Хусаинов А.А., *доктор физико-математических наук, профессор*
(Комсомольский-на-Амуре государственный университет)
Манохина Н.Н., *студент*
(Амурский гуманитарно-педагогический государственный университет)

ОПТИМАЛЬНАЯ ГЛУБИНА КОНВЕЙЕРА С КОНФЛИКТАМИ

Рассмотрены вопросы, связанные с производительностью вычислительного конвейера, при работе которого происходят конфликты, замедляющие обработку входных элементов данных заданного объема. Каждому конфликту присвоен его тип, определенный как число тактов конвейера, на которое этот конфликт замедляет работу конвейера. При заданных вероятностях типов конфликтов и заданном объеме данных построена аналитическая модель, которая позволяет вычислять математическое ожидание случайной величины времени обработки входных элементов. Оптимальная глубина конвейера при заданном числе входных элементов данных представляет собой минимальное число его ступеней, при которых математическое ожидание времени обработки этих элементов минимально. Показано, каким образом построенная модель применяется для расчета оптимальной глубины конвейера в различных случаях.

Ключевые слова: *Вычислительный конвейер; производительность конвейера; оптимальная глубина; конфликты в конвейере; случайная величина времени обработки; математическое ожидание; среднее время обработки.*

THE OPTIMUM DEPTH FOR A PIPELINE WITH HAZARDS

Problems related to the performance of a computational pipeline are considered, during which random hazards can occur that slow the processing of input data containing a given number of elements (amount of data). We assign to each hazard its own type defined as the number of pipeline cycles to which this hazard

Хусаинов А.А., Манохина Н.Н. Оптимальная глубина конвейера с конфликтами // Естественные и технические науки. 2018. Том 116, №2. С.170-175.

increases the processing time of the input data element. Given the probabilities of hazard types and a given amount of data, an analytical model is constructed that allows one to calculate the mathematical expectation of a random variable equal to processing time of input elements. The optimum depth of the pipeline for a given number of input data elements is the number of stages at which the mathematical expectation of the processing time of these elements is minimal. It is shown how the constructed model should be used in calculating the optimum depth of the pipeline in individual cases.

Keywords: *computational pipeline; pipeline performance; optimum depth; hazards in pipeline; random value of the processing time; mathematical expectation; average processing time.*

В работе рассмотрены вопросы, связанные с производительностью вычислительного конвейера с конфликтами. Построена аналитическая модель, позволяющая вычислять среднее время обработки n элементов при заданных вероятностях конфликтов. С помощью этой модели получены формулы для расчета оптимальной глубины конвейера в случаях, когда известны вероятности конфликтов. Новизна результатов в том, что в этих формулах учитывается объем данных, обрабатываемых конвейером.

Расчет производительности вычислительного конвейера необходим при проектировании процессоров [1]-[3], сигнальных процессоров [4], процессоров для вычисления операций над числами с плавающей точкой [5]. Он имеет большое значение при разработке программного обеспечения на основе многопоточных конвейеров, ступенями которых служат потоки, а также для вычислительных конвейеров, ступенями которых могут служить самые разнообразные компоненты, включая транспьютеры или полноценные компьютеры в кластерных системах [6]. Проблема расчета производительности актуальна. Она связана с нахождением оптимального числа ступеней конвейера.

Этой теме посвящено много работ (см. обзоры [7]-[8]), но в этих работах не учтена зависимость времени обработки от объема данных.

