Оптимальная глубина вычислительного конвейера при заданном объеме входных данных

А. А. Хусаинов*, Е. А. Титова
Комсомольский-на-Амуре государственный университет, Россия
*Контактный e-mail: husainov51@yandex.ru

Рассмотрен вычислительный конвейер, ступени которого могут иметь различные временные задержки, а время обмена с каналами одинаково. При заданной вероятности того, что входной элемент данных вызывает рестарт, построена аналитическая модель для расчета вероятнейшего и среднего значений времени обработки заданного числа входных элементов данных. На основании этой модели показано, что при заданных числе ступеней и тотальной логической задержке конвейер будет иметь максимальную производительность тогда и только тогда, когда задержки ступеней равны между собой. Для конвейера с одинаковыми задержками ступеней получена поправка известной формулы Доби — Флинна для оптимальной глубины вычислительного конвейера при заданной вероятности рестарта с учетом числа обрабатываемых элементов. Приведен пример, показывающий, что эта поправка может быть существенной. Показано, что эта формула дает оптимальную глубину на множестве всех конвейеров, включая конвейеры с различными задержками ступеней.

Ключевые слова: вычислительный конвейер, производительность, тотальный конфликт, рестарт конвейера, вероятнейшее значение, математическое ожидание, глубина конвейера.

Введение

Расчет производительности вычислительного конвейера имеет большое значение при проектировании процессоров [1, 2], сигнальных процессоров [3], процессоров для вычисления операций над числами с плавающей точкой [4]. Он полезен при разработке программного обеспечения, содержащего многопоточные конвейеры, ступенями которых служат потоки [5]. Расчет производительности необходим для вычислительных конвейеров, созданных на системах с распределенной памятью, ступенями которых могут служить самые разнообразные компоненты, включая транспьютеры и полноценные компьютеры [6].

Время обработки данных с помощью конвейера зависит от числа ступеней. К сожалению, уравнение для нахождения оптимального числа ступеней может не иметь решений в целых числах. Возможно, в связи с этим было введено определение, согласно которому глубиной конвейера называется вещественное число, близкое к количеству ступеней этого конвейера.

Мы рассматриваем задачу построения аналитической модели для расчета времени обработки входных данных заданного объема и на основе ее решения уточняем формулу для оптимальной глубины с учетом объема входных данных.

Равномерный конвейер состоит из ступеней, имеющих одинаковое время обработки одного элемента и одинаковое время записи результата обработки в выходной регистр. Время обработки одного элемента с помощью конвейера (тотальная логическая задержка) обозначено через t_p. Если время записи в выходной регистр t_o, а глубина конвейера равна p, то задержка ступени будет равна $t_p/p + t_o$. В работах [7, 8] описана формула для оптимальной глубины равномерного конвейера в случае известной вероятности рестарта $0 < b < 1$:

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

При доказательстве этой формулы неявно предполагалось, что объем входных данных бесконечен. Возникает естественный вопрос, как зависит оптимальная глубина от объема данных, поступающих в конвейер.

В работе [9] получена аналитическая модель для расчета производительности конвейера, входными данными которого являются машинные команды, обрабатываемые конвейером. Была получена формула для оптимальной глубины конвейера

$$\widetilde{p}_{opt} = \sqrt{\frac{N_I t_p}{\alpha\gamma N_H t_0}}, \tag{2}$$

где N_I — число входных элементов, обработанных без конфликтов; N_H — число конфликтов; α — средний "коэффициент суперскалярности" процессора [3]; $\gamma \in [0, 1]$ — коэффициент, зависящий от вида конфликтов в конвейере. Если $\gamma = 1$, то всякий конфликт является рестартом. Пусть $d = N_H/(N_I + N_H)$ — доля конфликтов. Тогда в случае $\alpha = \gamma = 1$ приходим к формуле, которую можно получить из формулы (1), подставляя вместо вероятности рестарта b долю конфликтов d. Формула будет показывать, что оптимальная глубина зависит не от числа входных элементов, а от доли конфликтов. В [3] указано, что при выводе формулы (2) в работе [9] были допущены некоторые пробелы. В результате получилось, что оптимальная глубина зависит только от отношения числа конфликтов к числу обрабатываемых элементов.

