

АНАЛИЗ ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЫ С ОБЩЕЙ ОПЕРАТИВНОЙ ПАМЯТЬЮ

О. И. Семенков
(Минск)

В работе рассматривается цифровой вычислительный комплекс, состоящий из n , в общем случае, разнородных цифровых вычислительных устройств (ЦВУ), работающих с общей оперативной памятью. Показано, что при организации такой вычислительной системы по асинхронному принципу в работе ЦВУ возникают задержки, вызванные простаиванием в очереди на обращение к оперативному запоминающему устройству (ОЗУ). Для двух дисциплин обслуживания — с приоритетом и без приоритета — получены зависимости, позволяющие определить среднее время задержки для любого ЦВУ комплекса. Результаты моделирования такой системы на вычислительной машине хорошо согласуются с расчетными данными.

Постановка задачи и математическая модель системы

Одним из возможных путей создания высокоэффективных вычислительных машин является объединение ряда, в общем случае, разнородных цифровых вычислительных устройств в единую вычислительную систему с общим объемом оперативной памяти. Такую систему назовем цифровым вычислительным комплексом (ЦВК). На рис. 1 приведена структурная схема ЦВК, состоящего, в част-

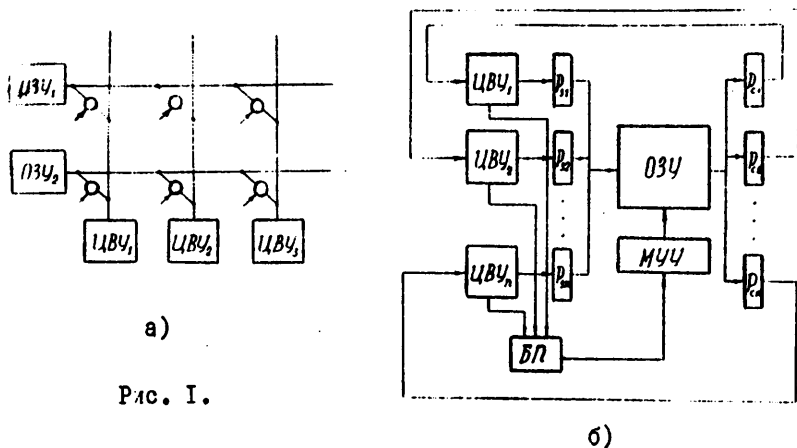


Рис. I.

ности, из двух блоков оперативной памяти с автономным управлением и единой адресной системой ($ОЗУ_1$ и $ОЗУ_2$) и трех цифровых вычислительных устройств ($ЦВУ_1 + ЦВУ_3$). При матричной схеме соединения элементов такой вычислительной системы и единой адресной системе оперативной памяти любое цифровое вычислительное устройство комплекса может обратиться к любому блоку $ОЗУ$, что обеспечивает максимальную гибкость при обмене информацией между отдельными $ЦВУ$.

В качестве исходной структуры для анализа примем схему ЦВМ с одним блоком (модулем) оперативной памяти, изображенную на рис. Iб. Каждое из устройств комплекса имеет свой регистр считывания $Р_1$ для приема информации и регистр записи $Р_2$, из которого производится запись информации в $ОЗУ$. Все $ЦВУ$ комплекса по своим программам производят параллельно во времени определенные операции над числами, полученными из $ОЗУ$, а результаты этих операций отсылают обратно в оперативную память. При этом вычислительные устройства комплекса решают либо самостоятельные задачи, либо соответствующие части одной общей задачи с обменом информацией через общую память.

Для обращения к $ОЗУ$ каждое устройство вырабатывает специальный сигнал, который поступает в блок приоритета (БП). Последовательность сигналов, вырабатываемых данным устройством, назовем потоком заявок на обращение к $ОЗУ$ со стороны этого устройства.

Система построена таким образом, что если к моменту прихода очередной заявки ОЗУ оказывается занятым выполнением другой заявки, то поступившая заявка запоминается в блоке приоритета до момента окончания обслуживания первой заявки, а соответствующее ЦВУ простаивает в течение времени, необходимого для окончания обслуживания первой заявки. Время задержки оказывает непосредственное влияние на производительность как отдельных устройств, так и всего комплекса, что, безусловно, следует учитывать при оценке эффективного быстродействия.

