B.Eng. Final Year Project

Literature Notes

Sam Moore, David Gow
Faculty of Engineering, Computing and Mathematics, University of Western Australia
March 2014

Contents

1 Literature Summaries **4**

1.1 Postscript Language Reference Manual [1] . . . 4

1.2 Portable Document Format Reference Manual [2] . . . 4

1.3 IEEE Standard for Floating-Point Arithmetic [3] . . . 4

1.4 Portable Document Format (PDF) — Finally... [4] . . . 5

1.5 Pixels or Perish [5] . . . 6

1.6 Embedding and Publishing Interactive, 3D Figures in PDF Files [6] . . . 7

1.7 27 Bits are not enough for 8 digit accuracy [**?**] . . . 7

1.8 What every computer scientist should know about floating-point arithmetic [7] . . . 7

1.9 Compositing Digital Images [8] . . . 8

1.10 Bresenham's Algorithm: Algorithm for computer control of a digital plotter [9] . . . 8

1.11 Quad Trees: A Data Structure for Retrieval on Composite Keys [10] . . . 9

1.12 Xr: Cross-device Rendering for Vector Graphics [11] . . . 9

1.13 Glitz: Hardware Accelerated Image Compositing using OpenGL [12] . . . 9

1.14 Boost Multiprecision Library [13] . . . 9

1.15 A CMOS Floating Point Unit [14] . . . 9

1.16 Simply FPU[15] . . . 10

1.17 Floating Point Package User's Guide [16] . . . 10

1.18 Low-Cost Microarchitectural Support for Improved Floating-Point Accuracy[17] . . . 10

1.19 Accurate Floating Point Arithmetic through Hardware Error-Free Transformations [18] . . . 10

1.20 Floating Point Unit from JOP [19] . . . 10

1.21 GHDL [20] . . . 11

1.22 On the design of fast IEEE floating-point adders [21] . . . 11

1.23 Re: round32 (round64 (X)) ?= round32 (X) [22] . . . 11

1.24 Basic Issues in Floating Point Arithmetic and Error Analysis [23] . . . 12

1.25 Charles Babbage [24, 25] . . . 12

1.26 GPU Floating-Point Paranoia [**?**] . . . 12

1.27 A floating-point technique for extending the available precision [26] . . . 12

1.28 Handbook of Floating-Point Arithmetic [27] . . . 12
1.28.1 A sequence that seems to converge to a wrong limit - pgs 9-10, [27] . . . 12
1.28.2 Mr Gullible and the Chaotic Bank Society pgs 10-11 [27] . . . 13
1.28.3 Rump's example pg 12 [27] . . . 13

2 General Notes **14**

2.1 Floating-Point [27, 7, 28, 29] . . . 14
2.2 Rounding Errors . . . 14
2.2.1 Results of calculatepi . . . 14

Chapter 1

Literature Summaries

1.1 Postscript Language Reference Manual [1]

Adobe's official reference manual for PostScript.

It is big.

1.2 Portable Document Format Reference Manual [2]

Adobe's official reference for PDF.

It is also big.

1.3 IEEE Standard for Floating-Point Arithmetic [3]

The IEEE (revised) 754 standard.

It is also big.

1.4 Portable Document Format (PDF) — Finally... [4]

This is not spectacularly useful, is basically an advertisement for Adobe software.

Intro

- Visual communications has been revolutionised by computing
- BUT there have always been problems in exchanging formats
- Filetypes like text, rich text, IGES, DXF, TIFF, JPEG, GIFF solve problems for particular types of files only
- PDF solves everything for everyone; can include text, images, animation, sound, etc

PDF Features

- Raster Image Process (RIP) — For printing (presumably also displaying on screen)
- Originally needed to convert to PS then RIP, with PS 3 can now RIP directly.
- Reduced filesize due to compression
- Four major applications - Stoy 1999[**?**]
  1. Download files from internet
  2. Files on CDs
  3. Files for outputing to printers
  4. Conventional [commercial scale?] printing
- List of various (Adobe) PDF related software
  - Includes software for PS that converts to/from PDF
  - So PS was obviously pretty popular before PDF
- Can Optimize for screen/printer [not clear how]
- Can compress for size

1.5 Pixels or Perish [5]

"The art of scientific illustration will have to adapt to the new age of online publishing" And therefore, JavaScript libraries (D^3) are the future.