В настоящей работе рассмотрены конвейеры, все конфликты которых приводят к рестартам. Приведены начальные сведения из теории вероятностей, описаны асинхронные вычислительные конвейеры, введена случайная величина времени обработки n элементов с помощью конвейера с рестартами. Построены аналитические модели для расчета вероятнейшего значения и математического ожидания этой случайной величины. С помощью вспомогательных утверждений 1 и 2 получены формулы для расчета глубины равномерного конвейера, при которой вероятнейшее значение этой случайной величины минимально (следствие 1), и для расчета глубины, при которой минимально математическое ожидание этой случайной величины (следствие 2). Доказана формула

$$\overline{p}_{opt}(n) = \sqrt{\frac{(1-b)t_p}{\left(b + \dfrac{1}{n-1}\right)t_0}}.$$

При $n \to \infty$ она приводит к формуле (1). Тем не менее она показывает, что для всякого $n > 1$ существует вероятность рестарта $b < 1$, делающая отношение глубин $p_{opt}/\overline{p}_{opt}(n)$ сколь угодно большим.

1. Предварительные сведения

1.1. Вероятнейшее и среднее значения конечной случайной величины

Конечное поле вероятностей задается как произвольное конечное множество Ω элементарных событий вместе с функцией $\mathbb{P} : \Omega \to \mathbb{R}$, удовлетворяющей условиям $\mathbb{P}(\omega) \geq 0$ для всех $\omega \in \Omega$ и $\sum\limits_{\omega\in\Omega} \mathbb{P}(\omega) = 1$. Для каждого $\omega \in \Omega$ значение $\mathbb{P}(\omega)$ называется его вероятностью. Конечной случайной величиной на Ω называется произвольная функция $\xi : \Omega \to \mathbb{R}$, определенная на конечном множестве. Для всякой конечной случайной величины ξ и произвольного числа $r \in \mathbb{R}$ определено число $\mathbb{P}[\xi = r] =_{def} \sum\limits_{\omega\in\xi^{-1}(r)} \mathbb{P}(\omega)$, которое интерпретируется как вероятность того, что ξ принимает значение r.

Например, для последовательности n независимых испытаний по схеме Бернулли в смысле [10, § 9] элементарное событие можно рассматривать как последовательность, состоящую из n нулей и единиц (единица соответствует благоприятному исходу испытания, а нуль — неблагоприятному). Отсюда число элементарных событий равно 2^n. Пусть b — вероятность благоприятного исхода в испытании. Тогда вероятность элементарного события, имеющего k единиц, равна $b^k(1-b)^{n-k}$. Это поле событий обозначим через $\Omega_{n,b}$. Функция $\xi : \Omega_{n,b} \to \mathbb{R}$ сопоставляет каждой последовательности нулей и единиц количество единиц. Вероятность того, что $\xi(\omega) = k$, будет равна $P_n(k) = C_n^k b^k (1-b)^{n-k}$.

Пусть ξ — конечная случайная величина. Ее математическое ожидание (или среднее значение) определяется как $\mathbb{M}\xi = \sum\limits_{\omega\in\Omega} \xi(\omega)\mathbb{P}(\omega)$. Ее вероятнейшим значением называется число $r \in \mathbb{R}$, для которого число $\mathbb{P}[\xi = r]$ максимально. Для каждого $x \in \mathbb{R}$ через $[x]$ будем обозначать целую часть числа x.

В дальнейшем нам понадобятся следующие сведения, доказательство которых содержится в [10, гл. 2, § 10] и [11, гл. IV, § 3]: пусть ξ_n — конечная случайная величина на $\Omega_{n,b}$, значения которой на каждой последовательности n испытаний по схеме Бернулли равны числу благоприятных исходов. Тогда верны следующие утверждения:

- Вероятнейшее значение ξ_n равно $[(n+1)b]$, если $(n+1)b$ не является целым. Если $(n+1)b$ — целое, то ξ_n будет иметь два вероятнейших значения: $(n+1)b-1$ и $(n+1)b$.
- Математическое ожидание случайной величины ξ_n равно $\mathbb{M}\xi_n = nb$.