Обработка инструкций с помощью конвейерного процессора и обработка данных с помощью конвейера замедляется в результате конфликтов между обрабатываемыми элементами данных. В работах [9]-[10] была получена формула для оптимального числа ступеней конвейера с рестартами – конфликтами, замедляющие работу конвейера на его полную задержку. Была введена пропускная способность G равномерного конвейера с рестартами и доказана формула:

$$G = \frac{1}{1 + (m-1)b} \cdot \frac{1}{\frac{t_p}{m} + t_o},$$

где m – число ступеней, b – частота рестарта, t_p – тотальная логическая задержка конвейера, t_o – время передачи элемента данных для ступени. Это позволило найти оптимальную глубину равномерного конвейера:

$$m_{opt} = \sqrt{\frac{(1-b)t_p}{bt_o}}.$$

В ряде работ были предложены модели для нахождения среднего времени обработки одного элемента с помощью конвейера. В работе Эммы и Дэвидсона [11] это время называется обратной пропускной способности конвейера BW^{-1} и измеряется в единицах CPI – числе тактов на инструкцию (cycles per instruction). Самая простая модель для нахождения обратной пропускной способности записывается с помощью формулы

$$BW^{-1} = 1 + \sum_{D=0}^{N_E-1} p_D (N_E - D)^+,$$

где N_E – число ступеней исполнительной части конвейера, а p_D – вероятность того, что зависимость по данным существует на расстоянии D для произвольно

выбранной инструкции. Слагаемое при $D = 0$ включено для того, чтобы охватить случай, когда инструкция перехода вычисляет собственное условие перехода. Хартстейн и Пузак [12] заметили, что при измерении в CPI график пропускной способности от глубины конвейера будет описываться прямой линией, и значит эта модель не годится для нахождения оптимальной глубины конвейера. В работе [12] была предложена следующая модель для расчета производительности в единицах TPI (time per instruction) с учетом конфликтов:

$$\frac{T}{N_I} = \frac{t_o}{\alpha} + \frac{\gamma N_H t_p}{N_I} + \frac{t_p}{\alpha p} + \frac{\gamma N_H t_o}{N_I} p \,,$$

где N_I – число входных элементов (машинных команд или данных); T – время обработки N_I элементов; N_H – число конфликтов; t_p – время обработки одного элемента конвейером, без времени записи в буферы; t_o – время обмена ступени конвейера с буфером; α – коэффициент суперскалярности процессора; $\gamma \in [0,1]$ – коэффициент, зависящий от типов конфликтов в конвейере. Это позволило установить, что оптимальная глубина равна

$$p_{opt} = \sqrt{\frac{N_I t_p}{\alpha \gamma N_H t_o}} \,.$$

Полученная формула для оптимальной глубины показывает, что оптимальная глубина не зависит от числа обрабатываемых инструкций, а зависит от отношения числа конфликтов к числу инструкций.

Мы решаем задачу нахождения метода вычисления оптимальной глубины при заданном объеме данных и заданных вероятностях конфликтов произвольных типов.

1. Случайная величина времени обработки n элементов

Вычислительный конвейер состоит из конечной последовательности вычислительных устройств $u_1, u_2, \cdots, u_m$, которые называются ступенями

Хусаинов А.А., Манохина Н.Н. Оптимальная глубина конвейера с конфликтами // Естественные и технические науки. 2018. Том 116, №2. С.170-175.

конвейера, и каналов $c_1, c_2, \cdots, c_m$ для передачи данных. Число ступеней конвейера m называется его глубиной. Каждый элемент входных данных последовательно обрабатывается всеми ступенями в порядке возрастания номеров ступеней. Ступень u_1 читает элемент входных данных, поступающих в конвейер, выполняет над этим элементом вычислительную операцию и результат записывает в канал c_1. Ступень u_2 считывает элемент данных из c_1, выполняет свою операцию и записывает результат в c_2 и т.д. Последняя ступень u_m читает элемент данных из c_{m-1}, выполняет операцию и записывает результат в c_m. Ступени, обрабатывающие различные элементы входных данных, могут работать параллельно.

