Федеральное агентство по образованию
Государственное образовательное учреждение
высшего профессионального образования
«Московский государственный технический университет имени
Н.Э.Баумана» (МГТУ им. Н.Э. Баумана)

В.В. Сюзев

Масштабирование данных и вычислительных алгоритмов в системах с фиксированной точкой

Учебное пособие

2008 г.

Оглавление

Предисловие

Преимущественное большинство управляющих ЭВМ в составе современных информационно-вычислительных системах реального времени используется в режиме работы с плавающей точкой, что упрощает программирование функциональных задач. Однако при решении задач с высокой инструментальной точностью необходим переход к представлению данных с фиксированной точкой, что создает дополнительные трудности при программировании, связанные с необходимостью сохранения диапазонов изменения физических величин и исключения возможных переполнений разрядных сеток при выполнении вычислительных операций. Обеспечивается это правильным выбором масштабов обрабатываемых величин и их преобразованием в процессе счета.

Настоящее пособие предназначено для оказания практической помощи студентам при изучении разделов, связанных с подготовкой машинных алгоритмов, входящих в состав учебной программы по дисциплине «Системы реального времени». Пособие в первую очередь предназначено для студентов специальности 230101 «Вычислительные машины, системы, комплексы и сети» и магистров по направлению 230100 «Информатика и вычислительная техника», однако может быть полезным и для студентов других специальностей, занимающихся проектированием систем реального времени и программированием задач для них.

1. Понятие о масштабировании и масштабах

Приведение всех исходных, промежуточных и конечных математических величин вычислительного алгоритма к диапазону изменения машинных переменных с учетом точности решения задач называется *масштабированием.*

При масштабировании математические величины алгоритма представляются соответствующими машинными переменными, над которыми затем в ЭВМ выполняются все действия, задаваемые формульной схемой алгоритма. Связь машинных переменных с математическими осуществляется с помощью масштабов. В ЭВМ, использующих дробную арифметику, в которой все числа по модулю меньше единицы, эта связь имеет вид

$$\bar{x} = m_x \cdot x, \qquad (1)$$

где x – некоторая математическая величина; $\bar{x}$ - ее машинное значение; m_x – масштаб x в дробной арифметике.

В случае применения целочисленной арифметики, где все числа целые, связь математической величины x с машинной $\bar{X}$ аналогична

$$\bar{X} = M_x x, \qquad (2)$$

здесь M_x – масштаб в целочисленной арифметике.

Масштабы определяются либо из предельных соотношений

$$m_x = \frac{\bar{x}_{\max}}{x_{\max}} = \frac{1-2^{-n}}{x_{max}}, \tag{3}$$

$$M_x = \frac{\bar{X}_{\max}}{x_{\max}} = \frac{2^n - 1}{x_{max}}, \tag{4}$$

и в этом случае их называют *предельными*, либо из условий

$$m_x^{\partial в} = \left\{ \frac{1-2^{-n}}{2^\kappa} < \frac{1-2^{-n}}{x_{\max}} < \frac{1-2^{-n}}{2^{\kappa-1}} \right\} \tag{5}$$

$$M_x^{\partial в} = \left\{ \frac{2^n - 1}{2^\kappa} < \frac{2^n - 1}{x_{\max}} < \frac{2^n - 1}{2^{\kappa-1}} \right\}. \tag{6}$$

Такой масштаб называется *двоичным*. Двоичные масштабы несколько ухудшают точность представления, так как используют не весь возможный диапазон, но в ряде случаев упрощают процесс масштабирования данных.

В формулах (3)-(6) n означает число разрядов в двоичном представлении машинных значений $\bar{x}$ и $\bar{X}$. При $n>>1$ эти формулы упрощаются:

$$m_x \cong 1/x_{\max} \tag{7}$$

$$M_x \cong 2^n / x_{\max}, \tag{8}$$

$$m_x^{\partial в} = \left\{ \frac{1}{2^\kappa} < \frac{1}{x_{\max}} < \frac{1}{2^{\kappa-1}} \right\}, \tag{9}$$

$$M_x^{\partial в} = \left\{ \frac{2^n}{2^\kappa} < \frac{2^n}{x_{\max}} < \frac{2^n}{2^{\kappa-1}} \right\}. \tag{10}$$

Масштабы имеют размерность, обратную размерности математической величины x. Машинные значения безразмерны.

Пример 1.1. Для величины x с максимальным значением x_{max} = 8000 м найти предельные и двоичные масштабы в дробной арифметике. Разрядность n = 8.

Решение.

$$m_x = \frac{1-2^{-n}}{x_{\max}} = \frac{1-2^{-8}}{8000} \approx 0{,}125 \cdot 10^{-3} \text{ 1/м},$$

$$m_x^{\partial в} = \left\{ \frac{1}{8192} < \frac{1}{8000} < \frac{1}{4096} \right\} = \frac{1}{8192} = 2^{-13} \text{ 1/м}.$$

Пример 1.2. Пусть x_{max} =240^0, n =12. Найти предельные и двоичные масштабы в целочисленной арифметике.

Решение.

$$M_x = \frac{2^n - 1}{x_{\max}} = \frac{2^{12} - 1}{240} \approx \frac{2^{12}}{240} = 17{,}056 \text{ 1/град},$$

$$M_x^{\partial в} = \left\{ \frac{2^{12}}{256} < \frac{2^{12}}{240} < \frac{2^{12}}{128} \right\} = \frac{2^{12}}{256} = 2^4 \text{ 1/град}.$$

Наряду с масштабами при масштабировании широко применяются обратные им величины:

$$\beta_x = 1/m_x, \tag{11}$$

$$B_x = 1/M_x, \tag{12}$$

которые носят название *цен*. Причем цена в дробной арифметике (β_x) имеет физический смысл *цены машинной единицы* (ЦМЕ) представления переменной x в машине, а цена в целочисленной арифметике (B_x) – *цены младшего разряда* (ЦМР) двоичного представления физической переменной. ЦМЕ и ЦМР имеют размерности, совпадающие с размерностями математических величин. Вычислить цены можно либо через масштабы по формулам (11) и (12), либо через предельные значения по формулам, легко получаемым из формул (3)-(10):

$$\beta_x = \frac{x_{\max}}{1-2^{-n}} \approx x_{\max} \tag{13}$$

$$\beta_x^{\partial в} = \left\{ \frac{2^{\kappa}}{1-2^{-m}} > \frac{x_{\max}}{1-2^{-n}} > \frac{2^{\kappa-1}}{1-2^{n}} \right\} \approx \left\{ 2^{\kappa} > x_{\max} > 2^{\kappa-1} \right\}, \tag{14}$$

$$B_x = \frac{x_{\max}}{2^n - 1} \approx \frac{x_{\max}}{2^n}, \tag{15}$$

$$B_x^{\partial в} = \left\{ \frac{2^{\kappa}}{2^n-1} > \frac{x_{\max}}{2^n-1} > \frac{2^{\kappa-1}}{2^n-1} \right\} \approx \left\{ \frac{2^{\kappa}}{2^n} > \frac{x_{\max}}{2^n} > \frac{2^{\kappa-1}}{2^n} \right\}. \tag{16}$$

Пример 1.3. Найти ЦМЕ для условий примера 1.1.
Решение. В соответствии с формулами (13) и (14)

$$\beta_x = \frac{8000}{1-2^{-8}} \approx 8000\,м,$$

$$\beta_x^{\partial в} = \left\{ \; 2^{13} > 8000 > 2^{12} \; \right\} = 2^{13} \;\; м.$$

Такой же ответ можно получить, если использовать результаты примера 1.1 и формулу (11):

$$\beta_x = \frac{1}{0{,}125 \cdot 10^{-3}} = 8000 \;\; м,$$

$$\beta_x^{\partial в} = \frac{1}{2^{-13}} = 2^{13} \;\; м.$$

Пример 1.4. Найти ЦМР для условий примера 1.2.
Решение. По формулам (15) и (16)

$$B_x = \frac{240}{2^{12}-1} \approx \frac{240}{4096} = 0{,}0596 \;\; град,$$

$$B_x^{\partial в} = \left\{ \; \frac{2^8}{2^{12}} > \frac{240}{2^{12}} > \frac{2^7}{2^{12}} \; \right\} = \frac{2^8}{2^{12}} = 2^{-4} \;\; град.$$

То же самое получаем, используя результаты примера 1.2,

$$B_x = \frac{1}{17{,}056} = 0{,}0596 \;\; град \;\; ,$$

$$B_x^{\partial в} = \frac{1}{2^4} = 2^{-4} \;\; град.$$

Взаимосвязь масштабов и цен в различных арифметиках можно выразить зависимостями, легко получаемыми из предыдущих формул:

$$m_x = \frac{1-2^{-n}}{2^n-1} M_x \approx \frac{1}{2^n} M_x; \quad M_x = \frac{2^{-n}-1}{1-2^n} m_x \approx 2^n m_x \tag{17}$$

$$\beta_x = \frac{2^n-1}{1-2^{-n}} B_x \approx 2^n B_x; \quad B_x = \frac{1-2^n}{2^{-n}-1} \beta_x \approx \frac{1}{2^n} \beta_x. \tag{18}$$

Эти зависимости позволяют, имея рассчитанные масштабы и цены в одной арифметике, легко перейти к масштабам и ценам в другой арифметике.

Пример 1.5. По результатам примеров 1.1 и 1.3 найти предельные масштабы и цены в целочисленной арифметике.

Решение. В соответствии с (17) и (18).

$$M_x = \frac{2^8-1}{1-2^{-8}} 0{,}125\cdot 10^{-3} \approx \frac{2^8}{8000} = 32\cdot 10^{-3} \quad 1/\text{м},$$

$$\beta_x = \frac{1-2^{-8}}{2^8-1} 8000 \approx \frac{8000}{2^8} = 31{,}25 \ \ м.$$

Наряду с расчетом масштабов и цен при масштабировании в ряде случаев, например, при хранении констант, возникает необходимость в определении машинного кода представления величин. Последовательность действий в этом случае такова: сначала по известным максимальным значениями находят масштабы или цены, а затем по формулам

$$\bar{x} = m_x x, \quad \overline{X} = M_x x, \qquad (19)$$

$$\bar{x} = \frac{x}{\beta x}, \quad \overline{X} = \frac{x}{B_x}$$

находят машинные значения в требуемой системе счисления.

