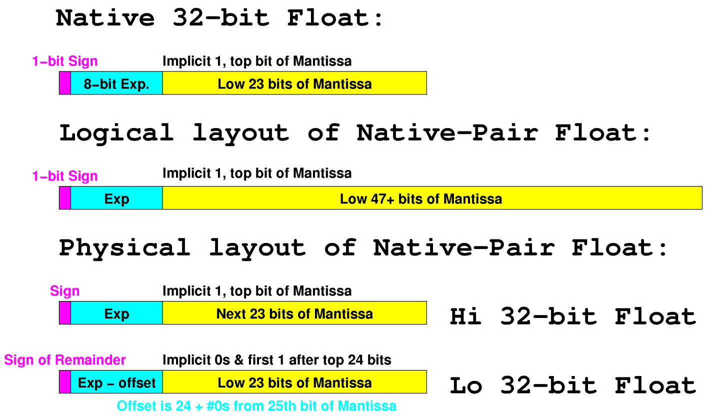
Horseshoes & Hand Grenades

"Close enough" is what floating-point arithmetic is all about, but that doesn't mean one can safely pretend to be operating on real numbers. Accuracy matters. The inconvenient fact is that what most systems allow the user to directly control is precision, not accuracy. Executing just a few arithmetic operations can make the relationship between precision and accuracy quite ponderous.

ADCs and DACs rarely attain 20-bit precision and there are 24 bits in an IEEE 32-bit SINGLE FLOAT's mantissa, thus many applications do not need hardware for directly implementing higher precisions. A similar argument has been made for 16-bit HALF FLOATs used to hold 8-bit or smaller digitized values. However, ignoring dynamic range issues, even 64-bit DOUBLE or 80-bit EXTENDED intermediate values can be insufficient to make some arithmetic sequences preserve the desired accuracy.

Native-Pair Arithmetic. One of the best ways to enhance accuracy is to augment a native-precision value with one or more residual error terms. This is most efficient using just a pair of native values:



Native 32-bit Float:

1-bit Sign
Implicit 1, top bit of Mantissa
8-bit Exp.
Low 23 bits of Mantissa

Logical layout of Native-Pair Float:

1-bit Sign
Implicit 1, top bit of Mantissa
Exp
Low 47+ bits of Mantissa

Physical layout of Native-Pair Float:

Sign
Implicit 1, top bit of Mantissa
Exp
Next 23 bits of Mantissa
Hi 32-bit Float

Sign of Remainder
Implicit 0s & first 1 after top 24 bits
Exp − offset
Low 23 bits of Mantissa
Lo 32-bit Float
Offset is 24 + #0s from 25th bit of Mantissa

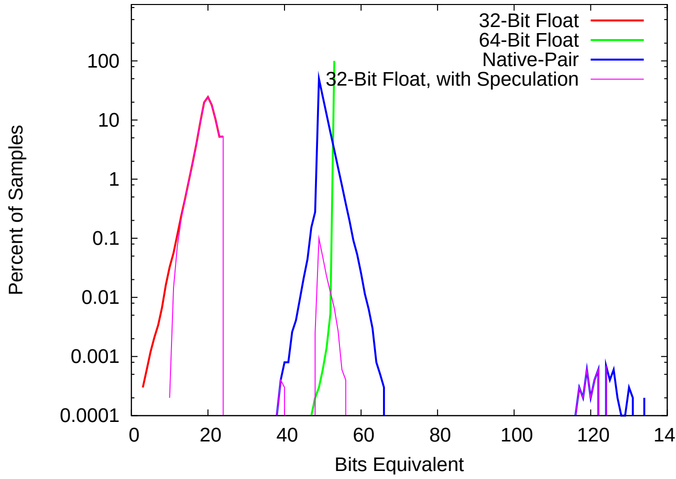
Native-pair operations are slower than native. The following table shows measured clocks for a Digital Signal Processor (TI320), SIMD Within A Register (ATHLON 3DNOW! and PENTIUM 4 SSE), and a Graphics Processing Unit (NVIDIA 6800).

Data Type	Target	+	−	×	/	√
32	DSP	1	1	1	42	51
pair(32)	DSP	11	11	25	112	119
32×2	SWAR	1	1	1	9	9
pair(32×2)	SWAR	24	28	27	57	40
pair(32×4)	SWAR	51	50	148	173	199
pair(64×2)	SWAR	45	48	48	50	–
$32 \times 4N$	GPU	1	1	–	4	20
pair$(32 \times 4N)$	GPU	11	11	18	35	28

The concept of error residuals has been around at least since Dekker's 1971 paper, *A floating-point technique for extending the available precision*. Our contributions center on algorithm and data layout changes tuning the performance for various modern targets, new analysis (native-pair values don't behave like ordinary higher-precision values), and simple architectural modifications to speed-up residual computations.

Speculative Precision. Native precision is good enough for many computations. Where it does not, it might fail only for relatively rare combinations of input values – in which case it might still be profitable to *try* using native precision. Our speculative precision work centers on language constructs that allow programmers to specify precisions and to mark blocks of code for speculative precision execution. Using a programmer-specified accuracy check, the compiled multiple-version code can dynamically recompute (or refine) insufficiently accurate results using increasingly higher precisions, including native-pair forms.

Native-pair works very well with speculative precision... in fact, on rare occassions SINGLE NATIVEPAIR yields far better accuracy than DOUBLE because the low 24 bits of a NATIVEPAIR mantissa can essentially separate from the high 24 bits to cover low bits that would have fallen off the bottom of a DOUBLE's 53-bit mantissa. The accuracies recorded summing many sets of 4,096 Gaussian-distributed pseudo-random numbers with zero mean and unit standard deviation are summarized below.



We also are investigating compiler methods for substitution of alternative algorithms based on accuracy. For example, recognizing a simple summation and replacing it with one of the many summation methods known to better preserve accuracy.

All of these tricks are particularly important for GPUs, which we are extending the SWARC language and SCC compiler system to target.

This document should be cited as:

```
@techreport{sc07accprec,
author={William Dieter and Henry Dietz},
title={Horseshoes and Hand Grenades},
institution={University of Kentucky},
address={http://aggregate.org/WHITE/sc07accprec.pdf},
month={Nov}, year={2007}}
```