1.2. Асинхронный вычислительный конвейер и время обработки данных

Под входными элементами мы подразумеваем инструкции или элементы массива данных, которые подаются на вход конвейера.

Вычислительный конвейер состоит из конечной последовательности

$$c_0 \xrightarrow{a_0} c_1 \xrightarrow{a_1} c_2 \to \cdots c_{p-1} \xrightarrow{a_{p-1}} c_p$$

каналов передачи данных c_i, $0 \leq i \leq p$, и вычислительных устройств a_j, $0 \leq j \leq p-1$. Вычислительные устройства называются также ступенями конвейера. Здесь p — некоторое произвольное положительное целое число. Канал c_0 называется входным каналом конвейера, а c_p — выходным.

Вычислительный конвейер называется асинхронным, если синхронизация работы ступеней управляется с помощью потока данных. Это означает, что каждое вычислительное устройство a_i, $0 \le i \le p-1$, выполняет цикл, состоящий из последовательных шагов:

- ожидание готовности данных для чтения из канала c_i;
- прием элемента данных из канала c_i;
- выполнение вычислительной операции;
- запись результата вычислительной операции в канал c_{i+1}.

Если на вход конвейера поступили n входных элементов, то для каждой из p ступеней число итераций будет равно n.

Компьютерная модель вычислительного конвейера, на основе которой проводились испытания, является многопоточным приложением. Потоки играют роль ступеней. Каналы организованы как каналы UNIX. Канал состоит из очереди, для которой запись и чтение осуществляются с помощью алгоритма Дейкстры для задачи о производителе и потребителе. В начале работы загружаются все потоки (ступени). Затем во входной канал загружаются n элементов данных. Каждый из потоков содержит цикл, аналогичный циклу вычислительного устройства. В конце работы все данные будут находиться в выходном канале. Конвейер, построенный с помощью многопоточного приложения, будет асинхронным.

Вычислительные устройства выполняют вычислительные операции параллельно. Они также могут одновременно производить чтение и запись, если не имеют общих каналов. Тем не менее для каждого $i = 0, \ldots, p$ чтение из канала c_i и запись в канал c_i никакие ступени не могут производить одновременно.

Обозначим через $\tau(a_i)$ время выполнения операции, выполняемой i-й ступенью. Пусть $r(c_i)$ — время чтения из канала c_i, а $w(c_i)$ — время записи в канал c_i. Заметим, что при проектировании конвейерных процессоров можно полагать $r(c_i) = 0$. Но в любом случае канал не освобождается до тех пор, пока находящиеся в нем данные не будут обработаны ступенью, читающей из него. Время обработки одного элемента данных с помощью i-й ступени при $0 \le i \le p-1$ равно $\tau_i = r(c_i) + \tau(a_i) + w(c_{i+1})$. Оно называется задержкой i-й ступени.

Обозначим

$$\sigma = \sum_{i=0}^{p-1} (r(c_i) + \tau(a_i) + w(c_{i+1})), \quad \mu = \max_{0 \le i \le p-1} (r(c_i) + \tau(a_i) + w(c_{i+1})).$$

Как известно, для равномерного конвейера, задержки ступеней которого равны h, а число ступеней p, время обработки n элементов можно вычислять по формуле $T(n) = (p+n-1)h$. Этот факт имеет следующее обобщение, простое доказательство которого можно найти в [12]: минимальное время обработки n входных элементов с помощью асинхронного вычислительного конвейера равно

$$T(n) = \sigma + (n-1)\mu. \tag{3}$$

2. Оптимальный конвейер

2.1. Вероятнейшее и среднее значения времени обработки данных

Формула (3) имеет место в том случае, когда обработка последовательности n входных элементов производится независимо различными ступенями конвейера и конвейер рабо-

тает непрерывно. В общем случае ступени могут конфликтовать. Существует несколько видов конфликтов для конвейера обработки команд процессора [2]. Нас будут интересовать конфликты, вызывающие рестарт.