The point is that we need to change from thinking about documents as paper to thinking of them as pixels. This kind of makes it related to our paper, because it is the same way we are justifying our project. It does mention precision, but doesn't say we need to get more of it.

I get the feeling from this that Web based documents are a whole bunch of completely different design philosophies hacked together with JavaScript.

This paper uses Metaphors a lot. I never met a phor that didn't over extend itself.

Intro

- Drawings/Pictures are ornaments in science but they are not just ornamental
- Processes have changed a lot; eg: photographic plates → digital images
- "we are about to turn the page — if not close the book — on yet another chapter in publishing history." (HO HO HO)
- It would be cool to have animated figures in documents (eg: Population pyramid; changes with time); not just as "supplements"
- In the beginning, there was PostScript, 1970s and 1980s, John Warnock and Charles Geschke, Adobe Systems
- PS is a language for vector graphics; objects are constructed from geometric primitives rather than a discrete array of pixels
- PS is a complete programming language; an image is also a program; can exploit this to control how images are created based on data (eg: Faces)
- PDF is "flattened" PS. No longer programable. Aspires to be "virtual paper".
- But why are we using such powerful computing machines just to emulate sheets paper? (the author asks)

Web based Documents

- HTML, CSS, JavaScript - The Axis of Web Documents
  - HTML - Defines document structure
  - CSS - Defines presentation of elements in document
  - JavaScript - Encodes actions, allows dynamic content (change the HTML/CSS)
- `<canvas>` will let you draw anything (So in principle don't even need all of HTML/CSS)
  - Not device independent
  - "Coordinates can be specified with precision finer than pixel resolution" **(TODO: Investigate this?)**
  - JavaScript operators to draw things on canvas are very similar to the PostScript model
- SVG — Same structure (Document Object Model (DOM)) as HTML
  - "Noun language"
  - Nouns define lines/curves etc, rather than paragraphs/lists
  - Also borrows things from PostScript (eg: line caps and joints)
  - IS device independent, "very high precision" **(TODO: Investigate)**
  - JavaScript can be used to interact with SVG too
- D^3 (Data Driven Documents) - A JavaScript library
  - Idea is to create or modify elements of a DOM document using supplied data
  - `https://github.com/mbostock/d3/wiki`
- We are in a new Golden Age of data visualisation
- Why do we still use PDFs?
  - PDFs are "owned" by the author/reader; you download it, store it, you can print it, etc
  - HTML documents are normally on websites. They are not self contained. They often rely on remote content from other websites (annoying to download the whole document).
- **Conclusion** Someone should open up PDF to accept things like D^3 and other graphics formats (links nicely with [6])
- Also, Harry Potter reference

1.6 Embedding and Publishing Interactive, 3D Figures in PDF Files [6]

- Linkes well with [5]; I heard you liked figures so I put a figure in your PDF
- Title pretty much summarises it; similar to [5] except these guys actually did something practical

1.7 27 Bits are not enough for 8 digit accuracy [?]

Proves with maths, that rounding errors mean that you need at least q bits for p decimal digits. $10^p < 2^{q-1}$

- Eg: For 8 decimal digits, since $10^8 < 2^{27}$ would expect to be able to represent with 27 binary digits
- But: Integer part requires digits bits (regardless of fixed or floating point represenetation)
- Trade-off between precision and range
  - $9000000.0 \rightarrow 9999999.9$ needs 24 digits for the integer part $2^{23} = 83886008$
- Floating point zero = smallest possible machine exponent
- Floating point representation:

$$y = 0.y_1y_2...y_q \times 2^n$$

- Can eliminate a bit by considering whether $n = -e$ for $-e$ the smallest machine exponent (???)
  - Get very small numbers with the same precision
  - Get large numbers with the extra bit of precision