Пример 1.6. Пусть x_{max} =20480 м, n=11. Записать число 3200 м в двоичном коде для машин с дробной и целочисленной арифметикой.

Решение. Для первого случая имеем

$$\beta_x = \frac{x_{\max}}{1-2^{-n}} = \frac{20480}{1-2^{-11}} \approx 20480 \ \text{м},$$

$$\bar{x} = \frac{x}{\beta_x} = \frac{3200}{20480} = \frac{10}{64} = 0{,}00101000000.$$

Во втором случае –

$$B_x = 2^{-n}\beta_x = \frac{20480}{2^{-11}} = 10 \ \text{м},$$

$$\bar{X} = \frac{x}{B_x} = \frac{3200}{10} = 320 = 00101000000.$$

Вопросы и задачи для самоподготовки

1. Что такое масштаб переменной? Каков физический смысл масштаба?
2. Что такое цена? Каков ее физический смысл?
3. Чем по физическому смыслу отличаются друг от друга цена машинной единицы и цена младшего разряда?
4. Чем отличаются предельные масштабы от двоичных?
5. Каковы достоинства и недостатки двоичных масштабов?
6. Как находятся двоичные коды констант для хранения в памяти?

Далее в задачах 7-16 определить предельные и двоичные масштабы и цены в дробной и целочисленной арифметике.

7. x_{max} = 60 м, n=3.
8. x_{max} = 10^0, n =4.
9. x_{max} =300 м/с, n =10.
10. x_{max}=360^0, n =9.
11. x_{max}=43^0C, n =10.
12. x_{max}=31 м/с^2, n =11.
13. x_{max}=15 с, n =10.
14. x_{max}=6 рад, n =12.
15. x_{max}=1700 об./с, n =8.
16. x_{max}=5600 кг, n =13.

Обратить внимание на то, что в задачах 7 и 8 число разрядов n мало и приближенными формулами в расчетах пользоваться нельзя.

В задачах 17-20 найти двоичное представление заданных констант при предельном масштабировании в обеих арифметиках.

17. x_{max} =7000м, n =10, x=3600 м.
18. x_{max} =400 м/с n =12, x_1 =300 м/с, x_2=120 м/с/
19. x_{max} =350^0, n =11, x_1 =270^0, x_2=60^0.
20. x_{max} =25 с, n =10, x =15 с.

2. *Масштабирование отдельных операций*

1. *Операция сдвига.* Пусть некоторая машинная величина $\bar{x}$ сдвигается на m разрядов вправо или влево. Такой сдвиг может рассматривать двояко. С одной стороны его можно интерпретировать как увеличение либо уменьшение в 2^m раз физической величины x (т.е. умножение x на 2^m при m>0 и на $2^{-/m/}$ при m<0). Масштаб (цена) результата z такой операции $z=2^m x$, очевидно, будет совпадать с масштабом (ценой) операнда x, т.е. $m_z = m_x$, $\beta_z = \beta_x$. Аналогично в целочисленной арифметике $M_z = M_x$, $B_z = B_x$. Машинные операции в этом случае таковы: $\bar{Z} = 2^m \bar{x}$; $\bar{Z} = 2^m \bar{X}$.

С другой стороны, можно считать, что физическая величина x при сдвиге не меняется, т.е. $z = x$. В этом случае $m_z = m_x 2^m$, $\beta_z = \beta_x 2^{-m}$. В целочисленной арифметике $M_z = M_x 2^m$, $B_z = B_x 2^{-m}$. Машинные операции имеют прежний вид.

Пример 2.1. Определить ЦМЕ и ЦМР $z = 2^3 x$ при n=14. $x_{\max}$ =8000 м. Проверить результаты при x =300 м.

Решение.

$$\beta_x = \frac{x_{\max}}{1-2^{-n}} \approx x_{\max} = 8000 \text{ м}; \quad B_x = \frac{x_{\max}}{2^n - 1} \approx x_{\max} 2^{-n} = 8000 \cdot 2^{-14} \text{ м},$$

$$\beta_z = \beta_x \cdot 2^{-3} = 1000 \text{ м}; \qquad B_z = B_x \cdot 2^{-3} = 1000 \cdot 2^{-14}.$$

Проверка:

$$\bar{x} = \frac{x}{\beta_x} = \frac{300}{8000} = \frac{3}{80}; \quad \bar{X} = \frac{x}{B_x} = \frac{3}{80} = \frac{3}{80} \cdot 2^{14};$$

$$\bar{z} = 2^3 \bar{x} = \frac{3 \cdot 8}{80} = 0,3; \qquad \bar{Z} = \bar{X} \cdot 2^3 = \frac{3 \cdot 8}{80} 2^{14} = 0,3 \cdot 2^{14};$$

$$z = \beta_z \cdot \bar{z} = 1000 \cdot 0,3 = 300 \text{ м}; \quad z = B_z \bar{z} = 1000 \cdot 2^{-14} \cdot 0,3 \cdot 2^{14} = 300 \text{ м}.$$

2. *Операция умножения на константу.* Умножение некоторой величины x на константу можно рассматривать с трех точек зрения. Во-первых, как перемножение двух величин, значение одной из которых неизменно. Это частный случай обычной операции умножения, масштабирование которой будет рассмотрено далее отдельно.

Во-вторых, умножение на константу можно интерпретировать как умножение машинного значения $\bar{x}$ на некоторый машинный множитель $к$ без изменения значения физической величины x . Машинная операция при этом будет иметь вид $\bar{z} = к\bar{x}$. Так как $\bar{x}$ возросло в $к$ раз, а x осталось при этом неизменным, то учитывая, что в дробной арифметике $x = \bar{x} / m_x$ $(x = \bar{x}\beta_x)$ необходимо масштаб произведения увеличить (цену уменьшить) в $к$ раз, т.е. $m_z = к m_x$, $\beta_z = \beta_x / к$. В целочисленной арифметике приходим к аналогичному результату, правда при этом следует еще учесть тот факт, что при перемножении двух n – разрядных машинных чисел $к$ и $\bar{x}$ результат в общем случае занимает $2n$-разрядную сетку. Если результат умножения затем усекается (округлятся) до n разрядов, то в целочисленной арифметике такое усечение приводит к уменьшению модуля машинного числа в 2^n раз. Для того, чтобы физическая величина произведения при этом не изменилась, масштаб результата должен быть уменьшен (цена увеличена) в 2^n раз. Поэтому в целочисленной арифметике $M_z = к M_x \cdot 2^{-n}$, $B_z = B_x 2^n / к$. Машинная операция в этом случае имеет вид: $\bar{Z} = к\bar{X} 2^{-n}$.

Формула $\bar{Z} = \bar{X} к 2^{-n}$ есть условная запись машинной операции умножения в целочисленной арифметике, в то время как реальная операция - $\bar{Z} = \bar{X}\bar{K}$. Однако при выполнении масштабирования и проверки его на бумаге использовать нужно именно условную запись, в противном случае результаты будут неверны.

В дробной арифметике ограничение разрядной сетки при умножении не влияет на расчет масштабов и цен.

Пример 2.2. Определить масштаб и цену $z = 0,3x$ при $x_{\max} = 3600$ м, n = 12. Арифметика дробная.

Решение.

$$m_x = \frac{1 - 2^{-n}}{x_{\max}} \approx \frac{1}{3600} 1/\text{м}; \quad \beta_x \approx 3600 \text{ м};$$

$$m_z = к m = 0,3 \frac{1}{3600} = \frac{1}{12000} \frac{1}{\text{м}}; \quad \beta_z = \frac{\beta_x}{к} = \frac{3600}{0,3} = 12000 \text{м}.$$

Проверка. Проверку выполним для x = 1500 м.

$$\bar{x} = \frac{x}{\beta_x} = \frac{1500}{3600} = \frac{15}{36}; \quad \bar{z} = к\bar{x} = 0,3 \frac{15}{36} = \frac{1}{8}; \quad \bar{z} = к\bar{x}$$

$$z = \beta_z \cdot \bar{z} = 12000 \cdot \frac{1}{8} = 1500 \text{ м} \equiv x.$$

Пример. 2.3. Определить масштаб и цену z=300x при $B_x = 3$м, n = 13. Арифметика целочисленная. Проверку сделать для x = 3000м.

Решение.

$$M_x = \frac{1}{B_x} = \frac{1}{3}\ \frac{1}{\text{м}};\ M_z = \kappa M_x 2^{-n} = 100 \cdot 2^{-13}\ \frac{1}{\text{м}};$$

$$B_z = B_x \cdot 2^n / \kappa = 3 \cdot 2^{13} / 300 = 0{,}01 \cdot 2^{13}\ \text{м}.$$

Проверка:

$$\overline{X} = \frac{x}{B_x} = \frac{3000}{3} = 1000;\ \overline{Z} = \kappa \overline{X} \cdot 2^{-n} = 3 \cdot 10^5 \cdot 2^{-13};$$

$$z = \overline{Z} B_z = 3 \cdot 10^5 \cdot 2^{-13} \cdot 0{,}01 \cdot 2^{13} = 3000\ \text{м} \equiv x.$$

В-третьих, умножение на константу можно вообще свести только к изменению масштабов (цен) без изменения машинной величины $\bar{x}$ (т.е. без выполнения самой операции умножения). Константу можно учесть следующим образом: так как $\bar{x}\beta_x = x$, а $\bar{x}\beta_z = x\kappa$, то $\beta_z = \kappa\beta_x$. Соответственно $m_z = m_x / \kappa$. Аналогично в целочисленной арифметике $B_z = \kappa B_x$, $M_z = M_x / K$. Машинные операции имеют простой вид: $\overline{Z} = \overline{X}$ и $\bar{z} = \bar{x}$.

Пример 2.4. Определить ЦМЕ $z = 0{,}12x$ при $\beta_x = 8000$ м. Проверку сделать для x=6000 м.

Решение. $\beta_z = \kappa \cdot \beta_x = 0{,}12 \cdot 8000 = 960$ м.

