполнения программы. Кривая для $E = 0$ соответствует С-режиму с покомандной синхронизацией.

На рис.5 показано, какие приращения емкости буфера команд необходимы для того чтобы обеспечить коэффициент загрузки процессора на уровне, достигнутом для А-режима. В пределах $\bar{K}_{\text{А}} = 0,65-0,7$ такие приращения составляют около 4 ячеек для $\ell \leq 20$.

Данные рис.6 показывают, что при определенных условиях компенсация падения коэффициента загрузки процессора при переходе к С-режиму может быть обеспечена присвоением команде обращения к памяти в С-режиме более высокого приоритета, чем в А-режиме. Это свидетельствует о возможности эффективной организации С-режима при сохранении емкости буфера команд, используемой для А-режима.

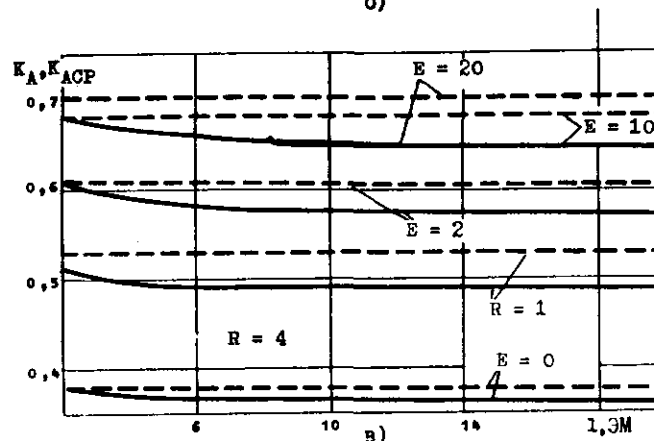
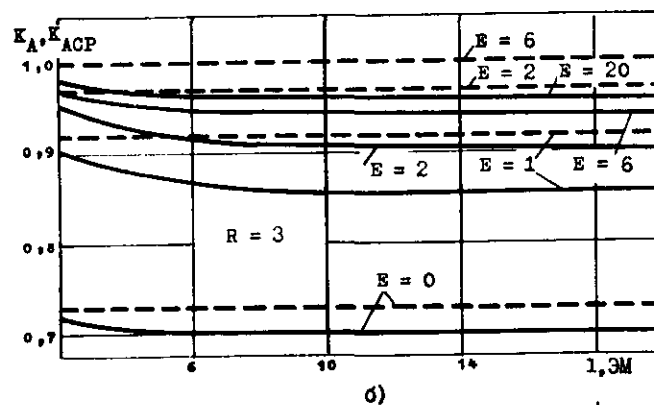
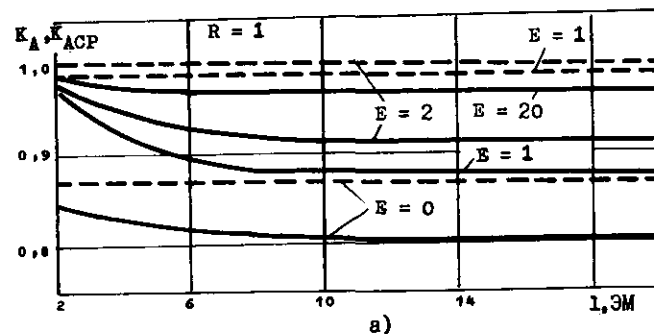
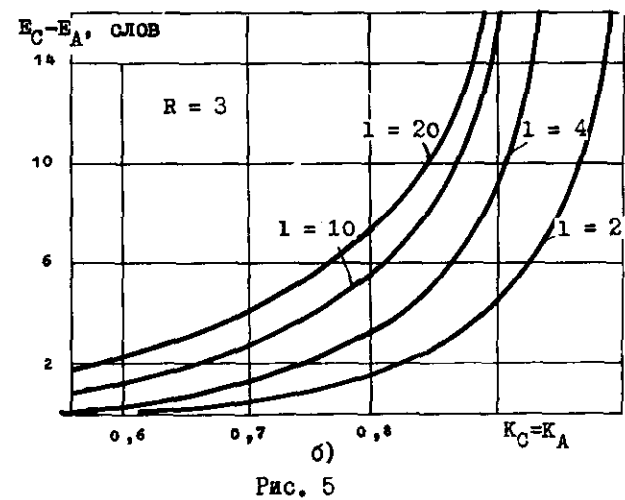
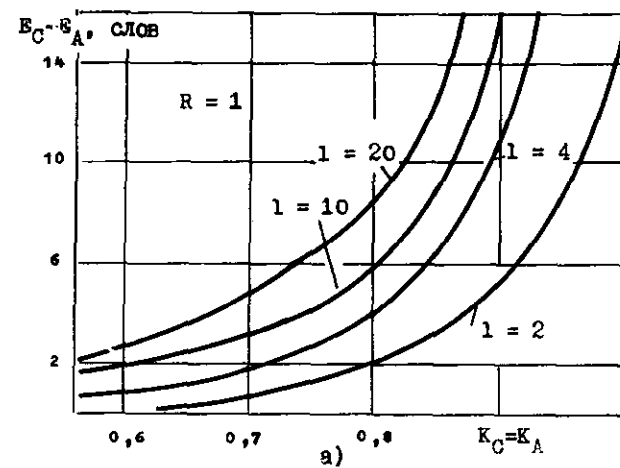
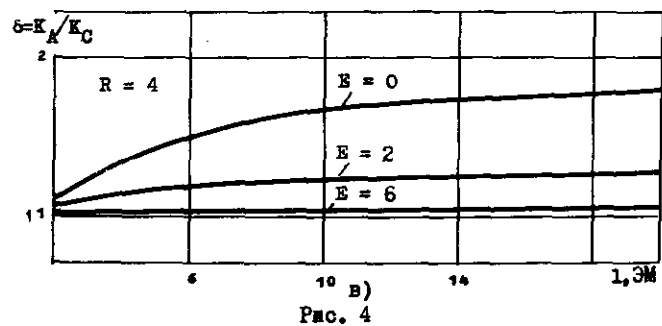
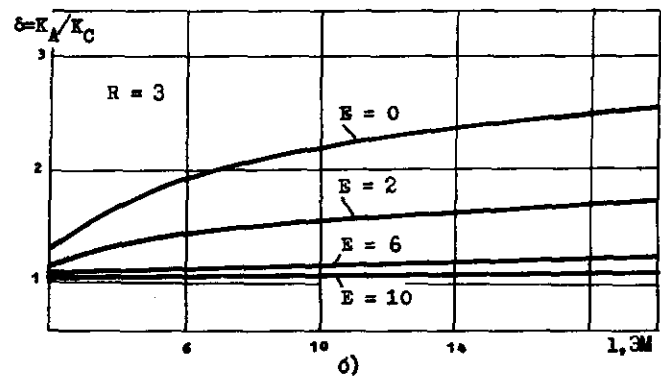
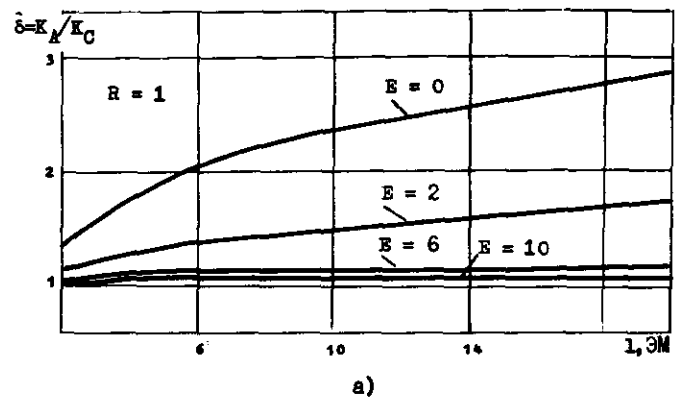