Элемент входных данных (или инструкций) вызывает рестарт или тотальный конфликт, если его обработка первой ступенью конвейера производится после завершения обработки предшествующего элемента каждой ступенью конвейера. Ниже повсюду b означает вероятность того, что входной элемент, не являющийся самым первым из n входных элементов, вызывает рестарт конвейера. Обработке $n \geq 1$ входных элементов с помощью конвейера можно сопоставить последовательность длины $n-1$, состоящую из нулей и единиц, где k-му элементу сопоставляется 1, если он вызвал рестарт, и 0 в противном случае. Рассмотрим конечное поле вероятностей, элементарные события которого состоят из этих последовательностей. Вероятность элементарного события равна $b^k(1-b)^{n-k-1}$, где k — число рестартов в этом событии, $0 \leq k \leq n-1$.

Согласно формуле (3) каждый входной элемент, вызывающий рестарт, добавляет к времени обработки число σ. Если он не вызывает рестарт, то добавляется время μ. Время обработки первого элемента равно σ. Отсюда следует, что время обработки n входных элементов равно $\sigma + k\sigma + (n-1-k)\mu = \sigma + (n-1)\mu + k(\sigma-\mu)$, где k — число рестартов. Отсюда вытекает, что время обработки n входных элементов конвейером можно рассматривать как случайную величину $\Theta_n : \Omega_{n-1,b} \to \mathbb{R}$, значение которой на элементарных событиях, содержащих k рестартов, равно $\sigma + (n-1)\mu + k(\sigma-\mu)$. Вероятность того, что эта случайная величина принимает значение $\sigma+(n-1)\mu+k(\sigma-\mu)$, будет равна $C_{n-1}^k b^k(1-b)^{n-k-1}$. Отсюда получаем следующие аналитические модели для расчета вероятнейшего и среднего времен обработки n элементов конвейером:

- Вероятнейшее время обработки n входных элементов с помощью конвейера равно $\sigma + (n-1)\mu + [nb](\sigma-\mu)$. В случае целого nb существует второе вероятнейшее значение $\sigma + (n-1)\mu + (nb-1)(\sigma-\mu)$.
- Среднее время обработки n элементов равно

$$\mathbb{M}\Theta_n = \sigma + (n-1)\mu + (n-1)b(\sigma-\mu).$$

Здесь $b < 1$ — вероятность того, что входной элемент вызывает рестарт конвейера.

Докажем вторую формулу. С этой целью рассмотрим случайную величину ξ_{n-1} на $\Omega_{n-1,b}$, значения которой на каждой последовательности $n-1$ испытаний по схеме Бернулли равны числу благоприятных исходов. Согласно изложенному в подразд. 1.1, $\mathbb{M}\xi_{n-1} = (n-1)b$. Имеет место равенство

$$\Theta_n(\omega) = \sigma + (n-1)\mu + (\sigma-\mu)\xi_{n-1}(\omega).$$

Отсюда

$$\mathbb{M}\Theta_n = (\sigma + (n-1)\mu) + (\sigma-\mu)\mathbb{M}\xi_{n-1} = \sigma + (n-1)\mu + (n-1)b(\sigma-\mu).$$

2.2. Зависимость оптимальной глубины конвейера от объема данных

Некоторые утверждения, доказанные ниже, будут справедливы как для математического ожидания дискретной случайной величины, так и для вероятнейшего значения. В связи с этим введем обозначение $\eta(n)$ как для вероятнейших значений случайной величины, равной количеству благоприятных исходов в $n-1$ испытаниях по схеме Бернулли, так и для математического ожидания этой случайной величины. Легко видеть,

что $\eta(n) \in \{nb-1, nb, (n-1)b\}$ в случае целых nb и $\eta(n) \in \{[nb], (n-1)b\}$ в остальных случаях.