Для определения среднего времени задержки для каждого из ЦВУ будем полагать, что потоки заявок всех устройств являются стационарными и независимыми.

Заявки устройств представим на оси времени в виде прямоугольных импульсов, которые поступают по соответствующим каналам в БП, не перекрываясь в каждом из этих каналов друг с другом. Длительность импульсов равна времени обращения соответствующих ЦВУ к ОЗУ.

Для анализа такой системы удобно использовать аппарат теории случайных дискретных потоков.

Процесс совпадения заявок и образования задержек рассмотрим на следующей модели.

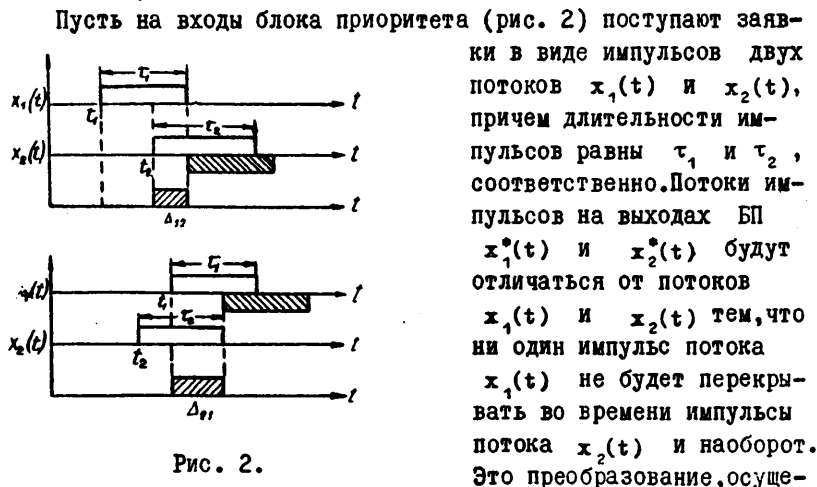


Рис. 2.

Это преобразование, осуществляемое блоком приоритета над входными потоками, описывается следующими выражениями:

$$\begin{aligned} x_1^*(t_1) &= x_1(t_1) \\ x_2^*(t_2) &= x_2(t_2 + \Delta_{12}) \end{aligned} \quad \text{при } t_1 \leq t_2 \leq t_1 + \tau_1 \quad (I)$$

$$\begin{aligned} x_1^*(t_1) &= x_1(t_1 + \Delta_{21}) \\ x_2^*(t_2) &= x_2(t_2) \end{aligned} \quad \text{при } t_2 \leq t_1 \leq t_2 + \tau_2 \quad (2)$$

Здесь Δ_{21} - время задержки заявки потока $x_1(t)$,

Δ_{12} - время задержки заявки потока $x_2(t)$.

Запоминающая часть БП хранит задержанную заявку до момента окончания обслуживания ранее пришедшей.

Исходя из стационарности потоков $x_1(t)$ и $x_2(t)$, нетрудно получить вероятность того, что произвольный момент времени ξ окажется в пределах основания импульса $x_1(t)$.

Для рассматриваемых потоков эти вероятности равны:

$$p_1 = P(t_1 \leq \xi \leq t_1 + \tau_1) = \frac{\tau_1}{\bar{T}_1}, \quad (3)$$

$$p_2 = P(t_2 \leq \xi \leq t_2 + \tau_2) = \frac{\tau_2}{\bar{T}_2}, \quad (4)$$

где \bar{T}_1 и \bar{T}_2 - математические ожидания длительностей интервалов между двумя соседними заявками потоков $x_1(t)$ и $x_2(t)$ соответственно.