1.8 What every computer scientist should know about floating-point arithmetic [7]

- Book: *Floating Point Computation* by Pat Sterbenz (out of print... in 1991)
- IEEE floating point standard becoming popular (introduced in 1987, this is 1991)
  - As well as structure, defines the algorithms for addition, multiplication, division and square root
  - Makes things portable because results of operations are the same on all machines (following the standard)
  - Alternatives to floating point: Floating slasi and Signed Logarithm (TODO: Look at these, although they will probably not be useful)
- Base β and precision p (number of digits to represent with) - powers of the base can be represented exactly.
- Largest and smallest exponents e_{min} and e_{max}
- Need bits for exponent and fraction, plus one for sign
- "Floating point number" is one that can be represented exactly.
- Representations are not unique! $0.01 \times 10^1 = 1.00 \times 10^{-1}$ Leading digit of one $\implies$ "normalised"
- Requiring the representation to be normalised makes it unique, **but means it is impossible to represent zero**.
  - Represent zero as $1 \times \beta^{e_{min}-1}$ - requires extra bit in the exponent
- **Rounding Error**
  - "Units in the last place" eg: 0.0314159 compared to 0.0314 has ulp error of 0.159
  - If calculation is the nearest floating point number to the result, it will still be as much as 1/2 ulp in error
  - Relative error corresponding to 1/2 ulp can vary by a factor of β "wobble". Written in terms of ϵ
  - Maths $\implies$ **Relative error is always bounded by** $\epsilon = (\beta/2)\beta^{-p}$
  - Fixed relative error $\implies$ ulp can vary by a factor of β . Vice versa
  - Larger $\beta \implies$ larger errors
- **Guard Digits**
  - In subtraction: Could compute exact difference and then round; this is expensive

  - Keep fixed number of digits but shift operand right; discard precision. Lead to relative error up to $\beta - 1$
  - Guard digit: Add extra digits before truncating. Leads to relative error of less than 2ϵ. This also applies to addition
- **Catastrophic Cancellation** - Operands are subject to rounding errors - multiplication
- **Benign Cancellation** - Subtractions. Error $< 2\epsilon$
- Rearrange formula to avoid catastrophic cancellation
- Historical interest only - speculation on why IBM used $\beta = 16$ for the system/370 - increased range? Avoids shifting
- Precision: IEEE defines extended precision (a lower bound only)
- Discussion of the IEEE standard for operations (TODO: Go over in more detail)
- NaN allow continuing with underflow and Infinity with overflow
- "Incidentally, some people think that the solution to such anomalies is never to compare floating-point numbers for equality but instead to consider them equal if they are within some error bound E. This is hardly a cure all, because it raises as many questions as it answers." - On equality of floating point numbers

1.9 Compositing Digital Images [8]

Perter and Duff's classic paper "Compositing Digital Images" lays the foundation for digital compositing today. By providing an "alpha channel," images of arbitrary shapes and images with soft edges or sub-pixel coverage information can be overlayed digitally, allowing separate objects to be rasterized separately without a loss in quality.

Pixels in digital images are usually represented as 3-tuples containing (red component, green component, blue component). Nominally these values are in the [0-1] range. In the Porter-Duff paper, pixels are stored as (R, G, B, α) 4-tuples, where alpha is the fractional coverage of each pixel. If the image only covers half of a given pixel, for example, its alpha value would be 0.5.

To improve compositing performance, albeit at a possible loss of precision in some implementations, the red, green and blue channels are premultiplied by the alpha channel. This also simplifies the resulting arithmetic by having the colour channels and alpha channels use the same compositing equations.

Several binary compositing operations are defined:

- over
- in
- out
- atop
- xor
- plus

The paper further provides some additional operations for implementing fades and dissolves, as well as for changing the opacity of individual elements in a scene.

The method outlined in this paper is still the standard system for compositing and is implemented almost exactly by modern graphics APIs such as `OpenGL`. It is all but guaranteed that this is the method we will be using for compositing document elements in our project.

1.10 Bresenham's Algorithm: Algorithm for computer control of a digital plotter [9]

Bresenham's line drawing algorithm is a fast, high quality line rasterization algorithm which is still the basis for most (aliased) line drawing today. The paper, while originally written to describe how to control a particular plotter, is uniquely suited to rasterizing lines for display on a pixel grid.

Lines drawn with Bresenham's algorithm must begin and end at integer pixel coordinates, though one can round or truncate the fractional part. In order to avoid multiplication or division in the algorithm's inner loop,

The algorithm works by scanning along the long axis of the line, moving along the short axis when the error along that axis exceeds 0.5px. Because error accumulates linearly, this can be achieved by simply adding the per-pixel error (equal to (short axis/long axis)) until it exceeds 0.5, then incrementing the position along the short axis and subtracting 1 from the error accumulator.