Проверка: $\bar{x} = \frac{x}{\beta_x} = \frac{6000}{8000} = \frac{3}{4};\ \bar{z} = \frac{z}{\beta_z} = \frac{0{,}12 \cdot 6000}{960} = \frac{3}{4}.$

3. *Операция сложения.* Пусть $z = x + y$, причем известны β_x и β_y (в общем случае $\beta_x \neq \beta_y$), а также максимальное значение суммы $z_{\max}$. Найдем цену результата операции и те дополнительные операции над машинными числами, которые необходимо осуществить при сложении.

Имеем $z = x + y = \bar{z}\beta_z = \bar{x}\beta_x + \bar{y}\beta_y$. Так как складывать можно только машинные числа с одним масштабом (одной ценой), необходимо преобразовать цены чисел x и y. В качестве величины цены, относительно которой будем выравнивать все цены, выберем наибольшую величину из β_x, β_y и $\beta_z^{пр} = \frac{z_{\max}}{1 - 2^{-n}}$. Очевидно цена суммы будет равна этой величине. Математически это можно сформулировать так:

$$\beta_z = \max\left\{\beta_z^{пр};\ \beta_x;\ \beta_y\right\}. \qquad (20)$$

Для преобразования цен x и y воспользуемся операцией умножения на машинную константу (см. второй случай предыдущей операции), т.е. умножим машинные значения $\bar{x}$ и $\bar{y}$ на соответствующие масштабные множители κ_x и κ_y. Поскольку в результате умножения должны получиться произведения $\bar{x}\kappa_x$ и $\bar{y}\kappa_y$ с одной и той же ценой β_z, причем $\beta_z = \beta_x / \kappa_x$ и $\beta_z = \beta_y / \kappa_y$, то

$$\kappa_x = \frac{\beta_x}{\beta_z};\ \kappa_y = \frac{\beta_y}{\beta_z}. \qquad (21)$$

Таким образом, в общем случае одна машинная операция сложения будет складываться из двух операций умножения $\bar{x}_1 = \bar{x}\kappa_x$, $y_1 = \bar{y}\kappa_y$ и одной операции суммирования $\bar{z} = \bar{x}_1 + \bar{y}_1$. В частном случае, если $\beta_z = \beta_x$ или $\beta_z = \beta_y$, один из коэффициентов равен 1 и умножение на него выполнять не надо. В самом простом случае, когда $\beta_x = \beta_y$ цены равны), цена β_z выбирается равной цене слагаемых, и в машинном выполнении операция сложения выполняется без предварительного умножения на масштабные множители (правда при этом следует иметь в виду возможность переполнения разрядной сетки).

При суммировании чисел в целочисленной арифметике аналогично

$$B_z = \max\left\{B_z^{mp};\ B_x;\ B_y\right\},\ B_z^{mp} = \frac{z\max}{2^n - 1}, \tag{22}$$

и

$$K_x = \frac{B_x}{B_z}2^n;\ \ K_y = \frac{B_y}{B_z}2^n. \tag{23}$$

Запись совокупности машинных операций здесь следующая:

$$\bar{X}_1 = \bar{X}K_x 2^{-n};\ \bar{Y}_1 = \bar{Y}K_{yz}2^n;\ \bar{Z} = \bar{X}_1 + \bar{Y}_1.$$

Приведенные выкладки выполнены относительно цен, однако подобные зависимости легко можно получить и для масштабов, учитывая их простую связь с ценами.

Пример 2.5. Выполнить масштабирование операции сложения $z = x + y$ при $x_{\max} = 10000$ м, $y_{\max} = 20000$ м, n=10. Арифметика дробная. Проверку сделать для чисел x=300 м, y=420 м.

Решение. $\beta_x \approx 10000$ м; $\beta_y \approx 20000$ м и $\beta_x \neq \beta_y$.

Тогда
$$\beta_z^{mp} = \frac{z_{\max}}{1 - 2^{-n}} \approx z_{\max} = x_{\max} + y_{\max} = 30000 \text{ м}.$$

и
$$\beta_z = \max\{30000;\ 10000;\ 20000\} = 30000 \text{ м}$$

$$\kappa_x = \frac{\beta_x}{\beta_z} = \frac{10000}{30000} = \frac{1}{3};\ \kappa_y = \frac{\beta_y}{\beta_z} = \frac{20000}{30000} = \frac{2}{3}.$$

Проверка:

$$\bar{x}_1 = \bar{x}\kappa_x = \frac{x}{\beta_x}\kappa_x = \frac{300}{10000}\cdot\frac{1}{3} = 0,01;$$

$$\bar{y}_1 = \bar{y}\kappa_y = \frac{y}{\beta_y}\kappa_y = \frac{420}{20000}\cdot\frac{2}{3} = 0,014;$$

$$\bar{z} = \bar{x}_1 + \bar{y}_1 = 0,01 + 0,014 = 0,024$$

$$\bar{z} = \bar{z}\beta_z = 0,004 \cdot 30000 = 720 \text{ м}.$$

Действительно, 300 м + 420 м = 720 м.

Пример2.6. Условия примера 2.5, только арифметика целочисленная.

Решение.

$$B_x \approx 10^4 \cdot 2^{-10} \text{ м},\ B_y \approx 2 \cdot 10^4 \cdot 2^{-10} \text{ м},\ B_z^{mp} = 3 \cdot 10^4 \cdot 2^{-10} \text{ м}$$

$$B_z = \max\{3 \cdot 10^4 \cdot 2^{-10};\ 10^4 \cdot 2^{-10}\} = 3 \cdot 10^4 \cdot 2^{-10} \text{ м};$$

$$K_x = \frac{B_x}{B_z}2^n = \frac{10^4 \cdot 2^{-10}}{3 \cdot 10^4 \cdot 2^{-10}} \cdot 2^{10} = \frac{1}{3}2^{10};\ K_y = \frac{B_y}{B_z}2^n = \frac{2 \cdot 10^4 \cdot 2^{-10}}{3 \cdot 10^4 \cdot 2^{-10}}2^{10} = \frac{2}{3}2^{10}.$$

Проверка:

$$\overline{X}_1 = \overline{X}K_x 2^{-n} = \frac{x}{B_x}K_x 2^{-10} = \frac{300}{10^4 \cdot 2^{-10}} \cdot \frac{1}{3}2^{10} \cdot 2^{-10} = 0,01 \cdot 2^{10} = 0,01 \cdot 2^{10};$$

$$\overline{Y}_1 = \overline{Y}K_y 2^{-n} = \frac{y}{B_y}K_y 2^{-10} = \frac{420}{2 \cdot 10^4 \cdot 2^{-10}} \cdot \frac{2}{3}2^{10} \cdot 2^{-10} = 0,014 \cdot 2^{10};$$

$$\overline{Z} = \overline{X}_1 + \overline{Y}_1 = 0,01 \cdot 2^{10} + 0,014 \cdot 2^{10} = 0,024 \cdot 2^{10};\ z = \overline{Z}B_z = 0,024 \cdot 2^{10} \cdot 3 \cdot 10^4 2^{-10} = 720 \text{ м}.$$

Очевидно, что все полученные результаты для операции сложения справедливы и для операции вычитания.

4. *Операция умножения.* Если $z = xy$, а β и β_y известны, то в дробной арифметике $z = \bar{z}\beta_z = \bar{x}\beta_x \cdot \bar{y}\beta_y$. Так как машинная операция умножения в этом случае имеет вид $\bar{z} = \bar{x}\bar{y}$, то нетрудно получить формулу для расчета цены произведения

$$\beta_z = \beta_x \cdot \beta_y. \qquad (24)$$

В целочисленной арифметике с учетом усечения произведения до n разрядов

$$B_z = B_x \cdot B_y 2^n, \qquad (25)$$

а машинная операция записывается так: $\overline{Z} = \overline{X} \cdot \overline{Y} \cdot 2^{-n}$.

Пример 2.7. Выполнить масштабирование $z = xy$ при $x_{\max} = 50$ м/с, $y_{\max} = 12$ с, n=10. Арифметика дробная. Проверку выполнить для x=25/с, y=10 с.

Решение. $\beta_x \approx 50$ м/с; $\beta_y \approx 12$ с; $\beta_z = \beta_x\beta_y$=50·12=600 м.

Проверка:

$$\bar{x} = \frac{x}{\beta_x} = \frac{25}{50} = 0,5;\quad \bar{y} = \frac{y}{\beta_y} = \frac{10}{12} = \frac{5}{6};$$

$$\bar{z} = \bar{x}\bar{y} = \frac{1}{2} \cdot \frac{5}{6} = \frac{5}{12};\quad z = \bar{z}\beta_z = \frac{5}{12} \cdot 600 = 250 \text{ м}.$$

Действительно, 25 м/с·10 с = 250 м.

Пример 2.8. Условие примера 2.7 только в целочисленной арифметике.

Решение. $B_x \approx 50 \cdot 2^{-10}$ м/с, $B_y \approx 12 \cdot 2^{-10}$ с.

$$B_z = B_x \cdot B_y \cdot 2^n = 0 \cdot 2^{-10} \cdot 12 \cdot 2^{-10} \cdot 2^{10} = 600 \cdot 2^{-10} \text{м}.$$

Проверка:

$$\overline{X} = \frac{x}{B_x} = \frac{25}{50 \cdot 2^{-10}} = 2^9;\quad \overline{Y} = \frac{y}{B_y} = \frac{10}{12 \cdot 2^{-10}} = \frac{5}{6}2^{10};$$

$$\overline{Z} = \overline{X} \cdot \overline{Y} \cdot 2^{-n} = 2^9 \cdot \frac{5}{6} \cdot 2^{10} \cdot 2^{-10} = \frac{5}{6} \cdot 2^9;\ z = \overline{Z}B_z = \frac{5}{6}2^9 \cdot 600 \cdot 2^{-10} = 250 \text{ м}.$$

5. *Операция деления.* Эта операция вызывает особые трудности при масштабировании, поскольку результат ее может выйти за пределы изменения делимого и делителя. Поэтому рассмотрим е более подробно, отдельно для дробной и целочисленной арифметики.