Учитывая независимость потоков, можно считать, что с вероятностью p_1 моменты прихода заявок потока $x_2(t)$ будут удовлетворять следующим неравенством:

$$t_1 \leq t_2 \leq t_1 + \tau_1. \quad (5)$$

Аналогично для потока $x_1(t)$ моменты прихода его заявок будут удовлетворять с вероятностью p_2 неравенствам

$$t_2 \leq t_1 \leq t_2 + \tau_2. \quad (6)$$

Из диаграммы рис. 3 видно, что всякий раз, как происходит совпадение заявок и выполняется неравенство (6) (событие А), заявка потока $x_1(t)$ задерживается на время Δ_{21} , которое является случайной равномерно распределенной величиной с условной плотностью вероятностей

$$w(\Delta_{21}/\Delta) = \begin{cases} \frac{1}{\tau_2} & \text{при } 0 \leq \Delta_{21} \leq \tau_2, \\ 0 & \text{при других значениях } \Delta_{21} \end{cases} \quad (7)$$

и условным математическим ожиданием

$$M[\Delta_{21}/\Delta] = \frac{\tau_2}{2}. \quad (8)$$

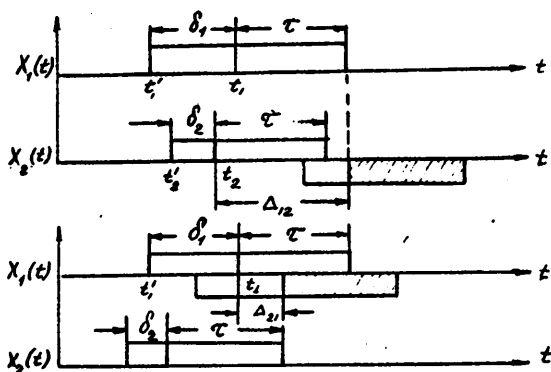


Рис. 3.

Используя формулу умножения вероятностей, получим следующее выражение для математического ожидания времени задержки импульсов потока $x_1(t)$

$$\overline{\Delta_{21}} \equiv \overline{\Delta_1} = \frac{\tau^2}{2T_2} \quad (9)$$

Аналогично для импульсов потока $x_2(t)$ имеем

$$\overline{\Delta_{12}} \equiv \overline{\Delta_2} = \frac{\tau^2}{2T_1} \quad (10)$$

Анализ системы с беспriorитетным обслуживанием

Обобщим результаты, полученные в предыдущем параграфе, на случай, когда с общей оперативной памятью через блок приоритета работают n цифровых устройств.

Теорема: Если на входы n -канального блока приоритета, каждая пара каналов которого описывается условиями (I) и (2), поступают n независимых стационарных потоков заявок $x_i(t)$ ($i = \overline{1, n}$), то среднее время задержки заявки на i -ом выходе блока определяется соотношением

$$\overline{\Delta_i} = \frac{n\tau^2}{4(n-1)} \sum_{s \neq i}^n \frac{1}{T_s} \quad (II)$$

Доказательство.

Пронумеруем случайные моменты времени поступления заявок в блок приоритета и составим следующую матрицу:

$$(t) = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ t_{21} & t_{22} & \dots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \dots & t_{nn} \end{bmatrix}. \quad (I2)$$

Элементом t_{ij} этой матрицы является случайный момент времени поступления заявки от i -го устройства, порядковый номер которого на оси времени от условного начала отсчета равен j .

Матрица (I2) содержит все возможные ситуации, которые могут возникнуть при совпадении заявок от n устройств. Так, например, вычеркивая в матрице i -ю строку и j -й столбец, получим минор этой матрицы размера $(n-1)$. Комбинируя каждый элемент первого столбца минора с элементами следующих по порядку столбцов, но расположенных в разных строках, получим все ситуации, в которых заявка i -го потока имеет j -й порядковый номер в очереди. Количество таких ситуаций, очевидно, равно

$$N = n - 1 \quad (I3)$$

и каждая из них появляется во времени равновероятно с вероятностью

$$p_1^1 = \frac{1}{(n-1)!}. \quad (I4)$$

Определим математическое ожидание времени задержки заявки i -го потока, занимающего с вероятностью

$$p_2^1 = \frac{1}{n-1} \quad (I5)$$

очередь с номерами $j = 2, 3, \dots, n$.