As this requires nothing but addition, it is very fast, particularly on the older CPUs used in Bresenham's time. Modern graphics systems will often use Wu's line-drawing algorithm instead, as it produces antialiased lines, taking sub-pixel coverage into account. Bresenham himself extended this algorithm to produce Bresenham's circle algorithm. The principles behind the algorithm have also been used to rasterize other shapes, including Bézier curves.

1.11 Quad Trees: A Data Structure for Retrieval on Composite Keys [10]

This paper introduces the "quadtree" spatial data structure. The quadtree structure is a search tree in which every node has four children representing the north-east, north-west, south-east and south-west quadrants of its space.

1.12 Xr: Cross-device Rendering for Vector Graphics [11]

Xr (now known as Cairo) is an implementation of the PDF v1.4 rendering model, independent of the PDF or PostScript file formats, and is now widely used as a rendering API. In this paper, Worth and Packard describe the PDF v1.4 rendering model, and their PostScript-derived API for it.

The PDF v1.4 rendering model is based on the original PostScript model, based around a set of *paths* (and other objects, such as raster images) each made up of lines and Bézier curves, which are transformed by the "Current Transformation Matrix." Paths can be *filled* in a number of ways, allowing for different handling of self-intersecting paths, or can have their outlines *stroked*. Furthermore, paths can be painted with RGB colours and/or patterns derived from either previously rendered objects or external raster images. PDF v1.4 extends this to provide, amongst other features, support for layering paths and objects using Porter-Duff compositing[8], giving each painted path the option of having an α value and a choice of any of the Porter-Duff compositing methods.

The Cairo library approximates the rendering of some objects (particularly curved objects such as splines) with a set of polygons. An `XrSetTolerance` function allows the user of the library to set an upper bound on the approximation error in fractions of device pixels, providing a trade-off between rendering quality and performance. The library developers found that setting the tolerance to greater than 0.1 device pixels resulted in errors visible to the user.

1.13 Glitz: Hardware Accelerated Image Compositing using OpenGL [12]

This paper describes the implementation of an `OpenGL` based rendering backend for the `Cairo` library.

The paper describes how OpenGL's Porter-Duff compositing is easily suited to the Cairo/PDF v1.4 rendering model. Similarly, traditional OpenGL (pre-version 3.0 core) support a matrix stack of the same form as Cairo.

The "Glitz" backend will emulate support for tiled, non-power-of-two patterns/textures if the hardware does not support it.

Glitz can render both triangles and trapezoids (which are formed from pairs of triangles). However, it cannot guarantee that the rasterization is pixel-precise, as OpenGL does not proveide this consistently.

Glitz also supports multi-sample anti-aliasing, convolution filters for raster image reads (implemented with shaders).

Performance was much improved over the software rasterization and over XRender accellerated rendering on all except nVidia hardware. However, nVidia's XRender implementation did slow down significantly when some transformations were applied.

1.14 Boost Multiprecision Library [13]

- "The Multiprecision Library provides integer, rational and floating-point types in C++ that have more range and precision than C++'s ordinary built-in types."
- Specify number of digits for precision as a template argument.
- Precision is fixed... **possible approach to project:** Use `boost::mpf_float<N>` and increase `N` as more precision is required?

1.15 A CMOS Floating Point Unit [14]

The paper describes the implentation of a FPU for PowerPC using a particular Hewlett Packard process (HP14B 0.5μm, 3M, 3.3V). It implements a "subset of the most commonly used double precision floating point instructions". The unimplemented operations are compiled for the CPU.

The paper gives a description of the architecture and design methods. This appears to be an entry to a student design competition.

Standard is IEEE 754, but the multiplier tree is a 64-bit tree instead of a 54 bit tree. " The primary reason for implementing a larger tree is for future additions of SIMD [Single Instruction Multiple Data (?)] instructions similar to Intel's MMX and Sun's VIS instructions".

HSPICE simulations used to determine transistor sizing.

Paper has a block diagram that sort of vaguely makes sense to me. The rest requires more background knowledge.

1.16 Simply FPU[15]

This is a webpage at one degree of seperation from wikipedia.

It talks about FPU internals, but mostly focuses on the instruction sets. It includes FPU assembly code examples (!)