а) *Дробная арифметика.* Пусть делимое – x с ценой β_x, а делитель – y с ценой β_y. Поскольку конструкция ЭВМ допускает деление только таких $\bar{x}$ и $\bar{y}$, при которых частное $\bar{z}$ меньше единицы, т.е. $\bar{z}=\bar{x}/\bar{y}<1$, то необходимым условием правильного деления является $|x|_{\max}/\beta_x < |y|_{\min}/\beta_y$, или

$$\beta_x \geq \beta_x^* = \left|\frac{x_{\max}}{y_{\min}}\right|\beta_y = |z|_{\max}\beta_y, \tag{26}$$

где β_x^* - ограниченное снизу β_x, приемлемое для выполнения операции без переполнения разрядной сетки.

Другими словами, если условие правильного деления не выполняется, то перед операцией деления цена делимого β_x должна быть приведена к ограниченному снизу значению β_x^*, только после этого можно приступить к выполнению самой операции деления. Преобразование цены выполняется путем умножения $\bar{x}$ на машинную константу $к$, величина которой определяется по формуле

$$к = \frac{\beta_x}{\beta_x^*}. \tag{27}$$

Если β_x является предельным масштабом, т.е $\beta_x = \beta_x^{пр}$, то формулу (27) можно записать еще и так:

$$к = \frac{\beta_x^{пр}}{\beta_x^*} = \left|\frac{y_{\min}}{y_{\max}}\right|\left|z_{\max}\right|. \tag{28}$$

Машинная операция деления будет иметь вид

$$\bar{z} = \frac{\bar{x}к}{\bar{y}}. \tag{29}$$

Найдем цену частного z. Очевидно, $z=\bar{z}\beta_z$. Умножим числитель и знаменатель последнего выражения на $к$: $\bar{z}\beta_z = \frac{\bar{x}к\beta_x}{\bar{y}\beta_н к}$ и сравним с (29). Тогда

$$\beta_z = \frac{\beta_x}{\beta_y к} = \frac{\beta_x^*}{\beta_y}. \tag{30}$$

В том случае, если условие деления выполняется, $к$=1 и

$$\beta_z = \frac{\beta_x}{\beta_y};\quad \bar{z} = \frac{\bar{x}}{\bar{y}} \tag{31}$$

Пример 2.9. Выполнить масштабирование $z=x/y$ при $x_{\max}=200$м, $y_{\max}=50c$, $y_{\min}=10c$, $n=12$. Проверку сделать для x=80 м, y=20 с.

Решение. Так как $n=12 >> 1$, $\beta_x \approx 200$ м; $\beta_y \approx 50c$.

$$\beta_x^* = \frac{x_{\max}}{y_{\min}}\beta_y = \frac{200\cdot 50}{10} = 10^3 \text{м} > \beta_x,$$

т.е. условие деления (~~1.13~~ 26) не выполняется. Тогда

$$\beta_z = \frac{\beta_x^*}{\beta_y} = \frac{1000}{50} = 20 \text{ м/с};\quad к = \frac{\beta_x}{\beta_x^*} = \frac{200}{1000} = \frac{1}{5}.$$

Машинная операция

$$\bar{z} = \frac{\bar{x}\kappa}{\bar{y}}.$$

Проверка:

$$\bar{x} = \frac{x}{\beta_x} = \frac{80}{200} = 0,4;\ \bar{y} = \frac{y}{\beta_y} = \frac{20}{50} = 0,4;$$

$$\bar{z} = \frac{0,4 \cdot 1}{0,4 \cdot 5} = 0,2;\ z = \beta_z \cdot \bar{z} = 20 \cdot 0,2 = 4 \text{ м/с}.$$

Действительно, 80 м/20 с=4 м/с.

б) *Целочисленная арифметика.* В этом случае машина оперирует с целыми числами $\bar{X}$ и $\bar{Y}$, для которых известны ЦМР B_x и B_y. Воспользуемся результатами дробной арифметики, учитывая, что $\bar{x} = 2^{-n}\bar{X}$, $\bar{y} = 2^{-n}\bar{Y}$, $\beta_x = 2^n B_x$ и $\beta_y = 2^n B_y$. Тогда в соответствии с (1.26) получаем $2^n B_x \geq 2^n B_x^* = \left|\frac{x_{\max}}{y_{\min}}\right| 2^n B_y$, из чего следует такое условие правильного выполнения операции деления и целочисленной арифметике:

$$B_x > B_x^* = \left|\frac{x_{\max}}{y_{\min}}\right| B_y = |z_{\max}| B_y. \tag{32}$$

Если условие не выполняется, то изменение цены B_x осуществляется умножением $\bar{X}$ на масштабный коэффициент

$$K = \frac{B_x}{B_x^*} 2^n. \tag{33}$$

Новое машинное число $\bar{X}^*$ будет равно $\bar{X}^* = \bar{X} \cdot K \cdot 2^{-n}$.

В целочисленной арифметике результат деления целых чисел тоже должен быть целым, поэтому он получается как бы увеличенным в 2^n раз. С учетом этого запись машинной операции деления здесь приобретает вид

$$\bar{Z} = \frac{\bar{X}^*}{\bar{Y}} 2^n = \frac{\bar{X} \cdot K \cdot 2^{-n}}{\bar{Y}} 2^n = \frac{\bar{X} \cdot K}{\bar{Y}}, \tag{34}$$

а ЦМР частного равна

$$B_z = \frac{B_x^*}{B_y} 2^{-n}. \tag{35}$$

Пример 2.10. Условие примера 2.9. Арифметика целочисленная.

Решение. $B_x = 200 \cdot 2^{-12}$ м; $B_y = 50 \cdot 2^{-12}$ с.

$$B_x^* = \left|\frac{x_{\max}}{y_{\min}}\right| B_y = \frac{200}{10} 50 \cdot 2^{-12} = 1000 \cdot 2^{-12} > B_x.$$

Условие правильного деления не выполняется, поэтому

$$B_z = \frac{B_x^*}{B_y} 2^{-n} = \frac{1000 \cdot 2^{-12}}{50 \cdot 2^{-12}} \cdot 2^{-12} = 20 \cdot 2^{-12} \text{ м/с};\ K = \frac{B_x}{B_x^*} = \frac{200 \cdot 2^{-12}}{1000 \cdot 2^{-12}} \cdot 2^{12} = \frac{1}{5} \cdot 2^{12}.$$

Проверка:

$$\bar{X} = \frac{x}{B_x} = \frac{80}{200 \cdot 2^{-12}} = 0,4 \cdot 2^{12};\qquad \bar{Y} = \frac{y}{B_y} = \frac{20}{50 \cdot 2^{-12}} = 0,4 \cdot 2^{12};$$

$$\overline{Z} = \frac{\overline{X} \cdot K \cdot 2^{-n}}{\overline{Y}} \cdot 2^{n} = \frac{\overline{X} \cdot K}{\overline{Y}} = \frac{0,4 \cdot 2^{12} \cdot 2^{12}}{0,4 \cdot 2^{12} \cdot 5} = 0,2 \cdot 2^{12};$$

$$z = \overline{Z} B_z = 0,2 \cdot 2^{12} \cdot 20 \cdot 2^{-12} = 4 \text{ м/с}.$$

Вопросы и задачи для самопроверки

1. Как тремя способами можно выполнить масштабирование операции умножения на константу?
2. В чем суть приема учета умножения на константу без выполнения самой операции умножения?
3. В чем особенности масштабирования операции сдвига?
4. Какие особенности имеет запись машинной операции умножения в целочисленной арифметике? В чем их причина?
5. В каких случаях при реализации операции сложения (вычитания) необходимо выполнять умножения на масштабные коэффициенты? В каких случаях операция сложения выполняется без них?
6. Что произойдет, если выполнить операцию деления машинных чисел без умножения делимого на масштабный коэффициент при невыполнении условия правильного деления?

В задачах 7-21 выполнить масштабирование заданных операций. Результаты проверить на указанных конкретных числах.

7. $z = 2^{-2}x$, $x_{max} = 300$ м/с, n=10, x=200 м/с.

   Арифметика дробная.

8. $z = 2^{-4}x$ (без изменения x), x_{max} =250 град., n=12, x=180 град.

   Арифметика целочисленная.

9. z = 0,8x (без изменения x), x_{max} =30 с, n = 11, x = 18 с.

   Арифметика дробная.

10. z = 250 x (без изменения x), x_{max} = 6 рад., n = 10, x = 4 рад.

    Арифметика целочисленная.

11. z = 1, 65 x (без умножения) , x_{max} = 3000 м^2, n = 15, x = 1200 м^2.

    Арифметика дробная.

12. z = 6, 25 x (без умножения), x_{max} =40 м/с^2, n = 12, x = 15 м/с^2.

    Арифметика целочисленная.

13. $z = x + y$, x_{max} = 9000 г, n = 10, x = 3000 г, y = 4500 г.

    Арифметика дробная.

14. $z = x - y$, x_{max} = 1500 м, y_{max} = 2000 м, n = 12, x = 1000 м, y = 300 м.

    Арифметика целочисленная.

15. $z = x + y$, $x_{max} = 30^0$, $y_{max} = 30^0$, $n = 10$, $x = 12^0 C$, $y = 8^0 C$.

    Арифметика целочисленная.

16. $z = x - y$, x_{max} = 600 м/с, y_{max} = 600 м/с, n=11, x =2 00 м/с, y = 250 м/с.

    Арифметика дробная.

17. $z = x + y - p$, $x_{max} = 360$ град., $y_{max} = 350$ град., $p_{max} = 355$ град., $n = 10$, $x = 100$ град., $y = 50$ град., $p = 200$ град.

Арифметика целочисленная.

18. $z = xy$, $x_{max} = 800$ м/с, $y_{max} = 7{,}5$ с, $n = 11$, $x = 200$ м/с, $y = 5$ с.

Арифметика дробная.

19. $z = xy$, $x_{max} = 60$ м/с^2, $y_{max} = 10$ с, $n = 12$, $x = 40$ м/с^2, $y = 8$ с.

Арифметика целочисленная.