По определению имеем

$$m[\Delta_j] \equiv \bar{\Delta}_i = \sum_{j=2}^n p_2^1 \bar{\Delta}_i^{(j)} = \frac{1}{n-1} \sum_{j=2}^n \bar{\Delta}_i^{(j)}. \quad (I6)$$

Здесь $\bar{\Delta}_i^{(j)}$ — математическое ожидание времени задержки, испытываемой заявкой i -го потока, когда она занимает j -ю очередь.

Совершенно очевидно, что для $j = 1$ $\bar{\Delta}_i^{(1)} \equiv 0$.

Определим $\bar{\Delta}_i^{(j)}$ для $j > 1$.

$j = 2$.

Общее количество ситуаций, в которых заявка i -го потока приходит второй после заявки k -го потока равно $(n-1)$. При этом от каждой заявки возникает задержка $\bar{\Delta}_{ki}$, вычисляемая по формуле (7). Математическое ожидание времени задержки в этом случае равно

$$\bar{\Delta}_i^{(2)} = (n-2)! \sum_{k \neq i}^n p_k^1 \bar{\Delta}_{ki} = \frac{1}{n-1} \sum_{k \neq i}^n \bar{\Delta}_{ki}. \quad (17)$$

$j = 3$.

Заявка i -го потока, приходя третьей после заявок k -го и l -го потоков, испытывает суммарную задержку, равную

$$\bar{\Delta}_{lki} = \bar{\Delta}_{lk} + \bar{\Delta}_{ki}, \quad (18)$$

где $\bar{\Delta}_{lk}$ и $\bar{\Delta}_{ki}$ также определяются по формуле (7).

Суммируя (18) по всем $l = \overline{1, n}$ и $k = \overline{1, n}$ за исключением $l = k$, $k \neq i$, $l = i$ и имея в виду, что общее количество случаев, когда заявка i -го потока приходит после заявок l -го и k -го потоков, равно $(n-3)!$, получаем

$$\bar{\Delta}_i^{(3)} = (n-3)! \sum_{k \neq i, l}^n \sum_{l \neq i}^n p_l^1 (\bar{\Delta}_{lk} + \bar{\Delta}_{ki}). \quad (19)$$

Раскрывая сумму (19) и принимая во внимание (17), имеем

$$\bar{\Delta}_i^{(3)} = \bar{\Delta}_i^{(2)} + \frac{1}{(n-1)(n-2)} \sum_{l \neq k, i}^n \sum_{k \neq i}^n \bar{\Delta}_{lk}. \quad (20)$$

Нетрудно получить следующую рекуррентную формулу для любого $j > 2$:

$$\bar{\Delta}_i^{(j)} = \bar{\Delta}_i^{(j-1)} + \frac{1}{(n-1)(n-2)} \sum_{l=1}^n \sum_{k=1}^n \bar{\Delta}_{lk}. \quad (21)$$

Пользуясь этой формулой и принимая во внимание (17) и (14), находим следующее выражение для математического ожидания времени задержки заявки i -го потока, когда она занимает j -ю очередь:

$$\bar{\Delta}_i^{(j)} = \frac{1}{n-1} \sum_{k \neq i}^n \bar{\Delta}_{ki} + \frac{j-2}{(n-1)(n-2)} \sum_{l=1}^n \sum_{k=1}^n \bar{\Delta}_{lk}, \quad (22)$$

где $j = \overline{2, n}$. Подставляя теперь (22) в (15) и производя

суммирование по j , получаем

$$\bar{\Delta}_1 = \frac{1}{n-1} \sum_{k \neq 1}^n \bar{\Delta}_{k1} + \frac{1}{2(n-1)} \sum_1^n \sum_k^n \bar{\Delta}_{1k}. \quad (23)$$

Это выражение связывает среднюю величину задержки заявки 1-го потока с попарными задержками всех потоков, включая и 1-й. Нетрудно установить, что

$$\bar{\Delta}_{k1} = \bar{\Delta}_{k2} = \dots = \bar{\Delta}_{kn} = \frac{\tau_k^2}{2T_k}. \quad (24)$$

Поэтому после подстановки (24) в (23) и выполнения несложных преобразований, получаем окончательно

$$\bar{\Delta}_1 = \frac{n}{4(n-1)} \sum_{k \neq 1}^n \frac{\tau_k^2}{T_k}. \quad (25)$$

Тем самым теорема доказана.

