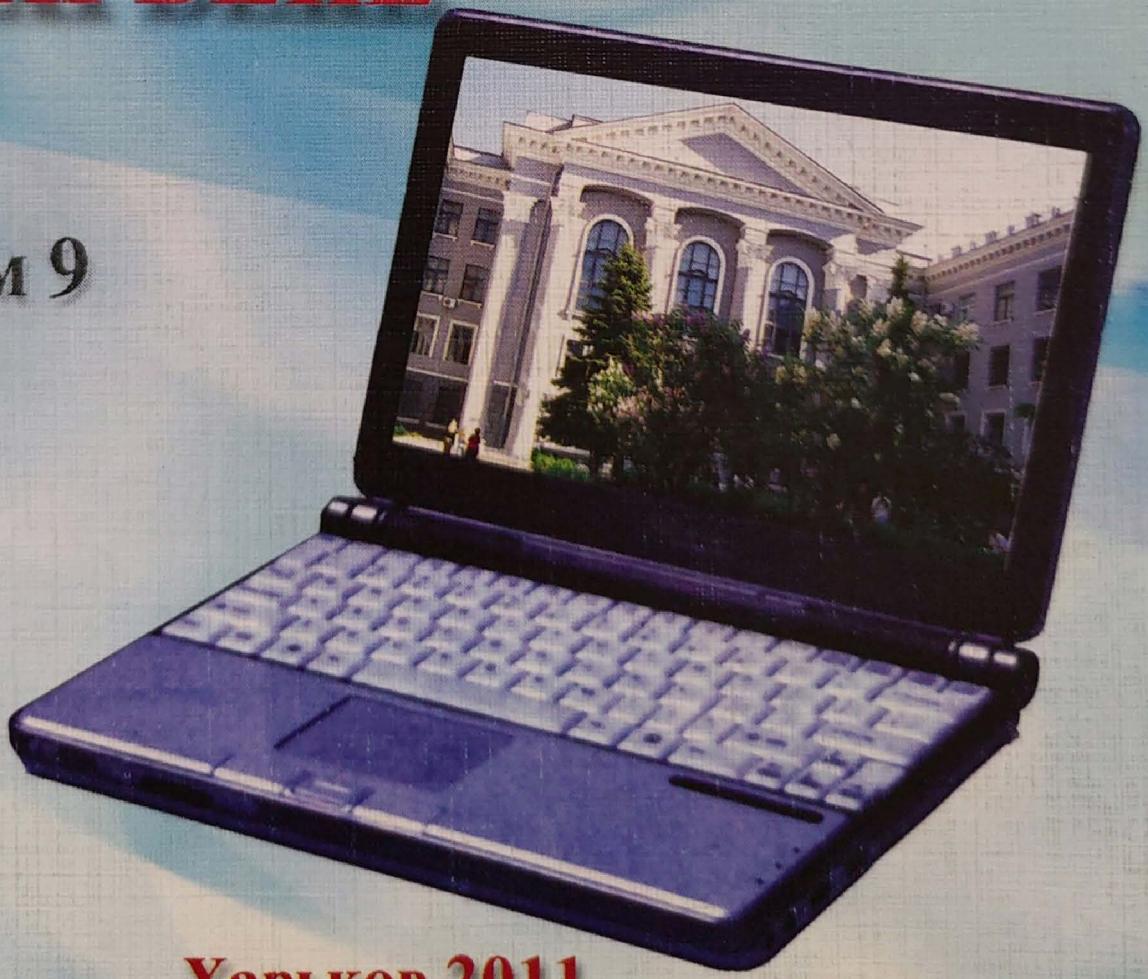
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ,
МОЛОДЕЖИ И СПОРТА УКРАИНЫ
ХАРЬКОВСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ
РАДИОЭЛЕКТРОНИКИ

МАТЕРИАЛЫ
XV МЕЖДУНАРОДНОГО
МОЛОДЕЖНОГО ФОРУМА

РАДИОЭЛЕКТРОНИКА И МОЛОДЕЖЬ В XXI ВЕКЕ

Том 9



Харьков 2011

Министерство образования и науки, молодежи и спорта Украины
ХАРЬКОВСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ
РАДИОЭЛЕКТРОНИКИ

МАТЕРИАЛЫ 15-го ЮБИЛЕЙНОГО МЕЖДУНАРОДНОГО МОЛОДЕЖНОГО ФОРУМА

«РАДИОЭЛЕКТРОНИКА И МОЛОДЕЖЬ В XXI веке»

18 – 20 апреля 2011 г.

Том 9

МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ

«ИНФОРМАЦИОННЫЕ ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ»

Харьков 2011

15-й Юбилейный Международный молодежный форум «Радиоэлектроника и молодежь в XXI веке». Сб. материалов форума. Т.9. - Харьков: ХНУРЭ. 2011. –672с.

В сборник включены материалы 15-го Юбилейного Международного молодежного форума «Радиоэлектроника и молодежь в XXI веке».

