

Re: Floating Point Issues

Source: <http://sci.tech-archive.net/Archive/sci.math.num-analysis/2008-01/msg00056.html>

- *From:* harper@xxxxxxxxxxxxxx (John Harper)
 - *Date:* 7 Jan 2008 09:56:25 +1300
-

In article <00dd5f63-496a-4cd9-a040-af5ddc6b62e1@xx>, user923005 <dcorbit@xxxxxxxx> wrote:

On Jan 2, 7:55 am, Bill Woessner <woess...@xxxxxxxx> wrote:

I need to evaluate an expression of the form $AB-CD$ and I only need the fractional portion of the result. Unfortunately, A and B are of the order 10^6 , C is of the order 10^9 and D is of the order 10^{-1} . So when I go to form my sum, which is of the order 10^{12} , I only get a few digits of precision in the fractional portion

If you can use C++, then I think that double-double will answer very nicely:

<http://www.cs.berkeley.edu/~yozo/>

Some Fortran compilers offer quadruple precision.

— John Harper, School of Mathematics, Statistics and Computer Science,
Victoria University, PO Box 600, Wellington 6140, New Zealand
e-mail john.harper@xxxxxxxx phone (+64)(4)463 5662 fax (+64)(4)463 5045