20. $z = \frac{x}{y}$, $x_{max} = 800$ м, $y_{max} = 15$ с, $y_{min} = 2$ с, $n = 10$, $x = 300$ м, $y = 6$ с.

Арифметика дробная.

21. $z = \frac{x}{y}$, $x_{max} = 10000$ м, $y_{max} = 10000$ м, $y_{min} = 100$ м, $n = 12$, $x = 500$ м, $y = 1000$ м.

Арифметика целочисленная.

3.. Масштабирование вычислительного алгоритма

Если в процессе масштабирования масштабы одних и тех же физических величин остаются неизменными и подвергаются преобразованиям только для согласования масштабов (например, при выполнении операции сложения), то такой способ масштабирования носит название *постоянного масштабирования* (масштабирования с постоянными масштабами). Если же на различных этапах вычислительного процесса масштаб одних и тех же физических величин выбирается различным, то говорят о *переменном масштабировании* (масштабировании с переменными масштабами). Переменное масштабирование в ряде случаев позволяет повысить точность вычислений, однако требует дополнительных затрат машинного времени при реализации алгоритма на ЭВМ. Примеры переменного масштабирования приведены ниже, в 4. Здесь же рассмотрим только примеры постоянного масштабирования задач в управляющей ЭВМ.

Пример 3.1. Выполнить масштабирование выражения $z = \frac{a^3 + b(c^2 + 6d^2)}{0{,}2fp}$ при $n = 12$ и следующих пределах изменения переменных в м: $a_{max} = 3000$, $b_{max} = 2000$, $c_{max} = 2500$, $d_{max} = 2800$, $f_{max} = 3000$, $p_{max} = 2500$, $f_{min} = 30$, $p_{min} = 25$. Арифметика дробная. Проверить правильность расчетов для значений $a = 300$ м, b=600 м, c =750 м, d=560 м, f=900 м, p =1500 м.

Решение. Вычислим цены всех переменных. Так как $n >> 1$, то β_a=300 м, β_b=2000 м, β_c = 2500 м, β_d=2800 м, β_f=3000 м, β_p=2500 м. Далее последовательность действий такова:

1) возведение c в степень 2: $z_1 = c^2$;

2) возведение d в степень 2: $z_2 = d^2$;

$$\beta z_2 = \beta^2 d = 7{,}84 \cdot 10^6 \text{ м}^2; \quad \overline{z}_2 = \overline{dd};$$

3) умножение на константу 6: $z_3 = z_2 \cdot 6$ (изменение масштаба);

$$\beta_{z_3} = 6\beta z_2 = 6 \cdot 7,84 \cdot 10^6 = 47,04 \cdot 10^6 \text{ м}^2;\ \bar{z}_3 = \bar{z}_2;$$

4) суммирование $z_4 = z_1 + z_3$;

так как $\beta z_1 \neq \beta z_3$, то $\bar{z}_4 = \kappa_1 \bar{z}_1 + \kappa_3 \bar{z}_3$;

$$\beta z_4^{пр} = z_{4\max} = c_{\max}^2 + 6d_{\max}^2 = 6,25 \cdot 10^6 + 47,04 \cdot 10^6 = 53,29 \cdot 10^6 \text{ м}^2;$$

$$\beta z_4 = max\{\beta z_4^{пр}; \beta z_1; \beta z_3\} = 53,29 \cdot 10^6 \text{ м}^2;$$

$$\kappa_1 = \frac{\beta z_1}{\beta z_4} = \frac{6,25 \cdot 10^6}{53,29 \cdot 10^6} = 0,117; \qquad \kappa_3 = \frac{\beta z_3}{\beta z_4} = \frac{47,04 \cdot 10^6}{53,29 \cdot 10^6} = 0,083;$$

5) вычисление произведения: $z_5 = z_4 \cdot b$;

$$\beta z_5 = \beta z_4 \cdot \beta_b = 53,29 \cdot 10^6 \cdot 2 \cdot 10^3 = 106,58 \cdot 10^9 \text{ м}^3;\ \bar{z}_5 = \bar{Z}_4 \cdot \bar{b};$$

6) возведение в степень 2: $z_6 = a^2$;

$$\beta_{z_6} = \beta_{a^2} = 9 \cdot 10^6 \text{ м}^2;\ \bar{z}_6 = \bar{a} \cdot \bar{a};$$

7) возведение в степень 3: $z_7 = z_6 \cdot a$;

$$\beta z_7 = \beta z_6 \cdot \beta_a = 27 \cdot 10^9 \text{ м}^3;\ \bar{z}_7 = \bar{z}_6 \cdot \bar{a};$$

8) суммирование $z_8 = z_7 + z_5$; $\beta z_7 \neq \beta z_5$, поэтому $\bar{z}_8 = \kappa_7 \bar{z}_7 + \kappa_5 \bar{z}_5$;

$$\beta z_8^{пр} = z_{8\max} = a_{\max}^3 + в_{\max} \cdot z_{4\max} = 27 \cdot 10^9 + 2 \cdot 10^3 \cdot 53,29 \cdot 10^6 = 133,58 \cdot 10^9 \text{ м}^3;$$

$$\beta z_8 = \max\{\beta z_8^{пр};\ \beta z_7;\ \beta z_5\} = 133,58 \cdot 10^9 \text{ м}^3;$$

$$\kappa_5 = \frac{\beta z_5}{\beta z_8} = \frac{106,58 \cdot 10^9}{133,58 \cdot 10^9} = 0,798; \qquad \kappa_7 = \frac{\beta z_7}{\beta z_8} = \frac{27 \cdot 10^9}{133,58 \cdot 10^9} = 0,202;$$

9) вычисление произведения: $z_9 = fp$;

$$\beta z_9 = \beta_f \cdot \beta_p = 3 \cdot 10^3 \cdot 2,5 \cdot 10^3 = 7,5 \cdot 10^6 \text{ м}^3;\ \bar{z}_9 = \bar{f} \cdot \bar{p};$$

10) умножение на константу 0,2: $z_{10} = 0,2 z_9$ (изменение масштаба);

$$\beta z_{10} = 0,2 \beta z_9 = 0,2 \cdot 7,5 \cdot 10^6 = 1,5 \cdot 10^6 \text{ м}^2;\ \bar{z}_{10} = \overline{z_9};$$

11) деление $z_{11} = \frac{z_8}{z_{10}}$;

$$\beta_{z_8}^* = \frac{|z_{8max}|}{|z_{10\min}|} \beta z_{10} = \frac{a_{\max}^3 + в_{\max}(c_{max}^2 + 6d_{\max}^{12})}{0,2 f_{\min} p_{\min}} = 133,58 \cdot 10^{13} \text{ м}^3;$$

$\beta z_8 < \beta z_8^*$, условие правильного деления не выполняется, поэтому

$$\beta z_{11} = \frac{\beta z_8^*}{\beta z_{10}} = \frac{133,58 \cdot 10^{13}}{1,5 \cdot 10^6} = 8,9 \cdot 10^8 \text{ м};$$

$$\bar{z}_{11} = \frac{\bar{z}_8 \cdot \kappa_8}{\bar{Z}_{10}};\ \kappa_8 = \frac{\beta z_8}{\beta z_8^*} = \frac{133,58 \cdot 10^9}{133,58 \cdot 10^{13}} = 10^{-4}$$

Таким образом, $\beta z = \beta z_{11} = 8,9 \cdot 10^8$ м, а расчетная формула с учетом масштабных множителей примет вид:

$$\bar{z} = \frac{[\kappa_7 \cdot \bar{a}^3 + \kappa_5 \bar{b}(\bar{c}^2 \kappa_1 + \bar{d}^2 \kappa_3)] \cdot \kappa_8}{\bar{f} \cdot \bar{p}}$$

Проверка: $\bar{a} = \frac{a}{\beta_a} = \frac{300}{3000} = 0,1; \qquad \bar{в} = \frac{b}{\beta_b} = \frac{600}{2000} = 0,3;$

$$\bar{c} = \frac{c}{\beta_c} = \frac{750}{2500} = 0,3; \qquad \bar{d} = \frac{d}{\beta_d} = \frac{560}{2800} = 0,2;$$

$$\bar{f} = \frac{f}{\beta_f} = \frac{900}{3000} = 0,3; \qquad \bar{p} = \frac{p}{\beta_p} = \frac{1500}{2500} = 0,6;$$

$$\bar{z} = \frac{[\kappa_7 \cdot \bar{a}^3 + \kappa_5 \bar{в}(\kappa_1 \bar{c}^2 + \kappa_3 \bar{d}^2)] \cdot \kappa_8}{\bar{f} \cdot \bar{p}} =$$

$$= \frac{10^{-4}[0,202(0,1)^3 + 0,798 \cdot 0,3 \cdot (0,117 \cdot 0,09 + 0,883 \cdot 0,04)]}{0,3 \cdot 0,6} = 0,621 \cdot 10^{-5};$$

$$z = \bar{z}\beta_z = 0,621 \cdot 10^{-5} \cdot 8,9 \cdot 10^8 \approx 5527 \text{ м}.$$

Непосредственный расчет по заданной математической формуле дает

$$z = \frac{300^2 + 600(750^2 + 6 \cdot 560^2)}{0,2 \cdot 900 \cdot 1500} \approx 5531 \text{ м}.$$

Некоторое различие в результатах связано с наличием погрешностей округления на промежуточных этапах вычислений.

В вычислительных алгоритмах часто встречаются различные математические функции. Если архитектура ЭВМ предусматривает вычисление этих функций, например, имеются специальные операции в системе команд или подпрограммы в математическом обеспечении ЭВМ, то в описаниях этих команд и программ приводятся масштабы и пределы изменения самих математических функций и их аргументов. В противном случае программисту самому приходится выбирать численный метод и строить на его основе алгоритм вычисления функций. Масштабы и цены функций и аргументов рассчитываются тогда так же, как масштабы и цены основного алгоритма.