Конвейер называется равномерным, если время обработки элемента ступенью, состоящее из времени выполнения операции и времени передачи результата операции в выходной канал, одинаково для всех ступеней. Это время называется временем задержки ступени или тактом конвейера. Сумма времен выполнения операций ступеней называется тотальной логической задержкой и обозначается через t_p. Время выполнения операции ступенью называется логической задержкой ступени. Для равномерного конвейера ступени имеют одинаковые логические задержки, которые равны $\frac{t_p}{m}$. Время записи результата операции обозначается через t_o. Время задержки ступени равно $h = t_o + \frac{t_p}{m}$.

Мы будем предполагать, что в каждый момент времени будет работать по крайней мере одна из ступеней. Процессу обработки n элементов с помощью равномерного конвейера соответствует последовательность $(x_1, \cdots, x_n)$, где x_i – разность между временем начала обработки i-го элемента и временем начала обработки (предшествующего) $(i-1)$-го элемента, при $i \geq 2$. Она всегда не меньше задержки одной ступени и не больше, чем сумма задержек всех ступеней конвейера. В данном случае время измеряется в тактах конвейера. Число x_1 равно глубине (количеству ступеней) конвейера.

Хусаинов А.А., Манохина Н.Н. Оптимальная глубина конвейера с конфликтами // Естественные и технические науки. 2018. Том 116, №2. С.170-175.

В случае равномерного конвейера с задержками ступеней равными одному такту, время начала обработки входного элемента отстает от времени начала предшествующего входного элемента на число тактов, равное одному из чисел $1, 2, \cdots, m$. Оно будет равно номеру ступени, после выполнения которой может начинаться обработка i-го элемента (с помощью первой ступени). В частности, если номер ступени равен m, то конфликт будет тотальным. Пусть b_1 – вероятность того, что нет конфликта. В этом случае отставание равно 1, ибо начальная обработка может выполняться только после срабатывания первой ступени. Для $1 \leq j \leq m$ обозначим через b_j вероятность того, что текущий элемент начинает обрабатываться через время j после начала обработки предшествующего элемента.

Если все ступени конвейера имеют равную задержку h=1, и число ступеней равно m, то приходим к задаче нахождения математического ожидания случайной величины ξ, определенной на множестве конечных последовательностей $\omega = (x_1, \cdots, x_n)$, состоящих из символов $x_i \in \{a_1, \cdots, a_m\}$, для всех $1 \leq i \leq m$, и $\xi(x_1, \cdots, x_n) = k_1 + 2k_2 + \cdots + mk_m$, где k_1 – число символов a_1 в ω, …, k_m – число символов a_m в ω.

Важно заметить, что первый элемент x_1 равен рестарту a_m и обрабатывается за время m. Поэтому случайная величина времени обработки n элементов конвейером равна $\Theta_n = m + \xi_{n-1}$, а математическое ожидание времени обработки n элементов будет равно

$M\Theta_n = M\xi = m + \sum_\omega P(x_1, \dots, x_{n-1})\xi(x_1, \dots, x_{n-1})$. Таким образом, мы приходим к задаче о нахождении математического ожидания случайной величины, определенной на последовательностях $n - 1$ независимых испытаний по схеме Бернулли с m исходами. Поскольку $P(x_1, \dots, x_{n-1}) = b_1^{k_1} \cdots b_m^{k_m}$, где k_m на единицу меньше числа конфликтов типа m, то получаем следующую предварительную формулу для математического ожидания времени обработки n элементов:

$$M\Theta_n = m + \sum_{k_1+\cdots+k_m=n-1} P_{n-1}(k_1,\dots,k_m) b_1^{k_1}\cdots b_m^{k_m}(k_1+2k_2+\cdots+m\,k_m),$$

где $P_{n-1}(k_1,\dots,k_m) = \frac{(n-1)!}{k_1!\cdots k_m!}$ обозначает число упорядоченных разбиений множества, состоящего из $n-1$ элементов на m подмножеств фиксированных размеров $k_1,\cdots,k_m$.