Нам известно, что вероятнейшее (среднее) значение времени обработки n элементов с помощью конвейера равно

$$T_\eta(n) = \sigma + (n-1)\mu + \eta(n)(\sigma - \mu),$$

где $\eta(n)$ — вероятнейшее (среднее) значение случайной величины $\xi_{n-1} : \Omega_{n-1,b} \to \mathbb{R}$ (см. подразд. 1.1).

Пусть $p > 1$ — число ступеней конвейера. Будем предполагать, что время обмена ступеней с каналами $r(c_i) + w(c_{i+1})$ одинаково для всех $i = 0, \ldots, p-1$. Обозначим это общее значение через t_o. Пусть $t_p = \sum\limits_{i=0}^{p-1} \tau(a_i)$ — так называемая тотальная логическая задержка конвейера. Она будет равна $\sigma - pt_o$. Буква "p" в обозначении t_p происходит от слова pipeline, а буква "o" в t_o — от слова overhead [4, 9].

Утверждение 1. *При заданных объеме данных n, количестве ступеней p, логической задержке конвейера t_p, вероятности рестарта $b < 1$ и времени обмена t_o значения $T_\eta(n)$ будут минимальными тогда и только тогда, когда $\tau(a_i) = t_p/p$ для всех $0 \le i \le p-1$. В этом случае они будут равны*

$$\overline{T}_\eta(n) = (pt_o + t_p)(1 + \eta(n) + \frac{n-1-\eta(n)}{p}).$$

Доказательство. Поскольку μ — максимальное от p неотрицательных чисел, а σ — их сумма, то имеет место неравенство $\mu \ge \sigma/p$. Оно превращается в равенство тогда и только тогда, когда $\tau_i = \mu$ для всех $0 \le i \le p-1$. Так как $b < 1$, то $[nb] \le n-1$. Тогда

$$T_\eta(n) = \sigma + (n-1-\eta(n))\mu + \eta(n)\sigma \ge \sigma + (n-1-\eta(n))\sigma/p + \eta(n)\sigma.$$

Отсюда следует, что наименьшее значение $T_\eta(n)$, взятое при различных наборах чисел τ_i, будет достигнуто в тех и только тех случаях, когда $\tau_i = \sigma/p$ для всех i из диапазона $0 \le i \le p-1$. Отсюда получаем $\tau(a_i) = t_p/p$ и искомую формулу для $\overline{T}_\eta(n)$. ■

Задача сводится к исследованию вероятнейшего и среднего времени для конвейера, все ступени которого имеют одинаковые логические задержки t_p/p и равные времена обращения к каналам $t_o = r(c_i) + w(c_{i+1})$. Теперь найдем p, при котором вероятнейшее время минимально.

Утверждение 2. *При заданных n, b, t_p, t_o глубина равномерного конвейера, при которой время обработки n элементов при $\eta(n)$ рестартах минимально, равна*

$$p_\eta(n) = \sqrt{\frac{(n-1-\eta(n))t_p}{(1+\eta(n))t_o}}. \tag{4}$$

Доказательство. Мы установили, что при фиксированных n, b, t_p, t_o время обработки n элементов при $\eta(n)$ рестартах будет функцией от p, принимающей значения

$$(pt_o + t_p)\left(1 + \eta(n) + \frac{n-1-\eta(n)}{p}\right).$$

Приравнивая производную этой функции нулю, получим уравнение для нахождения p, при котором эта функция имеет минимум:

$$p_\eta^2(n) = \frac{(n-1-\eta(n))t_p}{(1+\eta(n))t_o}.$$

Это приводит к доказываемой формуле (4). ■

Рассмотрим глубину, при которой вероятнейшее время обработки n элементов минимально. Если получатся два значения глубины с этим свойством, то будем брать меньшее из них. Полученную глубину будем называть оптимальной.

Следствие 1. *При заданных b, t_p, t_o и n оптимальная глубина равномерного конвейера равна*

$$\widetilde{p}_{opt}(n) = \sqrt{\frac{(n-1-[nb])t_p}{(1+[nb])t_o}}.$$

Подставляя $\eta(n) = (n-1)b$, получаем следующее следствие.