Пример 3.2.

$$z = 0,3\sqrt{\frac{5x^4 + y^4}{xy}}; \quad x_{\max} = 200 \text{ м}, \; y_{\max} = 300 \text{ м}, \; x_{\min} = 20 \text{ м}, \; y_{\min} = 30 \text{ м}, \; n = 10.$$

Арифметика целочисленная. Кроме того, $B_{\sqrt{a}} = \sqrt{B_a \cdot 2^{-n}}$. Запись машинной операции извлечения корня - $\sqrt{A \cdot 2^n}$. Проверку выполнить для x=120 м, y=150 м.

Решение. $B_x = 200 \cdot 2^{-10}$ м; $B_y = 300 \cdot 2^{-10}$ м.

1) Возведение x в степень 2: $z_1 = x^2$;

   $B_{z_1} = B_x^2 \cdot 2^{10} = 4 \cdot 10^4 \cdot 2^{-10} \text{м}^2$; $\overline{Z}_1 = \overline{X} \cdot \overline{X}$ (запись $\overline{Z}_1 = \overline{X} \cdot \overline{X} \cdot 2^{-10}$).

2) Возведение x в степень 3: $z_2 = z_1 \cdot x$;

   $B_{z_2} = B_{z_1} \cdot B_x \cdot 2^{10} = 8 \cdot 10^6 \cdot 2^{-10} \text{м}^3$; $\overline{Z}_2 = \overline{Z}_1 \cdot \overline{X}$ (запись $\overline{Z}_2 = \overline{Z}_1 \cdot \overline{X} \cdot 2^{-10}$).

3) Возведение x в степень4: $z_3 = z_2 \cdot x$;

   $B_{z_3} = B_{z_2} \cdot B_x \cdot 2^{10} = 16 \cdot 10^8 \cdot 2^{-10} \text{м}^4$; $\overline{Z}_3 = \overline{Z}_2 \cdot \overline{X}$ (запись $\overline{Z}_3 = \overline{Z}_2 \cdot \overline{X} \cdot 2^{-10}$).

4) Возведение x в степень 4: $z_4 = 5z_3$ (изменение масштаба);

   $B_{z_4} = 5B_{z3} = 80 \cdot 10^8 \cdot 2^{-10} \text{м}^4$; $\overline{Z}_4 = \overline{Z}_3$

5) Возведение y в степень 2: $z_5 = y^2$;

$B_{z_5} = B_y \cdot B_y \cdot 2^{10} = 9 \cdot 10^4 \cdot 2^{-10}$ м^2; $\overline{Z}_5 = \overline{Y} \cdot \overline{Y}$ (запись $\overline{Z}_5 = \overline{Y} \cdot \overline{Y} \cdot 2^{-10}$).

6) Возведение y в степень 3: $z_6 = z_5 \cdot y$;

$B_{z_6} = B_{z_5} \cdot B_y \cdot 2^{10} = 27 \cdot 10^6 \cdot 2^{-10}$ м^3; $\overline{Z}_6 = \overline{Z}_5 \cdot \overline{Y}$ (запись $\overline{Z}_6 = \overline{Z}_5 \cdot \overline{Y} \cdot 2^{-10}$).

7) Возведение y в степень 4: $z_7 = z_6 \cdot y$;

$B_{z_7} = B_{z6} \cdot B_y \cdot 2^{10} = 81 \cdot 10^8 \cdot 2^{-10}$ м^4; $\overline{Z}_7 = \overline{Z}_6 \cdot \overline{Y}$ (запись $\overline{Z}_7 = \overline{Z}_6 \cdot \overline{Y} \cdot 2^{-10}$).

8) Сложение $z_8 = z_4 + z_7$;

так как $B_{z_4} \neq B_{z_7}$, то $\overline{Z}_8 = \overline{Z}_4 K_4 + \overline{Z}_7 K_7$ (запись $\overline{Z}_8 = \overline{Z}_4 K_4 \cdot 2^{-10} + \overline{Z}_7 K_7 \cdot 2^{-10}$);

$$B_{z_8}^{Пр} = \frac{5x^4_{\max} + y^4_{\max}}{2^{10}} = 161 \cdot 10^8 \cdot 2^{-10} \text{ м}^4; \quad B_{z_8} = \max\left\{B_{z_8}^{Пр};\ B_{z_7}^{Пр};\ B_{z_4}\right\} = 161 \cdot 10^8 \cdot 2^{-10} \text{ м}^4;$$

$$K_4 = \frac{B_{z_4}}{B_{z_8}} 2^{10} = \frac{80 \cdot 10^8 \cdot 2^{-10}}{161 \cdot 10^8 \cdot 2^{-10}} 2^{10} = \frac{80}{161} 2^{10}; \qquad K_7 = \frac{B_{z_7}}{B_{z_8}} 2^{10} = \frac{81 \cdot 10^8 \cdot 2^{-10}}{161 \cdot 10^8 \cdot 2^{-10}} 2^{10} = \frac{81}{161} 2^{10}.$$

9) Вычисление произведения $z_9 = xy$;

$B_{z_9} = B_x B_y 2^{10} = 6 \cdot 10^4 \cdot 2^{-10}$ м^2; $\overline{Z}_9 = \overline{X} \cdot \overline{Y}$ (запись $\overline{Z}_9 = \overline{X} \cdot \overline{Y} \cdot 2^{-10}$).

10) Деление $z_{10} = z_8 / z_9$;

$$B_{z_8}^* = \frac{|z_{8\max}|}{|z_{9\max}|} B_{z_9} = \frac{|5x^4_{\max} + y^4_{\max}|}{|x_{\min} y_{\min}|} 6 \cdot 10^4 \cdot 2^{-10} = 161 \cdot 10^{10} \cdot 2^{-10} \text{ м}^4.$$

$B_{z_8} < B_{z_8}^*$, поэтому $B_{z_{10}} = \frac{B_{z_8}^*}{B_{z_9}} 2^{-n} = \frac{161 \cdot 10^{10} \cdot 2^{-10}}{6 \cdot 10^4 \cdot 2^{-10}} 2^{-10} = \frac{161}{6} \cdot 10^6 \cdot 2^{-10}$ м^2.

$\overline{Z}_{z_{10}} = \frac{\overline{Z}_8 \cdot K_8}{\overline{Z}_9}$ (запись $Z_{10} = \frac{\overline{Z}_8 K_8 \cdot 2^{-10}}{\overline{Z}_9} 2^{10}$); $K_8 = \frac{B_{z_8}}{B_{z_8}^*} 2^{10} = \frac{161 \cdot 10^8 \cdot 2^{-10} \cdot 2^{10}}{161 \cdot 10^{10} \cdot 2^{-10}} = 10^{-2} \cdot 2^{10}$.

11) Извлечение корня $z_{11} = \sqrt{z_{10}}$;

$Bz_{11} = \sqrt{Bz_{11} \cdot 2^{-n}} = \sqrt{\frac{161}{6} 10^6 \cdot 2^{-10} \cdot 2^{-10} \cdot 2^{-10}} = \sqrt{\frac{161}{6} 10^3 \cdot 2^{-10}}$; $\overline{Z}_{11} = \sqrt{\overline{Z}_{10}}$ (запись $\overline{Z}_{11} = \sqrt{\overline{Z}_{10} \cdot 2^{10}}$).

12) Умножение на константу 0,3: $z = z_{12} = 0,3 z_{11}$ (изменение масштаба);

$B_z = B_{z_{12}} = 0,3 \cdot B_{z_{11}} = 0,3 \sqrt{\frac{161}{6} 10^3 \cdot 2^{-10}}$ м; $\overline{Z} = \overline{Z}_{12} = \overline{Z}_{11}$.

Расчетная формула:

$\overline{Z} = \sqrt{\frac{(\overline{X}^4 K_4 + \overline{Y}^4 K_7) K_8}{\overline{X} \cdot \overline{Y}}}$ (запись для проверки $\overline{Z} = \sqrt{\frac{(\overline{X}^4 K_4 \cdot 2^{-40} + \overline{Y}^4 K_7 \cdot 2^{-40}) K_8}{\overline{X} \cdot \overline{Y} \cdot 2^{-10}}} \cdot 2^{10}$).

Проверка:

$$\overline{Z} = \sqrt{\frac{[(0,6 \cdot 2^{10})^4 \cdot \frac{80}{161} \cdot 2^{10} \cdot 2^{-40} + (0,5 \cdot 2^{10})^4 \frac{81}{161} 2^{10} \cdot 2^{-40}] \cdot 10^{-2} \cdot 2^{10}}{0,6 \cdot 2^{10} \cdot 0,5 \cdot 2^{10} \cdot 2^{-10}}} \cdot 2^{10} = 5,652186 \cdot 10^{-2} \cdot 2^{10};$$

$$z = \overline{Z}B_z = 5,652186 \cdot 10^{-2} \cdot 2^{10} \cdot 0,3 \cdot \sqrt{\frac{161}{6}} \cdot 10^3 \cdot 2^{-10} = 87,8353 .$$

Непосредственный расчет по исходной формуле дает

$$z = 0,3\sqrt{\frac{5 \cdot 120^4 + 150^4}{120 \cdot 150}} \approx 87,8365 .$$

Вопросы и задачи для самоподготовки

1. Какое масштабирование называется постоянным?
2. Чем переменное масштабирование отличается от постоянного?
3. Перечислить достоинства и недостатки переменного масштабирования.

В задачах 4-14 выполнить постоянное масштабирование заданных математических выражений. Выполнить проверку для конкретных значений переменных, указанных в скобках.

4. $z = a + bx + cx^2$, $a_{\max} = 8000$ м, $b_{\max} = 200$ м/с, $c_{\max} = 60$ м/с^2, $x_{\max} = 25$ с, $n = 10$.

Арифметика целочисленная.

($a = 3000$ м, $b = 120$ м/с, $c = 20$ м/с^2, $x = 10$ с).

5. $z = \dfrac{ax^2 + 3b^2}{(a+b)^2}$; $a_{\max} = 200$, $x_{\max} = 300$, $b_{\max} = 250$, $a_{\min} = 20$, $b_{\min} = 25$, $n = 12$.

Арифметика дробная. ($a = 50$, $x = 30$, $b = 40$).

6. $z = \dfrac{2xy}{\sqrt{3x^2 + 5y^2}}$; $x_{\max} = 300^0$, $y_{\max} = 360^0$, $x_{\min} = 10^0$, $y_{\min} = 10^0$, $n = 10$.