2. Среднее значение случайной величины времени обработки n элементов

Используя соотношение $\sum_{k=0}^{n} kC_n^k x^{k-1} = \frac{d}{dx}\sum_{k=0}^{n} C_n^k x^k = n(1+x)^{n-1}$, получаем следующее вспомогательное равенство:

Лемма 1. $\sum_{k=0}^{n} kC_n^k x^k (1-x)^{n-k} = nx.$

Доказательство. С помощью преобразований левой части равенства получим $(1-x)^{n-1}\sum_{k=0}^{n} kC_n^k\, x\left(\frac{x}{1-x}\right)^{k-1} = (1-x)^{n-1}x\,n\left(1+\frac{x}{1-x}\right)^{n-1} = nx.$

Предложение 1. Имеет место равенство

$$\sum_{k_1+\cdots+k_m=n} P_n(k_1,\dots,k_m) b_1^{k_1}\cdots b_m^{k_m} k_1 = nb_1$$

Доказательство. Сумма равна

$$\sum_{k_1=0}^{n} k_1 C_n^{k_1} b_1^{k_1} \sum_{k_2+\cdots+k_m=n-k_1} P_{n-k_1}(k_2,\dots,k_m) b_2^{k_2}\cdots b_m^{k_m} =$$

$$= \sum_{k_1=0}^{n} k_1 b_1^{k_1} C_n^{k_1} (b_2+\cdots+b_m)^{n-k_1} = \sum_{k_1=0}^{n} k_1 C_n^{k_1} b_1^{k_1}(1-b_1)^{n-k_1}.$$

Откуда с помощью леммы 1 приходим к искомому равенству.

Теорема 1. Математическое ожидание случайной величины времени обработки n элементов с помощью равномерного конвейера, состоящего из m

ступеней, равно $M\Theta_n = \big(m + (n-1)(b_1 + 2b_2 + \cdots + mb_m)\big)h$, где b_i – вероятность i-го конфликта, а h - время задержки ступени (время такта конвейера).

Доказательство. Для нахождения $M\Theta_n$ применим предварительную формулу для математического ожидания времени обработки n элементов (приведенную в п.1), а затем воспользуемся теоремой 1, подставляя вместо b_1 вероятности остальных конфликтов. Умножая левые части полученных равенств на коэффициенты $1, 2, \cdots, m$ и складывая, получим искомое.

Пример 1. Для конвейера, все конфликты которого - рестарты, при вероятности рестарта $b = b_m$ и времени задержки ступени $h = 1$ получим формулу $M\Theta_n = m + (n-1)(1 - b + mb)$.

3. Метод нахождения оптимальной глубины

Рассмотрим конвейер, полученный с помощью разложения некоторой функции в композицию ступеней. Предположим, что время обработки элемента с помощью этой функции равно t_p, и пусть на обмен данными каждая ступень расходует время t_o. Предположим, что время выполнения операций ступеней одинаково. Тогда один элемент обрабатывается ступенью за время $h = t_o + \frac{t_p}{m}$, где m – число ступеней. Среднее значение (математическое ожидание) времени обработки n элементов с помощью конвейера будет равно

$$T(m) = \big(m + (n-1)(b_1 + 2b_2 + \cdots + mb_m)\big)\big(t_o + \frac{t_p}{m}\big).$$

Мы будем предполагать, что задана вероятность b возникновения конфликта, и вероятности $b_2, b_3, \cdots, b_m$ равны между собой, причем каждая из них равна $\frac{b}{m-1}$. Положим $b_1 = 1 - b$. Получим *конвейер с одинаковыми временами конфликтов ступеней*.

Хусаинов А.А., Манохина Н.Н. Оптимальная глубина конвейера с конфликтами // Естественные и технические науки. 2018. Том 116, №2. С.170-175.

Наша задача – при заданных b, t_o, t_p, n найти глубину m конвейера, при которой среднее время обработки n элементов минимально. Эта глубина называется оптимальной.