Обычно время обращения в ОЗУ есть величина постоянная для всех ЦВУ. Поэтому формула (25) приобретает следующий вид:

$$\bar{\Delta}_1 = \frac{n \tau^2}{4(n-1)} \sum_{k \neq 1}^n \bar{q}_k, \quad (26)$$

где $\bar{q}_k = \frac{1}{T_k}$ — среднее быстродействие k -го ЦВУ.

Таким образом, средняя величина времени задержки определяется с одной стороны количеством цифровых устройств, входящих в комплекс, а с другой — интенсивностями потоков заявок на обращение ОЗУ. Физически величина $\bar{\Delta}_1$ представляет собой время, на которое в среднем увеличивается длительность каждой операции в данном ЦВУ. Следовательно, если вычислительное устройство обладает номинальным быстродействием

$$Q_1 = \frac{1}{T_1} \quad \left[\frac{\text{средних операций}}{\text{сек.}} \right], \quad (27)$$

то в результате работы в комплексе с общим ОЗУ его быстродействие снизится до величины

$$Q_1' = \frac{1}{T_1 + \bar{\Delta}_1}. \quad (28)$$

Относительное уменьшение быстродействия определяется следующим выражением:

$$\varepsilon_1 = \frac{Q_1'}{Q_1} = \frac{1}{1 + \frac{n \tau^2}{4(n-1)} \sum_{k \neq 1}^n \frac{1}{T_k}}. \quad (29)$$

Полученные результаты для структуры комплекса с одним модулем ОЗУ нетрудно распространить на более общий случай структуры с m независимыми модулями, объединенными единой адресной системой (рис. 1).

Из структуры задач, решаемых ЦВУ, а также из принятого распределения оперативной памяти между ЦВУ можно определить матрицу интенсивностей потоков заявок q .

Элементом этой матрицы \bar{q}_{ji} является средняя интенсивность потока заявок от i -го ЦВУ к j -му модулю ОЗУ. Очевидно, что

$$\bar{q}_i = \sum_{j=1}^n \bar{q}_{ji}, \quad i = \overline{1, n}. \quad (30)$$

Воспользовавшись выражением (26), определим среднюю величину времени задержки в i -м ЦВУ при обращении к j -му модулю ОЗУ:

$$\bar{\Delta}_{ji} = \frac{n\tau^2}{4(n-1)} \sum_{s \neq i}^n \bar{q}_{js}. \quad (31)$$

Переходя к матричным обозначениям, матрицу задержек можно определить из следующего выражения:

$$\bar{\Delta} = \frac{n\tau^2}{4(n-1)} (1 - E)\bar{q}, \quad (32)$$

где E — единичная матрица,

\bar{q} — матрица интенсивностей потоков заявок.

Анализ системы с приоритетным обслуживанием

Цифровой вычислительный комплекс, обслуживаемый общей оперативной памятью, относится к классу систем массового обслуживания, в которых прерывание обслуживания невозможно, так как в противном случае это привело бы к искажению или даже к полной потере информации, хранящейся в ОЗУ. Поэтому в силу специфики потоков заявок вычислительных устройств, характеризующихся дискретной функцией плотности вероятностей длительностей интервалов между моментами поступления соседних заявок, в цифровом комплексе можно организовать приоритетное обслуживание устройств методом "защитных" промежутков времени. Этот метод позволяет, не прерывая обслуживания заявок с

малым приоритетом, обеспечить в то же время приоритетное обслуживание заявок, имеющих любую более высокую степень приоритета.