It is probably not that useful, I don't think we'll end up writing FPU assembly?

FPU's typically have 80 bit registers so they can support REAL4, REAL8 and REAL10 (single, double, extended precision).

1.17 Floating Point Package User's Guide [16]

This is a technical report describing floating point VHDL packages `http://www.vhdl.org/fphdl/vhdl.html`

In theory I know VHDL (cough) so I am interested in looking at this further to see how FPU hardware works. It might be getting a bit sidetracked from the "document formats" scope though.

The report does talk briefly about the IEEE standard and normalised / denormalised numbers as well.

See also: Java Optimized Processor[19] (it has a VHDL implementation of a FPU).

1.18 Low-Cost Microarchitectural Support for Improved Floating-Point Accuracy[17]

Mentions how GPUs offer very good floating point performance but only for single precision floats. (NOTE: This statement seems to contradict [**?**].

Has a diagram of a Floating Point adder.

Talks about some magical technique called "Native-pair Arithmetic" that somehow makes 32-bit floating point accuracy "competitive" with 64-bit floating point numbers.

1.19 Accurate Floating Point Arithmetic through Hardware Error-Free Transformations [18]

From the abstract: "This paper presents a hardware approach to performing ac- curate floating point addition and multiplication using the idea of error- free transformations. Specialized iterative algorithms are implemented for computing arbitrarily accurate sums and dot products."

The references for this look useful.

It also mentions VHDL.

So whenever hardware papers come up, VHDL gets involved... I guess it's time to try and work out how to use the Opensource VHDL implementations.

This is about reduction of error in hardware operations rather than the precision or range of floats. But it is probably still relevant.

This has the Fast2Sum algorithm but for the love of god I cannot see how you can compute anything other than $a + b = 0 \forall a, b$ using the algorithm as written in their paper. It references Dekker[26] and Kahn; will look at them instead.

1.20 Floating Point Unit from JOP [19]

This is a 32 bit floating point unit developed for JOP in VHDL. I have been able to successfully compile it and the test program using GHDL[20].

Whilst there are constants (eg: `FP_WIDTH = 32, EXP_WIDTH = 8, FRAC_WIDTH = 23`) defined, the actual implementation mostly uses magic numbers, so some investigation is needed into what, for example, the "52" bits used in the sqrt units are for.

1.21 GHDL [20]

GHDL is an open source GPL licensed VHDL compiler written in Ada. It had packages in debian up until wheezy when it was removed. However the sourceforge site still provides a `deb` file for wheezy.

This reference explains how to use the `ghdl` compiler, but not the VHDL language itself.

GHDL is capable of compiling a "testbench" - essentially an executable which simulates the design and ensures it meets test conditions. A common technique is using a text file to provide the inputs/outputs of the test. The testbench executable can be supplied an argument to save a `vcd` file which can be viewed in `gtkwave` to see timing diagrams.

Sam has successfully compiled the VHDL design for an FPU in JOP[19] into a "testbench" executable which uses standard i/o instead of a regular file. Using unix domain sockets we can execute the FPU as a child process and communicate with it from our document viewing test software. This means we can potentially simulate alternate hardware designs for FPUs and witness the effect they will have on precision in the document viewer.

Using `ghdl` the testbench can also be linked as part a C/C++ program and run using a function; however there is still no way to communicate with it other than forking a child process and using a unix domain socket anyway. Also, compiling the VHDL FPU as part of our document viewer would clutter the code repository and probably be highly unportable. The VHDL FPU has been given a seperate repository.

1.22 On the design of fast IEEE floating-point adders [21]

This paper gives an overview of the "Naive" floating point addition/subtraction algorithm and gives several optimisation techniques:

TODO: Actually understand these...

- Use parallel paths (based on exponent)
- Unification of significand result ranges
- Reduction of IEEE rounding modes
- Sign-magnitude computation of a difference
- Compound Addition
- Approximate counting of leading zeroes
- Pre-computation of post-normalization shift

They then give an implementation that uses these optimisation techniques including very scary looking block diagrams.

They simulated the FPU. Does not mention what simulation method was used directly, but cites another paper (TODO: Look at this. I bet it was VHDL).

The paper concludes by summarising the optimisation techniques used by various adders in production (in 2001).

This paper does not specifically mention the precision of the operations, but may be useful because a faster adder design might mean you can increase the precision.