Арифметика дробная.

$\beta_{\sqrt{a}} = \sqrt{\beta a}$; $\overline{\sqrt{a}} = \sqrt{\overline{a}}$; ($x = 1000^0$, $y = 60^0$).

7. $z = \sqrt{\left(\dfrac{a+b}{c+d}\right)^2} + 1,64t$; $a_{\max} = 300$, $b_{\max} = 500$, $c_{\max} = 200$, $d_{\max} = 200$, $c_{\min} = 30$, $d_{\min} = 10$, $t_{\max} = 300$, $n = 10$.

Арифметика дробная.

$\beta_{\sqrt{a}} = \sqrt{\beta a}$; $\overline{\sqrt{a}} = \sqrt{\overline{a}}$; ($a = 100$, $b = 200$, $c = 50$, $d = 50$, $t = 1000$).

8. $z = \pi(R + r)l + \pi R^2 + \pi r^2$, $R_{\max} = 50$ см, $r_{\max} = 20$ см, $l_{\max} = 100$ см, $n = 12$.

Арифметика целочисленная. ($R = 20$ см, $r = 10$ см, $l = 60$ см).

9. $z = \dfrac{2}{a}\sqrt{P(p-a)(p-b)(p-c)}$, где $p = \dfrac{a+b+c}{2}$;

$a_{\max} = 60$ см, $b_{\max} = 80$ см, $c_{\max} = 100$ см, $a_{\min} = 10$ см, n=10.

Арифметика дробная.

$\beta_{\sqrt{a}} = \sqrt{\beta a}$; $\overline{\sqrt{a}} = \sqrt{\overline{a}}$; ($a = 30$ см, $b = 50$ см, $c = 40$ см).

10. $z = \dfrac{2\sqrt{bcp(p-a)}}{b+c}$, где $p = \dfrac{a+b+c}{2}$; $(b+c)_{\min} = 20$ см, далее условия примера 9.

11. $z=\frac{1}{2}\sqrt{2b^2+2c^2-a^2}$, далее условия примера 9.

12. $z=\sqrt{2r\ -2r\sqrt{r^2-\frac{a^2}{4}}}$; $a_{\max}=60$ см, $r_{\max}=100$ см, $n=12$.

Арифметика дробная.

$\beta_{\sqrt{a}}=\sqrt{\beta_a}$; $\overline{\sqrt{a}}=\sqrt{\overline{a}}$; ($a=20$ см, $r=30$ см).

13. $z=\sqrt{x^2+y^2+H^2}$, $x_{\max}=y_{\max}=800$м, $H_{\max}=300$ м, $n=10$.

Арифметика целочисленная.

$B_{\sqrt{a}}=\sqrt{B_a\cdot 2^{-n}}$; $\overline{\sqrt{a}}=\sqrt{\overline{A}\cdot 2^n}$; ($x=5000$ м, $y=3000$ м, $H=1000$ м).

14. $z=b\cdot\sin(23^0+1,4x)$; $x_{\max}=280^0$, $\beta_{\sin}=1$, $\beta_{аргумента}=360^0$, $b_{\max}=6$м, $n=11$.

Арифметика дробная.

4. Особенности переменного масштабирования данных

Как уже отмечалось, переменное масштабирование характеризуется изменением масштабов одних и тех же физических величин на различных этапах вычислительного алгоритма. Применяется это в тех случаях, когда выбранные масштабы величин не являются предельными и имеется запас по разрядной сетке в двоичных представлениях этих величин, и когда необходимо повысить точность вычислений на различных диапазонах изменения одних и тех же физических величин.

Пример 4.1. Выполнить масштабирование $z=t(200-20t)$ при $t=0-10$, $B_t=2^{-7}$, $n=13$. Арифметика целочисленная.

Решение. Выполним сначала постоянное масштабирование z.

1) Умножение 20 на t (умножение на константу): $z_1=20t$;

$$Bz_1=20B_t=20\cdot 2^{-7};\ \overline{z}_1=\overline{t}.$$

2) Вычитание $z_2=200-z_1$.

Так как 200 есть константа, выберем ее цену равной Bz_1. Тогда $Bz_2=Bz_1$, а машинный эквивалент 200 будет равен: $200=200:20\cdot 2^{-7}=10\cdot 2^7$. Машинная операция будет иметь вид:

$$\overline{Z}_2=10\cdot 2^7-\overline{Z}_1.$$

3) Умножение $z_3=z_2t$;

$B_{z_3}=B_{z_2}B_t\cdot 2^{13}=20\cdot 2^{-7}\cdot 2^{-7}\cdot 2^{13}=10$, $\overline{Z}_3=\overline{Z}_2\cdot\overline{T}$ (запись $\overline{Z}_3=\overline{Z}_2\cdot\overline{T}\cdot 2^{-13}$).

Таким образом, цена результата равна 10. Определим количество значащих разрядов, которое будет занимать максимальное значение z . Анализируя функцию z, можно показать, что на интервале 0-10 ее максимальное значение равно 500. Тогда $\overline{Z}_{\max}=\frac{500}{10}=50=000000110010$, т.е.

результат занимает всего 6 разрядов из 13 отведенных Точность представления z получена явно ниже возможной.

Потеря точности в данном примере произошла при выполнении операции умножения. Действительно, максимальное значение z_2 , равное 100, при цене $20\cdot 2^{-7}$ занимает 10 младших разрядов 13-разрядной сетки, а соответствующее ему значение $t = 5$ тоже 10 разрядов. Их произведение в $2n$ – разрядной сетке займет не более 20 младших разрядов. Поэтому последующее усечение до 13 разрядов приведет к потере 13 значащих разрядов произведения.

Точность результата можно повысить, изменив масштабы z_2 и t перед выполнением операции умножения. Сделаем это следующим образом. Так как $\overline{Z}_2$ может занимать не более 10 разрядов, сдвинем его влево на 3 разряда, что не вызовет переполнения разрядной сетки. Максимальное t, равное 10, будет занимать 11 младших разрядов, поэтому $\overline{T}$ можно сдвинуть влево без переполнения на 2 разряда. Цены после этого будут равны $B_{z_2} = 20\cdot 2^{-7}\cdot 2^{-3} = 20\cdot 2^{-10}$, $B_t = 2^{-7}\cdot 2^{-2} = 2^{-9}$. Цена результата z станет равной $B_z = B_{z_2}\cdot B_t\cdot 2^{13} = 20\cdot 2^{-10}\cdot 2^{-9}\cdot 2^{13} = 10\cdot 2^{-5}$. В этом случае максимальное значение z будет занимать уже не менее 11 двоичных разрядов. Расчетная формула для z при переменном масштабировании примет вид:

$$\overline{Z} = (2^2\cdot\overline{T})[(10\cdot 2^7 - \overline{Z}_1)\cdot 2^3].$$

Дополнительные затраты машинного времени связаны только со сдвигом влево $\overline{T}$ и $\overline{Z}_2$ общим число на 5 разрядов и не являются существенными.

Пример 4.2. Выполнить масштабирование в целочисленной арифметике

алгоритма $x = x_0 + \dot{x}t,\ y = y_0 + \dot{y}t,\ z = x/y$

при

$$x_{0_{max}} = y_{0_{max}} = 1024\text{м},\ \dot{x}_{max} = \dot{y}_{max} = 64\ \text{м/с},\ t_{max} = 16\text{с},\ B_x = B_y = B_{x_0} = B_{y_0} = 0{,}5\ \text{м},\ n = 12.$$

Кроме того, известно, что x и y в алгоритме связаны между собой таким образом, что никогда $|x|$ не может быть больше $|y|$.

Решение. Выберем цены переменных $\dot{x}$, $\dot{y}$ и t из предельных соотношений:

$$B_{\dot{x}} = B_{\dot{y}} = \frac{\dot{x}_{max}}{2^n} = 64\cdot 2^{-12} = 2^{-6}\ \text{м/с};\qquad B_t = \frac{t_{max}}{2^n} = 2^{-8}\text{с}.$$

Теперь выполним постоянное масштабирование алгоритма:

1) умножение $z_1 = \dot{x}t$;

$$B_{z_1} = B_{\dot{x}}B_t\cdot 2^n = 2^{-5}\cdot 2^{-8}\cdot 2^{12} = 2^{-2}\ \text{м}.$$

Машинная операция - $\overline{Z}_1 = \overline{\dot{X}}\cdot\overline{T}$; запись машинной операции - $\overline{Z}_1 = \overline{\dot{X}}\cdot\overline{T}\cdot 2^{-12}$.

2) вычисление $x = x_0 + z_1$.

Так как $B_{x_0} \neq B_{z_1}$, то перед операцией сложения необходимо выравнить цены слагаемых. В данном примере это возможно путем сдвига $\overline{Z}_1$ на 1 разряд вправо (сдвиг вправо без изменения величины числа). Тогда $\overline{Z}_1$ будет равен 2^{-1} м, т.е.

равен B_{x_0}. В этом случае машинная операция $\overline{X} = \overline{X}_0 + (\overline{\dot{X}} \cdot \overline{T}) \cdot 2^{-1}$. Условная запись ее: $\overline{X} = \overline{X} + (\overline{\dot{X}} \cdot \overline{T} \cdot 2^{-12}) \cdot 2^{-1}$. Цена B_x=0,5 м;

3) вычисление y.

Поскольку цены $\dot{y}$ и y равны соответственно ценам $\dot{x}$ и x, то вычисление y выполняется аналогично x, т.е. $\overline{Y} = \overline{Y}_0 + (\overline{\dot{Y}} \cdot \overline{t}) \cdot 2^{-1}$. Условная запись машинной операции: $\overline{Y} = \overline{Y}_0 + (\overline{\dot{Y}} \cdot \overline{t} \cdot 2^{-12}) \cdot 2^{-1}$. Цена B_y=0,5 м;

4) деление $z = x / y$.

