Investigating Numerical Stability of Posit Number System and Offering an Alternative Model for Arithmetic

*Note: Sub-titles are not captured in Xplore and should not be used

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

***Abstract*—The IEEE floating-point standard format has maintained almost exclusive adherence for over 30 years despite its frequent criticisms. A particular concern with this standard is that the vast majority of applications do not utilize the wide dynamic range that it offer, implying that the bits should be organized to provide increased precision to numbers more commonly found in computation such as those that are close to unity. A recent proposition for implementing this design suggests the use of a run of identical bits after the sign bit in addition to a smaller sized exponent to encode the binary point as opposed to a fixed partition of exponent bits as is the case in floating point. This effectively creates tapered precision around unity by offering a smaller scaling component in this area which consequently allows for more bits to represent the fraction.**

The objective of this paper is to provide an introductory study into the practical effects of this tapered precision and to test the hypothesis that we can achieve greater numerical stability if we favor precision closer to unity. We investigate a handful of common applications including matrix inversion and image processing. We demonstrate that an LU factorization based direct solve for linear systems as well as Fast Fourier Transform derive moderate improvements to quality of solution when swapping to posit from traditional IEEE floating point wherearas the method of conjugate gradients for solving linear systems iteratively requires too large a dynamic range to gain much of an advantage. Furthermore, we show that the proposed use of loss-free addition does not endow posit with a unique advantage over IEEE floating point.

***Index Terms*—Posit, Floating-Point, Quire, Regime**

I. Introduction

and that applications that do depend on a wide dynamic range are more tolerant of coarseness in numerical precision.

good place to introduce bfloat and talk about why bfloat might be a better matchup against posit16 than half precision IEEE.

In this paper we emphasize the single precision case, but we also experiment with half precision as well as different posit configurations for 32 bits. We demonstrate that CG makes use of a wide range of numbers and consequently derives little benefit from switching to posit. However, image convolution and direct solve which are more stable in their nature obtain a modest performance increase from the switch to posit. We argue that ES=3 is a better option universally than ES=2 for single precision. Furthermore, we show that the use of quire while offering a substantial performance advantage, does not provide posit with a unique advantage over IEEE floating point.

Identify applicable funding agency here. If none, delete this.

II. Background

Introduce float/posit format, benefit of lowering ES, quire, other observations for posit such as only one NaR, and reciprocate by twos complementing lower bits. Also g

A. IEEE-754 Technical Standard

B. Posit Format

III. Related Work

IV. Conjugate Gradient Method

A. Overview of Algorithm

CG is a widely used iterative method for solving linear system, $Ax = b$, when the system of equations representing by the matrix A is positive-definite and symmetric. The algorithm is defined below.

...

The residual of the system is defined as $r_i = b - Ax_i$ while the error of the system is defined as $e_i = \hat{x} - x_i$ where $\hat{x} = A^{-1}b$. Because the matrix in question is symmetric and positive-definite, we may define an inner product $\langle \cdot \rangle_A$ with respect to A such that $\langle x, y \rangle_A = \langle Ax, y \rangle$. At each iteration $\|e_i\|_A = \sqrt{\langle e_i, e_i \rangle_A}$ is minimized by orthogonal projection (line 5) with respect to a basis for $\Re_n$ that is orthonormal with respect to $\langle \cdot \rangle_A$.

In its nature, CG is sensitive to numerical instability and poorly conditioned matrices. This is largely a result of the gradual loss of orthogonality (with respect to $\langle \cdot \rangle_A$) of $p_0, p_1, ..., p_i$ as well as the discrepancy between the computed

residual r_i and the true residual $b - Ax_i$ that results from accumulated roundoff error and the recurrence relation which is used to compute r_{i+1} from r_i (line 6). Despite these issues, the algorithm is appealing when performed iteratively, such that the algorithm terminates once $\|r_i\|$ dips under a specified tolerance bound.

B. Setup of Experiment

In this section we compare the performance of CG, defined by the number of iterations to converge, when swapping IEEE float for posit. Following *Ghysels, we chose $\hat{x} = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})^T$ such that $\|\hat{x}\| = 1$ and we assume convergence only when the norm of our residual $\|b - A\hat{x}\|$, drops below $\|b\| * 10^{-5}$. In practice, we would compute r_i using finite precision arithmetic for performance concerns; however, due to the discrepancy between r_i and the true residual $\hat{b} - Ax_i$ resulting from the limitations of finite precision arithmetic, it is possible for CG to assume convergence before the true residual falls under our convergence metric. This is especially problematic when dropping from double precision to single precision. In situations where this has a measurable effect, we will demonstrate a side to side comparison of the convergence plot where convergence is decided using finite precision, and a second experiment where convergence is decided by the true residual which is computed using infinite precision arithmetic: in this case all computation within the algorithm itself is limited to finite precision and only the stopping point may be affected.

All experiments are performed with real symmetric, positive-definite matrices of varying condition numbers that were obtained from (ghysels*). To avoid bias in our experimentation, we first load these matrices into memory by representing them using infinite precision after which we drop to finite precision.

C. Results and Discussion

V. Image Correlation

A. Introduction

In this section we consider the task of performing a correlation of two images where the first image represents the kernel and is used to locate shapes and patterns in the second image. This procedure is widely used in image processing and is exploited in machine learning by the use of convolutional neural networks as an effective procedure for image recognition. A fundamental result in signal processing is that a multiplication in the frequency domain is analagous to a convolution in the time/spatial domain, meaning that we can achieve the same result much faster by first taking the discrete fourier transform of our image and kernel (with proper padding) and perform a pointwise multiplication followed by an inverse fourier transform.

This fast-convolution procudure is fairly stable under heavy computation which makes it a tempting canditate for lower precision arithmetic. Given the recent emergence of low precision arithmetic in machine learning algorithms, we believe this would be an interesting area to explore the performance of half precision posit compared with half precision floating-point as well as Google's bfloat16.

B. Setup

We begin by taking an image as well an isolated section of the image to use as a kernel. The Kernel is zero-padded appropriately and the DFT of both images are computed using Cooley-Tukey algorithm followed by a pointwise multiplication and inverse DFT. We apply this same procedure using infinite precision arithmetic and use the mean-squared-error between finite and infinite precision as our metric for comparison. The mean squared error is shown alongside the computed result.

C. Results and Discussion

VI. Direct Solve

A. overview

A stable and straightforward approach for solving a dense linear system is to take advantage of the LU factorization and solve a pair of triangular systems directly. This more robust method does not suffer nearly as much from loss of precision. Since the method is not iterative by defualt, we are mainly interested in minimizing the final residual.

To exacerbate the numerical challenge we also chose to experiment with the solution to non-singular Vandermonde matrices. These matrices are relevant to the problem of polynomial interpolation which is well known to be an extremely ill conditioned for higher degree polynomials. This poses a unique challenge for the underlying numerical representation which should highlight any potential advantages that can be derived from using posit.

Similar to our experiments with image convolution, here we are also interested in the use of lower precision arithmetic. Much literature already exists for the use of low precision arithmetic as part of a direct solve. A frequently exploited trick is to compute an initial result through the LU decomposition, constituting the bulk of the computation, using lower precision and to iteratively improve this solution using higher precision. Effectively using low precision to condition the problem before solving it completely. Since we are most interested in exercising the limits of our underlying numerical stability using different number representation formats however, we will instead perform the entire computation using half precision.

Acknowledgment

References

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.