UNIVERSITY OF DUBLIN, TRINITY COLLEGE

Abstract

Faculty of Engineering, Mathematics and Science
Department of Computer Science and Statistics

Master of Science in Computer Science

Scientific computing with non-standard floating point types

May 2015

Author: Vlăduţ Mădălin Druţa

Supervisor: Dr. David Gregg

This study examined the possible use of non-standard floating point types for scientific computing. The question of this thesis is: "Is there anything to be gained by supporting non-standard floating point data types?".

There are several gaps in the literature that this thesis will aim to address. There could exist potential in the use of non-standard floating point types. This thesis investigates in particular the non-standard floating point type of 48-bit size. As long as there is no need for the full precision of floating point standard size of 64, the 48-bit non-standard type requires less memory, reduces the amount of data movement and might be faster than the standard size of 64-bit.

The initial findings showed that the non-standard (f48-bit) without the use of Streaming SIMD (Single Instruction Multiple Data) Extensions (SSE) is slower than using the standard 64 bit floating point. However, using SSE intrinsics the non-standard 48-bit floating point is competitive with the standard 64-bit. The results shown are good for a floating-point type that is not supported in hardware.