Замечание 1. Мы рассматриваем условия, которым должен удовлетворять класс конвейеров, не связанные с последовательностями входных элементов. Существует задача нахождения оптимальной глубины конвейера для класса последовательностей входных элементов с заданным набором вероятностей существования зависимых элементов в этих последовательностях. Формула для $T(m)$ позволяет построить модель для решения этой задачи, если вместо вероятностей b_i, для всех $2 \leq i \leq m$, подставить p_{m-i+1} - вероятности того, что на расстоянии $m - i + 1$ существует элемент, от которого зависит произвольно выбранный элемент. Определение расстояния дано в [11].

Следствие 1. Оптимальная глубина равномерного конвейера с одинаковыми вероятностями конфликтов ступеней равна

$$m_{\text{opt}} = \sqrt{\frac{(n-1)t_p}{\left(1 + \frac{n-1}{2}b\right)t_o}}.$$

Доказательство. Поскольку $b_1 + \cdots + b_m = 1$ и $b_2 = \cdots = b_m = \frac{b}{m-1}$, то

$$T(m) = \left(m + (n-1)(b_1 + 2b_2 + \cdots + mb_m)\right)\left(t_o + \frac{t_p}{m}\right) =$$

$$= \left(m + (n-1)\left(1 + \frac{b}{m-1}\left(1 + 2 + \cdots + (m-1)\right)\right)\right)\left(t_o + \frac{t_p}{m}\right) =$$

$$= \left(m + (n-1)\left(1 + \frac{bm}{2}\right)\right)\left(t_o + \frac{t_p}{m}\right).$$

Хусаинов А.А., Манохина Н.Н. Оптимальная глубина конвейера с конфликтами // Естественные и технические науки. 2018. Том 116, №2. С.170-175.

Приравнивая производную нулю $\frac{dT}{dm} = t_o + \frac{(n-1)b}{2} t_o + (n-1)(-\frac{t_p}{m^2}) = 0$, получим равенство $m^2 = \frac{t_p(n-1)}{t_o(1+\frac{(n-1)b}{2})}$, приводящее к искомой формуле.

Следствие 2. При $n \to \infty$ получаем оптимальную глубину конвейера с одинаковыми вероятностями конфликтов $m_{opt} = \sqrt{\frac{2t_p}{bt_o}}$ без учета объема данных.

Следствие 3. Если вероятности $b_1, \cdots, b_m$ равны между собой, то

$$m_{opt} = \sqrt{\frac{(n-1)t_p}{(n+1)t_o}}.$$

Доказательство. В случае, когда вероятности $b_1, b_2, \cdots, b_m$ равны, получим $b_i = \frac{1}{m}$, откуда среднее время обработки $T(m)$ будет равно

$$\big(m + (n-1)(b_1 + 2b_2 + \cdots + mb_m)\big)\big(t_o + \frac{t_p}{m}\big) = (m + \frac{(n-1)(m+1)}{2})(t_o + \frac{t_p}{m}).$$

Поскольку в этом случае производная $\frac{d}{dm}T(m)$ будет равна $\frac{1}{2}((n+1)t_o - \frac{(n-1)t_p}{m^2})$. Уравнение $\frac{dT}{dm} = 0$ приводит к $m_{opt}^2 = \frac{(n-1)t_p}{(n+1)t_o}$.

В работе [13, следствие 2] было доказано следующее утверждение с помощью модели для среднего времени обработки n элементов, отличной от представленной в теореме 1.

Следствие 4. Если каждый конфликт приводит к рестарту, и вероятность конфликта равна b, то оптимальная глубина равна $m_{opt} = \sqrt{\frac{(n-1)(1-b)t_p}{(1+(n-1)b)t_o}}$.