Следствие 2. *Глубина равномерного конвейера, при которой математическое ожидание времени обработки n входных элементов конвейером минимально, равна*

$$\overline{p}_{opt}(n) = \sqrt{\frac{(1-b)t_p}{\left(b+\dfrac{1}{n-1}\right)t_o}}.$$

Пример. Сравним оптимальную глубину конвейера

$$p_{opt} = \sqrt{\frac{(1-b)t_p}{bt_o}},$$

принадлежащую Доби и Флинну [7], с глубиной, полученной в теореме 2. Отношение будет равно

$$\frac{p_{opt}}{\overline{p}_{opt}(n)} = \sqrt{1+\frac{1}{b(n-1)}}.$$

Для любого $n > 1$ существует вероятность b, делающая это отношение сколь угодно большим. Например, при числе инструкций $n = 11$ и вероятности рестарта $b = 1/30$ отношение будет равно $p_{opt}/\overline{p}_{opt}(n) = 2$. В этом случае оптимальная глубина, полученная в теореме 2, в два раза меньше оптимальной глубины, полученной по формуле (1).

Следствие 3. *При заданной тотальной логической задержке t_p, времени обмена с каналом t_o, вероятности рестарта $b < 1$ и объеме данных n конвейер, имеющий минимальное вероятнейшее (среднее) время обработки n элементов, будет равномерным конвейером с глубиной $\widetilde{p}_{opt}(n)$ (соответственно $\overline{p}_{opt}(n)$).*

Доказательство. Пусть конвейер, имеющий минимальное $T(n)$, состоит из p ступеней с логическими задержками ступеней $\tau(a_0), \cdots, \tau(a_{p-1})$. Тогда, согласно утверждению 1, логические задержки ступеней будут равны между собой. Доказываемое утверждение вытекает из следствий 1 и 2. ■

Заключение

Для конвейера с рестартами, ступени которого имеют различные временные задержки, построена аналитическая модель расчета вероятнейшего и среднего значений времени обработки n элементов. Показано, что эти вероятнейшее и среднее значения времени будут минимальными тогда и только тогда, когда задержки ступеней равны между собой. Получены формулы для расчета глубины конвейера, при которой вероятнейшее (соответственно среднее) значение времени обработки n элементов минимально.

Данная работа появилась в ходе компьютерных экспериментов, основанных на многопоточных приложениях (см. подразд. 1.2). Просматриваются два направления развития полученных результатов. Во-первых, надо будет учесть любые конфликты, для которых заданы вероятности. Второе направление — обобщить полученные результаты на другие вычислительные системы с конвейерным параллелизмом: псевдоконвейеры, двумерные конвейеры, волновые процессоры.

Список литературы / References

[1] **Patterson, D.A., Hennessy, J.L.** Computer organization and design. Amsterdam: Elsevier, 2014. 793 p.

[2] **Хамахер К., Вранешич З., Заки С.** Организация ЭВМ. СПб.: Питер; Киев: Изд. группа BHV, 2003. 848 с.
Hamacher, C., Vranesic, Z., Zaky, S. Computer organisation. Boston: McGraw Hill Comp., 2002. 818 p.

[3] **Беляев А.А.** Теория, разработка и создание проблемно-ориентированных процессорных ядер с оптимальным вычислительным конвейером и многоядерных сигнальных процессоров на их основе: Дис. ... д-ра техн. наук. М.: ОАО НПЦ "Электронные вычисл.-информ. сист.", 2012. 377 с.
Belyaev, A.A. Theory, development and design of the problem-oriented processor cores with optimal computational pipeline and multi-core signal processors based on them: Dis. ... d-ra tekhn. nauk. M.: OAO NPZ "Elektronnye Vychisl.-Inform. Sist.", 2012. 377 p. (In Russ.)