Так как цены x и y равны, то по условию задачи следует, что $|\overline{X}| < |\overline{Y}|$, и условие правильного деления выполняется при любых значениях $\overline{X}$ и $\overline{Y}$. Поэтому $B_z = \frac{B_x}{B_y} 2^{-n} = \frac{0,5}{0,5} \cdot 2^{-12} = 2^{-12}$, а машинная операция примет вид $\overline{Z} = \frac{\overline{X}}{\overline{Y}}$. Условная запись ее такова: $\overline{Z} = \frac{\overline{X}}{\overline{Y}} \cdot 2^n$.

Таким образом, при постоянном масштабировании цена результата равна $B_z = 2^{-12}$, расчетная формула имеет вид:

$$\overline{Z} = \frac{\overline{X}_0 + (\overline{\dot{X}} \cdot \overline{T}) \cdot 2}{\overline{Y}_0 + (\overline{\dot{Y}} \cdot \overline{T}) \cdot 2^{-1}}^{-1},$$

а ее условная запись

$$\overline{Z} = \frac{\overline{X}_0 + (\overline{\dot{X}} \cdot \overline{T} \cdot 2^{-12}) \cdot 2}{\overline{Y}_0 + (\overline{\dot{Y}} \cdot \overline{T} \cdot 2^{-12}) \cdot 2^{-1}}^{-1} 2^{12}.$$

Проверим результаты масштабирования для двух вариантов значений переменных. В первом используем значения переменных, близкие к максимальным, а во втором – близкие к минимальным.

Вариант 1. Пусть $x_0 = y_0 = 1024$ м, $\dot{y} = 3847 \cdot 2^{-6}$ м/с, $\dot{x} = 3207 \cdot 2^{-6}$ м/с, $t = 10$ с. Тогда $\overline{X}_0 = \overline{Y}_0 = \frac{1024}{0,5} = 2048$; $\overline{\dot{Y}} = \frac{\dot{y}}{B_y} = \frac{3847 \cdot 2^{-6}}{2^{-8}} = 3847$; $\overline{\dot{X}} = \frac{\dot{x}}{B_x} = \frac{3207 \cdot 2^{-6}}{2^{-6}} = 3207$. Основная доля погрешности в результате $\overline{Z}$ будет внесена при выполнении операций умножения на $\overline{Y}$ и $\overline{X}$ на t. Так умножение $\overline{\dot{Y}}$=3847=111100000111 на $t = 10 \cdot 2^8 = 101000000000$ с усечением результата до 12 разрядов дает машинное число 100101100100 = 2404, а не 2404,4. Аналогично $\overline{\dot{X}}$=3207=110010000111—2004 вместо 2004,4. Поэтому

$$\overline{Z} = \frac{\overline{X}_0 + (\overline{\dot{X}} \cdot \overline{t} \cdot 2^{-12}) \cdot 2}{\overline{Y}_0 + (\overline{\dot{Y}} \cdot \overline{t \cdot 2^{-12}}) \cdot 2^{-1}}^{-1} 2^{12} = \frac{2048 + 1002}{2048 + 1202} 2^{12} = 3844,$$

$$z = \overline{Z} B_z = 3844 \cdot 2^{-12} = 0,93847656$$

Действительное значение z равно

$$z=\frac{x_0+\dot{x}t}{y_0+\dot{y}t}=\frac{1024+3207\cdot 2^{-6}\cdot 10}{1024+3847\cdot 2^{-6}\cdot 10}=\frac{1024+501{,}1}{1024+601{,}1}=0{,}93846532.$$

Вариант 2. Пусть $x_0=1$ м, $y_0=1{,}5$ м, $\dot{y}=15\cdot 2^{-6}$ м/с, $\dot{x}=6\cdot 2^{-6}$ м/с, t = 4 с. Машинные величины $\overline{X}_0$=2, $\overline{Y}_0$=3, t =1024, $\dot{X}$=6, $\dot{\overline{Y}}$=15. В этом случае при умножении с усечением $\dot{\overline{X}}$ на $\overline{T}$ результат равен $6\cdot 1024\cdot 2^{-12}=1$ (вместо 1,5) и при последующем масштабном сдвиге вправо на 1 разряд (множитель 2^{-1} в расчетной формуле) окажется равным 0. При умножении $\dot{\overline{Y}}$ на $\overline{T}$ с последующим масштабным сдвигом получим: $15\cdot 2^{10}\cdot 2^{-12}\cdot 2^{-1}=15/8=1$. Следовательно,

$$\overline{Z}=\frac{2+0}{3+1}\cdot 2^{12}=2048;\quad z=2^{11}\cdot 2^{-12}=0{,}5.$$

Действительное значение

$$z=\frac{1+6\cdot 2^{-6}\cdot 4}{1{,}5+15\cdot 2^{-6}\cdot 4}=0{,}5641;\quad \Delta z=0{,}0641;\ \delta z=11{,}35\%.$$

Таким образом, погрешность результата в зоне малых значений переменных существенно больше. Чтобы ее уменьшить, применим переменное масштабирование. Для этого разобьем интервалы изменения исходных данных x_0, y_0, $\dot{x}$ и $\dot{y}$ на два диапазона. Первый будет включать их значения от максимальной величины до максимальной, уменьшенной в 2^6 раз, а второй – от максимальной, уменьшенной в 2^6 раз, до минимальной. Например, для $\dot{x}$ такое разбиение математически выражается следующими неравенствами:

$\dot{x}_{\max}\ge\dot{x}\ge\dot{x}_{\max}\cdot 2^{-6}$; $\dot{x}_{\max}\cdot 2^{-6}>\dot{x}\ge\dot{x}_{\min}$. Для машинных чисел первый и второй диапазон определяется соответственно такими объединенными неравенствами:

$$2^{12}\ge\overline{X}_0;\overline{Y}_0;\dot{\overline{X}};\dot{\overline{Y}}\ge 2^6;\ 2^6>\overline{X}_0;\overline{Y}_0;\dot{\overline{X}};\dot{\overline{Y}}\ge 1.$$

Далее будем поступать следующим образом. Если хотя бы одна исходная величина попадает в первый диапазон, будем использовать методику и формулы постоянного масштабирования. А если все они попадают во второй диапазон, то выполним предварительный сдвиг машинных чисел $\overline{X}_0$, $\overline{Y}_0$, $\dot{\overline{X}}$, $\dot{\overline{Y}}$ на 6 разрядов влево, уменьшив их цены в 2^{-6} раз, и только после этого воспользуемся расчетной формулой постоянного масштабирования. Поскольку в расчетной формуле необходимо делать масштабный сдвиг вправо на 1 разряд произведений $\dot{\overline{X}}\cdot\overline{T}$ и $\dot{\overline{Y}}\cdot\overline{T}$, его можно учесть без выполнения, если предварительный сдвиг $\dot{\overline{X}}$ и $\dot{\overline{Y}}$ делать на 5 разрядов влево, тем самым уменьшив число дополнительных сдвигов. С учетом этого расчетная формула для 2-го диапазона примет вид:

$$\overline{Z}=\frac{\overline{X}_0\cdot 2^6+(\dot{\overline{X}}\cdot 2^5\cdot\overline{T}\cdot 2^{-12})}{\overline{Y}_0\cdot 2^6+(\dot{\overline{Y}}\cdot 2^5\cdot\overline{T}\cdot 2^{-12})}2^{12}.$$

Цена результата останется прежней. Временные затраты на реализацию алгоритма возрастут на время анализа диапазонов и время дополнительных сдвигов, однако точность будет повышена. Покажем это на примере данных второго варианта. Они попадают все во второй диапазон, поэтому

$$\overline{Z} = \frac{2\cdot 2^6 + 6\cdot 2^{10}\cdot 2^5\cdot 2^{-12}}{3\cdot 2^6 + 15\cdot 2^{10}\cdot 2^5\cdot 2^{-12}}\cdot 2^{12} = \frac{(2\cdot 2^6 + 6\cdot 2^3)\cdot 2^{12}}{3\cdot 2^6 + 15\cdot 2^3} = 2310{,}56 \approx 2310.$$

Величина $z = \overline{Z}\cdot B_z = 0{,}56365$, $\Delta z = 0{,}000135$, $\delta z = 0{,}024\%$.

Единственным источником погрешности в данном примере остается усечение результатов деления. Если деление делается с округлением, то получим $\overline{Z} = 2311$ и $\Delta z = -0{,}00011$, $\delta z = 0{,}019\%$.

В свою очередь второй диапазон данных также можно разбить на два диапазона, повысив точность и для других пределов суммирования величин. Рациональное число диапазонов выбирают в каждой задаче свое в зависимости от ее условий.

Вопросы и задачи для самоподготовки

1. Как переменные масштабы используются для повышения точности решения?
2. Почему вообще возможно переменное масштабирование?
3. Почему стремятся машинные множители представлять в двоично-рациональном виде?

Далее в задачах 4-5 выполнить постоянное и переменное масштабирование, сравнить их результаты.

4. $z = v(10 - 5v)$ при $\beta_v = 2^3$, $v = 0{,}5 - 2$, $n >> 1$. Арифметика дробная.
5. $z = c_0 - c_1 t + c_2 t^2$ при $c_0 = 0 - 1000$ м, $c_1 = 0 - 100$ м/с,
   $c_2 = 0 - 20$ м/с^2, $t = 0 - 8$с, n = 10, $\beta_{c_2} = 1$ м, $\beta_{c_1} = 2^{-3}$ м/с, $\beta_{c_2} = 2^{-4}$ м/с^2, $\beta_t = 2^{-5}$ с.

В задачах 6-7 выполнить масштабирование, разбив интервал изменения данных за два диапазона так, как это было сделано в примере 4.2.

6. $x = x_0 - xt$; $z = \frac{x}{y}$, $x_{0_{max}} = 1024$м, $x_{max} = 32$ м/с, $t_{max} = 8$с, $y = const = 1024$ м, $n = 12$.
   Арифметика дробная.
7. $x = x_0 + xt$, $y = y_0 + yt$, $z = \frac{x}{y}$, $x_{0_{max}} = y_{0_{max}} = 2048$м, $x_{max} = y_{max} = 127$ м/с, $t_{max} = 32$с,
   $B_x = B_y = B_{x_0} = B_{y_0} = 1$м, $n = 13$, $|x| < |y|$.
   Арифметика целочисленная.