Доказательство. В этом случае вероятности равны $b_m = b$ и $b_1 = 1 - b$. С помощью теоремы 1 получим

$$T(m) = \big(m + (n-1)(b_1 + 2b_2 + \cdots + mb_m)\big)\big(t_o + \frac{t_p}{m}\big) =$$

$$= \big(m + (n-1)(1 - b + mb)\big)\Big(t_o + \frac{t_p}{m}\Big).$$

Производная будет равна $\frac{d}{dm}T(m) = t_o + (n-1)bt_o - \frac{(n-1)(1-b)t_p}{m^2}$. Приравнивая ее нулю, приходим к доказываемому утверждению.

Заключение

Мы получили формулу для среднего значения случайной величины времени обработки n элементов с помощью конвейера, имеющего m ступеней (теорема 1). Это позволило нам найти оптимальное число ступеней в нескольких случаях (следствия 1–4). В перспективе развитие полученных результатов для других вычислительных систем с конвейерным параллелизмом: псевдо-конвейеров, двумерных конвейеров, волновых процессоров.

ЛИТЕРАТУРА

1 Коуги, П.М. Архитектура конвейерных ЭВМ [Текст]: Пер. с англ. / П.М. Коуги – М.: Радио и связь, 1985. – 360 с.

2 Patterson, D.A. Computer Organization and Design / D.A. Patterson, J. L. Hennessy – Amsterdam: Elsevier, 2014. – 793 p.

3 Хамахер, К. Организация ЭВМ / К. Хамахер, З. Вранешич, С. Заки – 5-е изд. СПб.: Питер; Киев: Издательская группа BHV, 2003. - 848 с.

4 Беляев, А.А. Теория, разработка и создание проблемно-ориентированных процессорных ядер с оптимальным вычислительным конвейером и многоядерных сигнальных процессоров на их основе: дис. ... доктора техн. наук. - М.: ОАО Научно-производственный центр ``Электронные вычислительно-информационные системы'', 2012. - 377 с.

5 Merchant, F. Accelerating BLAS and LAPACK via Efficient Floating Point Architecture Design / F. Merchant, A. Chattopadhyay, S. Raha, S.K. Nandy, R.

Narayan. Preprint, arxiv: 1610.08705v2 [cs.AR] - New York: Cornell University Library, 2016. - 7 p.

6 Горшенин А.К. Параллелизм в микропроцессорах / А.К. Горшенин, С.В. Замковец, В. Н. Захаров // Системы и средства информ. 2014. Т. 24, № 1. С. 46-60.

7 Hartstein A. Optimum power/performance pipeline depth / A. Hartstein, T. R. Puzak – Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture. Washington: IEEE Computer Society, 2003. P. 117-128.

8 Yao, J. Optimal pipeline depth with pipeline stage unification adoption / J. Yao, S. Miwa, H. Shimada // ACM SIGARCH Computer Architecture News. Vol. 35, No. 5. 2007. P. 3-9.

9 Dubey, P.K. Optimal Pipelining / P.K. Dubey, M.J. Flynn // J. Parallel and Distributed Computing. 1990. Vol. 8, No. 1. P. 10-19.

10 Flynn, M.J. Deep-Submicron Microprocessor Design Issues / M.J. Flynn, P. Hung, K.W. Rudd // IEEE Micro. 1999. Vol. 19, No. 4. P. 11-22.

11 Emma, P.G. Characterization of Branch and Data Dependencies in Programs for Evaluating Pipeline Performance / P.G. Emma, E.S. Davidson // IEEE Transactions on Computers, 1987. Vol. 7. P. 859-875.

12 Hartstein, A. The optimum pipeline depth for a microprocessor / A. Hartstein, T.R. Puzak // ACM Sigarch Computer Architecture News. IEEE Computer Society, 2002. Vol. 30, No. 2. P. 7-13.

13 Хусаинов, А.А. Оптимальная глубина вычислительного конвейера при заданном объеме входных данных / А.А. Хусаинов, Е.А. Титова // Вычислительные технологии, 2018. Т.23, № 1. С. 96-104.