Издание подготовлено факультетом компьютерных наук
Харьковского национального университета радиоэлектроники (ХНУРЭ)

61166 Украина, Харьков, просп. Ленина, 14
тел.: (057) 7021397
факс: (057) 7021515

E-mail: innov@kture.kharkov.ua

ОСОБЕННОСТИ РЕАЛИЗАЦИИ ФОРМАТА IEEE 754 ЧИСЕЛ С ПЛАВАЮЩЕЙ ТОЧКОЙ В КОМПИЛЯТОРАХ ЯЗЫКА C/C++

Мнушка О.В.
Научный руководитель – д.т.н., проф. Никонов О.Я.
Харьковский национальный автомобильно-дорожный университет
(61002, Харьков, ул. Петровского, 25,
каф. Информатики, тел. (057) 707-37-74)
e-mail: mnushka@live.com

The paper presents the results of the implementation of IEEE-754 in C/C+ compilers. It is shown that the data presented in this format, ambiguous interpreted by different compilers, standard IEEE 754—2008 partially supported by modem compilers, so to overcome the limitations on the accuracy of the calculations necessary to use arbitrary precision arithmetic.

Стандартом IEEE 754-2008 определены четыре формата двоичных чисел - binary!6, binary32, binary64, binary128, и три формата десятичных чисел - decimal32, decima!64, decimal128, а также способы расширения базовых форматов для повышения точности вычислений [I]. В настоящее время компиляторы C/C++ поддерживают данные форматы частично и ориентированы на предыдущую версию стандарта [2]. В работе проведено исследование поддержки данного формата в распространенных реализациях компиляторов C/C++, среди которых: Microsoft Visual C++ - из состава Microsoft Visual Studio 2008(2010); gcc 4.x (Debian 4.4.5-8) н 4.5.0 (mingw); Intel C Compiler 11.1.054 (Windows), 11.1.064 (Debian amd64), 11.1.074 (Debian x86). Исследование проводилось с целью выяснения необходимости применения алгоритмов вычисления с произвольной точностью при разработке программы компьютерного математического моделирования.

Таблица 1. Количество байт для хранения вещественного числа

Компилятор	float	double	long double
MS VC++ x86	4	8	8(!)
MS VC++ x86_64	4	8	8(!)
gcc/g++ x86	4	8	12/16[1]
gcc/g++ x86_x64	4	8	12[2]/16
icc x86	4	8	8[3]/12[4]/16[5]
icc x86_x64	4	8	8[3]/16[6]

Примечание: [1] с ключом -m128bit-long-double; [2] с ключом -m96bit-long-double; [3] Visual Studio по умолчанию; [4] Linux (libstdc++5); [5] Visual Studio с ключом /Qlong-double; [6] Linux x86_x64 (libstdc++5).

Анализ полученных данных (табл.1) показывает, что: 1) для данных в формате long double компилятором выделяется различное количество байт для хранения одного экземпляра переменной в памяти; 2) неоднозначная

интерпретация компиляторами C++ этого формата данных требует аккуратности и обоснованности его применения; 3) большое число байт в соответствии со стандартом [1] должно обеспечивать большую точность представления чисел.

Была проведена оценка результатов перехода от формата double к long double и его влияние на точность представления данных. Для этого вычислим число π по известной из тригонометрии формуле: $\pi = 4 \cdot arctg(1.0)$ (табл.2). Данное число с точностью до 40 разрядов: 3,1415926535 8979323846 2643383279 5028841971…

Таблица 2. Результат вычисления числа π

Компилятор	double	long double
MS VC++ 10.0	3,1415926535 89793100…	3,1415926535 89793100…
gcc/g++ x86	3,1415926535 89793115…	3,1415926535 89793238512… =-8879609370493449500 0000000000000000000000.0[1)]
gcc/g++ x86_x64	3,1415926535 89793115…	3,1415926535 89793238512…
icc 11.1.054 x86	3,1415926535 89793115…	3,1415926535 89793238512…
icc 11.1.054 x86_x64	3,1415926535 89793115…	3,1415926535 89793238512… -8.87960937049289e+043[2)] 0.000000[1)]

Примечание: [1)] вывод оператора pintf(); [2)] вывод оператора cout; жирным шрифтом выделен первый несовпадающий разряд.

Анализ полученных данных позволяет сделать следующие выводы: 1) переход от формата double к long double позволяет получить дополнительно 3 точных цифры результата; 2) несмотря на различный объем выделяемой памяти, переменные формата long double компиляторами icc и gcc/g++ преобразуются в формат extended (по IEEE754-1985); 3) результат может содержать «информационный мусор» (см. табл.2), который может быть неверно истолкован; 4) компилятор MS VC++ не определяет способов работы с форматами long double размером больше 8 байт, что вызывает проблемы с интероперабельностью программ; 6) переменные типа long double (больше 8 байт) могут некорректно выводиться как стандартным оператором (printf()), так и потоковым (cout) в ОС Windows. Могут не работать или оба способа (icc), или один из них (printf() в gcc/g++), что по-видимому связано с ограничениями соответствующих модулей ОС, отвечающих за работу с консолью. Результат, который будет отображаться на экране в данном случае, непредсказуем.

1. IEEE Standard for Floating-Point Arithmetic. – New York, 2008. – 70 P.
2. IEEE Standard for Binary Floating-Point Arithmetic. – New York, 1985. – 23 P.

Л

М

Н

Підп. до друку 28.03.11. Формат 60x841/16. Спосіб друку – ризографія.
Умов. друк. арк. 39,0. Облік. вид. арк.35,0. Тираж 352 прим.
Зам. № 2-292. Ціна договірна.

ХНУРЕ. Україна. 61166, Харків, просп. Леніна, 14

Віддруковано в навчально-науковому
видавничо-поліграфічному центрі ХНУРЕ
Харків, просп. Леніна, 14