1.23 Re: round32 (round64 (X)) ?= round32 (X) [22]

I included this just for the quote by Nelson H. F. Beebe:

"It is too late now to repair the mistakes of the past that are present in millions of installed systems, but it is good to know that careful research before designing hardware can be helpful."

This is in regards to the problem of double rounding. It provides a reference for a paper that discusses a rounding mode that eliminates the problem, and a software implementation.

It shows that the IEEE standard can be fallible!

Not sure how to work this into our literature review though.

1.24 Basic Issues in Floating Point Arithmetic and Error Analysis [23]

These are lecture notes from U.C Berkelye CS267 in 1996.

1.25 Charles Babbage [24, 25]

Tributes to Charles Babbage. Might be interesting for historical background. Don't mention anything about floating point numbers though.

1.26 GPU Floating-Point Paranoia [?]

This paper discusses floating point representations on GPUs. They have reproduced the program *Paranoia* by William Kahan for characterising floating point behaviour of computers (pre IEEE) for GPUs.

There are a few remarks about GPU vendors not being very open about what they do or do not do with

Unfortunately we only have the extended abstract, but a pretty good summary of the paper (written by the authors) is at: `www.cs.unc.edu/~ibr/projects/paranoia/`

From the abstract:

"... [GPUs are often similar to IEEE] However, we have found that GPUs do not adhere to IEEE standards for floating-point op- erations, nor do they give the information necessary to establish bounds on error for these operations ... "

and "...Our goal is to determine the error bounds on floating-point op- eration results for quickly evolving graphics systems. We have cre- ated a tool to measure the error for four basic floating-point opera- tions: addition, subtraction, multiplication and division."

The implementation is only for windows and uses glut and glew and things. Implement our own version?

1.27 A floating-point technique for extending the available precision [26]

This is Dekker's formalisation of the Fast2Sum algorithm originally implemented by Kahn.

$$\begin{aligned} z &= \mathrm{RN}(x+y) \\ w &= \mathrm{RN}(z-x) \\ zz &= \mathrm{RN}(y-w) \\ \implies z+zz &= x+y \end{aligned}$$

There is a version for multiplication.

I'm still not quite sure when this is useful. I haven't been able to find an example for x and y where $x+y \neq \mathrm{Fast2Sum}(x,y)$.

1.28 Handbook of Floating-Point Arithmetic [27]

This book is amazingly useful and pretty much says everything there is to know about Floating Point numbers. It is much easier to read than Goldberg or Priest's papers.

I'm going to start working through it and compile their test programs.

1.28.1 A sequence that seems to converge to a wrong limit - pgs 9-10, [27]

$$u_n = \begin{cases} u_0 = 2 \\ u_1 = -4 \\ u_n = 111 - \frac{1130}{u_{n-1}} + \frac{3000}{u_{n-1}u_{n-2}} \end{cases}$$

The limit of the series should be 6 but when calculated with IEEE floats it is actually 100 The authors show that the limit is actually 100 for different starting values, and the error in floating point arithmetic causes the series to go to that limit instead.