Рис. 3



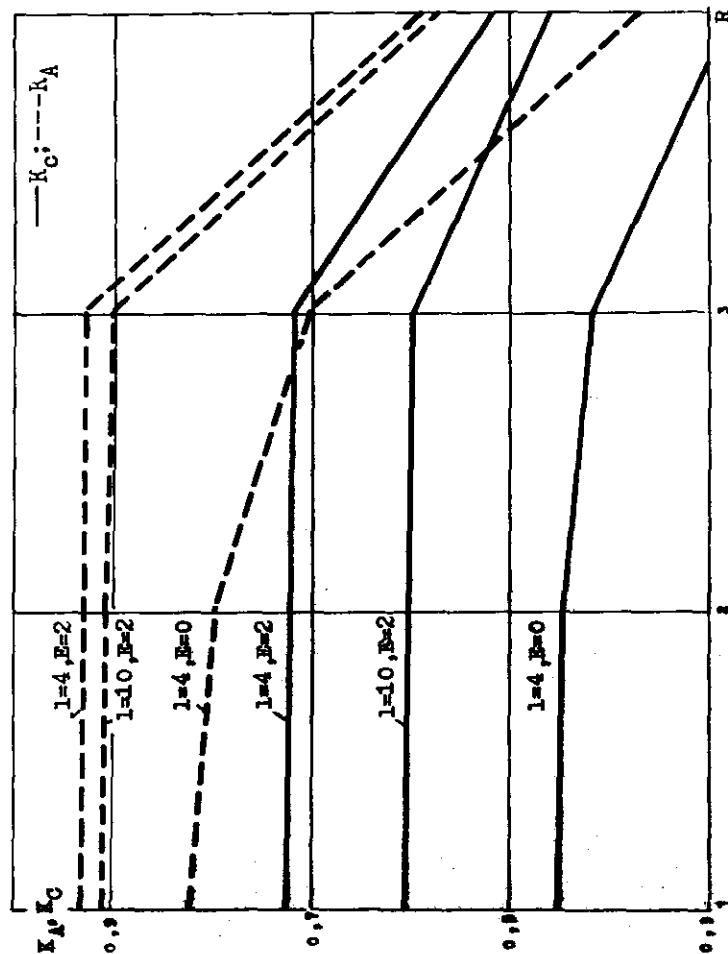


Рис. 6

Результаты статистического моделирования подтверждают эффективность предложенных архитектурных мер повышения скорости работы системы в С-режиме. Рассмотрены два архитектурных приема: увеличение емкости сверхоперативной памяти команд и изменение приоритета обращения за командой в системе запросов к оперативной памяти. Первый прием применим в системах с программируемой микроструктурой [4]. Эти системы позволяют программно изменять емкость буфера команд в соответствии с числом машин, занятых в А- и С-режимах. Второй прием позволяет использовать работу в С-режиме при традиционной структуре ЭВМ третьего и четвертого поколений. Он обеспечивает постоянное значение коэффициента загрузки процессора независимо от числа занятых в работе машин.

В моделировании и обсуждении результатов принимала участие Л.К.Игнатенко, за что автор выражает ей благодарность.

Л и т е р а т у р а

1. FLYNN M.J. Very High-Speed Computing System. - "Proc.IEEE", 1966, v.54, n 12, p.1909-1919.
2. ДИМИТРИЕВ Ю.К. О потерях времени при работе однородной вычислительной системы в синхронно-программном режиме. - В кн.: Теория однородных вычислительных систем. (Вычислительные системы. Вып. 63.) Новосибирск, 1975, с.69-85.
3. ДИМИТРИЕВ Ю.К. О способе реализации ветвящихся программ при работе однородной вычислительной системы в синхронно-программном режиме. - В кн.: Вопросы теории и построения вычислительных систем. (Вычислительные системы. Вып. 60.) Новосибирск, 1974, с. 103-114.
4. ВИШНЕВСКИЙ А.П. Об объединении цифровой вычислительной машины со структурно однородным логическим устройством. - В кн.: Вычислительные системы. Вып. 26. Новосибирск, 1967, с. 57-67.

Поступила в ред.-изд.отд.
29 декабря 1976 года