Сущность его состоит в том, что до момента поступления t_1 очередной заявки I-го потока (рис. 3) в блок приоритета по тому же каналу поступает предварительный сигнал в момент t_1' . Таким образом, с момента t_1' блок приоритета оказывается уже занятым для других заявок и выполнение их переносится к моменту $t_1 + \tau$. Интервал времени

$$\delta_1 = t_1 - t_1' \quad (33)$$

между поступлениями предварительного сигнала и сигнала заявки I-го потока назовем "защитным" промежутком времени.

Определим величину времени задержки, которая возникает в результате совпадения заявок двух потоков, имеющих защитные промежутки времени длительностью δ_1 и δ_2 .

Как видно из диаграммы (рис. 3), плотность вероятностей величины времени задержки для заявок потока $x_2(t)$ по аналогии с (7) можно написать в виде

$$W(\Delta_{12}/\Delta) = \begin{cases} \frac{1}{\tau + (\delta_1 - \delta_2)} & \text{при } 0 \leq \Delta_{12} \leq \tau + (\delta_1 - \delta_2), \\ 0 & \text{при других значениях } \Delta_{12}. \end{cases} \quad (34)$$

Следовательно, математическое ожидание времени задержки заявок этого потока определяется аналогично (8) следующим выражением:

$$\bar{\Delta}_{12}(\delta) = \frac{[\tau + (\delta_1 - \delta_2)]^2}{2\bar{T}_1} \quad (35)$$

Выполнив те же преобразования для заявок потока $x_1(t)$, получим

$$\bar{\Delta}_{21}(\delta) = \frac{[\tau + (\delta_1 - \delta_2)]^2}{2\bar{T}_2} \quad (36)$$

Как видно из (35) и (36), имеется возможность производить перераспределение задержек между ЦБУ комплекса в зависимости от величины "защитного промежутка". В частном случае, при $\delta_2 = 0$ получаем следующие зависимости средних величин времени задержки для заявок обоих потоков в функции длительности защитного промежутка времени δ_1 :

$$\bar{\Delta}_{12}(\delta_1) = \frac{(\tau + \delta_1)^2}{2\bar{T}_1}, \quad (37)$$

$$\bar{\Delta}_{21}(\delta_1) = \frac{(\tau + \delta_1)^2}{2\bar{T}_2} \quad (38)$$

Как видим, в случае $\delta_1 = \tau$ заявки потока $x_1(t)$ получают абсолютный приоритет при обращении к МОЗУ за счет возрастания времени задержки заявок второго потока. На рис. 4 представле-

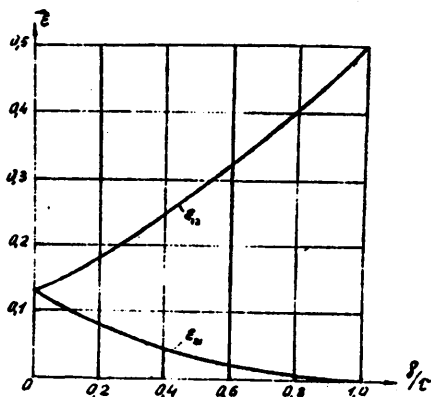


Рис. 4.

ны графики, иллюстрирующие зависимости относительных средних величин времени заявок обоих потоков в функции относительной длительности защитного промежутка времени δ_1/τ , построенные по формулам (37) и (38) при

$$\frac{\tau}{\bar{T}_1} = \frac{\tau}{\bar{T}_2} = 0,5.$$

В случае трех и более устройств, объединенных в комплекс с общим ОЗУ, среднее время задержки заявок 1-го устройства определяется по формуле (23), где величина первого слагаемого зависит от длительности защитного промежутка времени заявок данного устройства, в то время как второе слагаемое от него не зависит. Следовательно, для обеспечения абсолютного приоритета заявок какого-либо устройства необходимо наряду с заданием определенной величины защитного промежутка времени обеспечить соответствующий порядок опроса каналов блока приоритета, например, в порядке увеличения их номеров. Таким образом, приоритет устройства в этом случае будет определяться как величиной защитного промежутка времени заявок, так и номером канала, по которому они поступают в блок приоритета.