[4] **Merchant, F., Chattopadhyay, A., Raha, S., Nandy, S.K., Narayan, R.** Accelerating BLAS and LAPACK via efficient floating point architecture design. Preprint, arxiv: 1610.08705v2 [cs.AR]. New York: Cornell Univ. Libr., 2016. 7 p.

[5] **Горшенин А.К., Замковец С.В., Захаров В.Н.** Параллелизм в микропроцессорах // Системы и средства информатики. 2014. Т. 24, № 1. С. 46–60.
Gorshenin, A.K., Zamkovets, S.V., Zakharov, V.N. Parallelism in Microprocessors // Systems and Means of Informatics. 2014. Vol. 24, No. 1. P. 46–60. (In Russ.)

[6] **Зайцев Г.В.** Вычислительный макроконвейер с переменным тактом работы // Цифровая обработка сигналов. 2006. № 1. С. 38–44.
Zaytsev, G.V. Computational macro pipeline with variable operation cycle // Digital Signal Processing. 2006. No. 1. P. 38–44. (In Russ.)

[7] **Dubey, P.K., Flynn, M.J.** Optimal pipelining // J. Parallel and Distributed Computing. 1990. Vol. 8, No. 1. P. 10–19.

[8] **Flynn, M.J., Hung, P., Rudd, K.W.** Deep-submicron microprocessor design issues // IEEE Micro. 1999. Vol. 19, No. 4. P. 11–22.

[9] **Hartstein, A., Puzak, T.R.** The optimum pipeline depth for a microprocessor // ACM Sigarch Computer Architecture News. 2002. Vol. 30, No. 2. P. 7–13.

[10] **Гнеденко Б.В.** Курс теории вероятностей. М.: Едиториал УРСС, 2005. 448 с.
Gnedenko, B.V. The Theory of Probability. Rhode Island: AMS Chelsea Publ., 2005. 529 p.

[11] **Феллер В.** Введение в теорию вероятностей. Т. 1. М.: Мир, 1984. 528 с.
Feller, W. An introduction to probability theory. Vol. 1. New York: John Wiley & Sons, 1967. 509 p.

[12] **Хусаинов А.А., Чернов А.М., Маевская Е.Д., Романченко А.А.** Модели для расчета времени работы вычислительных конвейеров // Матер. XXIII Междунар. науч.-практ. конф. "Актуальные проблемы науки". М.: Спутник+, 2016. С. 83–91.
Husainov, A.A., Chernov, A.M., Maevskaya, E.D., Romanchenko, A.A. Models for calculating the operating time of computational pipelines // Proc. of the Intern. Sci. and Pract. Conf. "Actual Problems of Science". Moscow: Sputnik+, 2016. P. 83–91.

Поступила в редакцию 10 мая 2017 г.,
с доработки — 11 октября 2017 г.

Optimal depth of the computational pipeline for a given amount of input data

Husainov, Ahmet A.*, Titova, Ekaterina A.

Komsomolsk-na-Amure State University, Komsomolsk-on-Amur, 681013, Russia

*Corresponding author: Husainov, Ahmet A., e-mail: husainov51@yandex.ru

A computational pipeline is considered, the stages of which can have different logic delays, and the overhead clock of the channels is the same. Given the probability that the data element calls a restart, an analytical model is constructed to calculate the most probable and expected times for processing input data elements for a given amount of input data.

Based on this model, it is shown that for a given number of stages and total logical delay, the pipeline has maximum performance if and only if the delays of the stages are equal to each other. For the pipeline with the same delays of stages, a correction is made for the well-known Duby — Flynn formula for the optimum depth of the computational pipeline with a given probability of restart, taking into account the number of elements being processed. The criterion of optimality can be the minimum most probable data processing time or the minimum value of the mathematical expectation the data processing time. An example is given showing that this correction can be significant.

It is shown that for given restart probability, total logical delay, channel exchange time and data volume, this formula gives the optimum depth on the set of all pipelines, including pipelines with different stage delays.

Keywords: computational pipeline, performance, total conflict, restart, most probable value, mathematical expectation, pipeline depth.

Received 10 May 2017
Received in revised form 11 October 2017