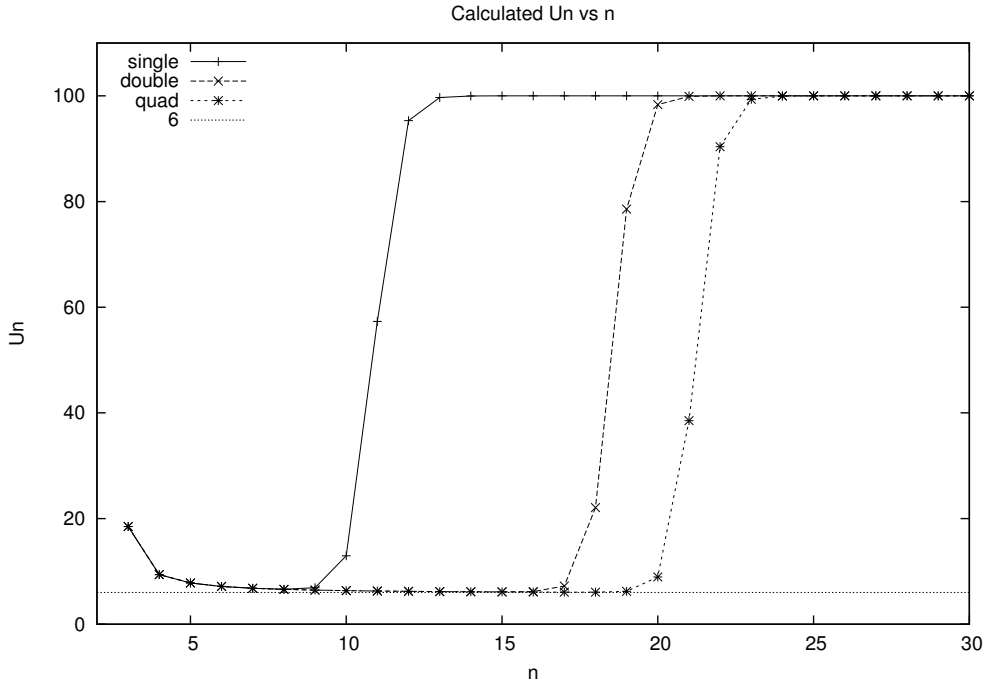

Figure 1.1: Output of Program 1.1 from *Handbook of Floating-Point Arithmetic*[27] for various IEEE types

1.28.2 Mr Gullible and the Chaotic Bank Society pgs 10-11 [27]

This is an example of a sequence involving e. Since e cannot be represented exactly with FP, even though the sequence should go to 0 for $a_0 = e - 1$, the representation of $a_0 \neq e - 1$ so the sequence goes to $\pm\infty$.

To eliminate these types of problems we'd need an *exact* representation of all real numbers. For *any* FP representation, regardless of precision (a finite number of digits) there will be numbers that can't be represented exactly hence you could find a similar sequence that would explode.

IE: The more precise the representation, the slower things go wrong, but they still go wrong, **even with errorless operations**.

1.28.3 Rump's example pg 12 [27]

This is an example where the calculation of a function $f(a, b)$ is not only totally wrong, it gives completely different results depending on the CPU. Despite the CPU conforming to IEEE.

Chapter 2

General Notes

2.1 Floating-Point [27, 7, 28, 29]

A set of FP numbers is characterised by:

1. Radix (base) $\beta \geq 2$
2. Precision
3. Two "extremal" exponents $e_min < 0 < e_max$ (generally, don't have to have the 0 in there)

Numbers are represented by **integers**: (M, e) such that $x = M \times \beta^{e-p+1}$

Require: $|M| \leq \beta^p - 1$ and $e_min \leq e \leq e_max$.

Representations are not unique; set of equivelant representations is a cohort.

β^{e-p+1} is the quantum, $e - p + 1$ is the quantum exponent.

Alternate represetnation: (s, m, e) such that $x = (-1)^s \times m \times \beta^e$ m is the significand, mantissa, or fractional part. Depending on what you read.

2.2 Rounding Errors

They happen. There is ULP and I don't mean a political party.

TODO: Probably say something more insightful. Other than "here is a graph that shows errors and we blame rounding".

2.2.1 Results of calculatepi

We can calculate pi by numerically solving the integral:

$$\int_0^1 \left(\frac{4}{1+x^2}\right) dx = \pi$$

Results with Simpson Method:

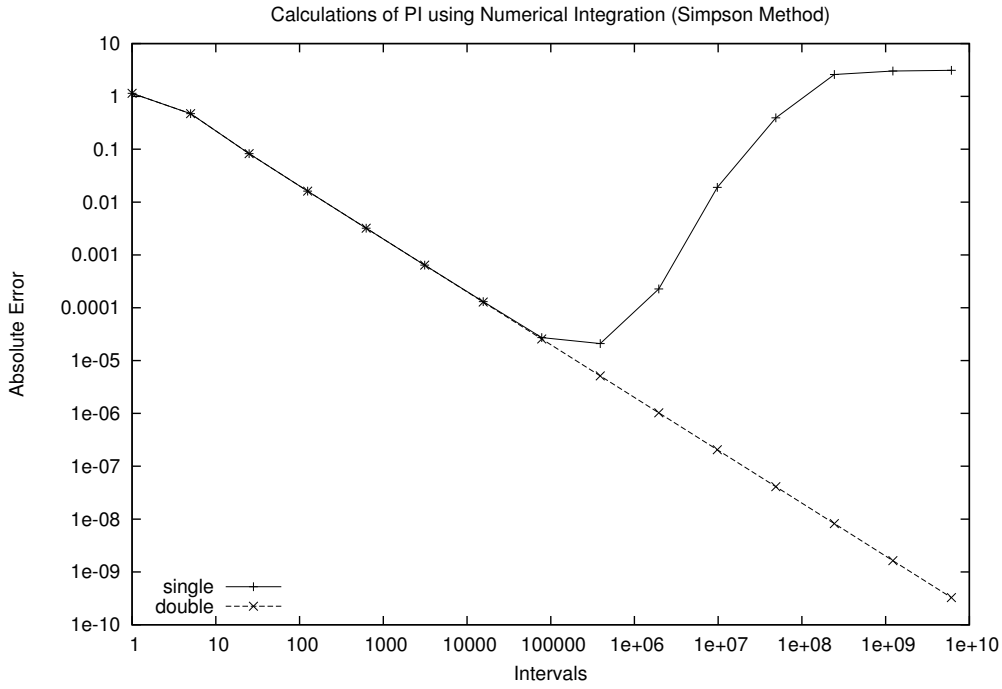

Figure 2.1: Example of accumulated rounding errors in a numerical calculation

Tests with `calculatepi` show it's not quite as simple as just blindly replacing all your additions with Fast2Sum from Dekker[26]. ie: The graph looks exactly the same for single precision. `calculatepi` obviously also has multiplication ops in it which I didn't change. Will look at after sleep maybe.

Bibliography

[1] Adobe Systems Incorporated. *PostScript Language Reference*. Addison-Wesley Publishing Company, 3rd edition, 1985 - 1999.

[2] Adobe Systems Incorporated. *PDF Reference*. Adobe Systems Incorporated, 6th edition, 2006.

[3] Ieee standard for floating-point arithmetic. *IEEE Std 754-2008*, pages 1–70, Aug 2008.

[4] Michael A. Wan-Lee Cheng. Portable document format (pdf) – finally, a universal document exchange technology. *Journal of Technology Studies*, 28(1):59 – 63, 2002.

[5] Brian Hayes. Pixels or perish. *American Scientist*, 100(2):106 – 111, 2012.

[6] David G. Barnes, Michail Vidiassov, Bernhard Ruthensteiner, Christopher J. Fluke, Michelle R. Quayle, and Colin R. McHenry. Embedding and publishing interactive, 3-dimensional, scientific figures in portable document format (pdf) files. *PLoS ONE*, 8(9):1 – 15, 2013.

[7] David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Comput. Surv.*, 23(1):5–48, March 1991.

[8] Thomas Porter and Tom Duff. Compositing digital images. In *ACM Siggraph Computer Graphics*, volume 18, pages 253–259. ACM, 1984.

[9] Jack E Bresenham. Algorithm for computer control of a digital plotter. *IBM Systems journal*, 4(1):25–30, 1965.

[10] Raphael A. Finkel and Jon Louis Bentley. Quad trees a data structure for retrieval on composite keys. *Acta informatica*, 4(1):1–9, 1974.

[11] Carl Worth and Keith Packard. Xr: Cross-device rendering for vector graphics. In *Linux Symposium*, page 480, 2003.

[12] Peter Nilsson and David Reveman. Glitz: Hardware accelerated image compositing using OpenGL. In *USENIX Annual Technical Conference, FREENIX Track*, pages 29–40, 2004.

[13] John Maddock and Christopher Kormanyos. Boost multiprecision library. `http://www.boost.org/doc/libs/1_53_0/libs/multiprecision/doc/html/boost_multiprecision/`.

[14] Michael J. Kelley, Matthew A. Postiff, Advisor Richard, and B. Brown. A cmos floating point unit, 1997.

[15] Raymond Filiatreault. Simply fpu. `http://www.website.masmforum.com/tutorials/fptute/index.html`, 2003.

[16] David Bishop. Floating point package user's guide. *EDA Industry Working Groups*, 2008. Technical Report.

[17] William R. Dieter, Akil Kaveti, and Henry G. Dietz. Low-cost microarchitectural support for improved floating-point accuracy. *IEEE Comput. Archit. Lett.*, 6(1):13–16, January 2007.

[18] Edin Kadric, Paul Gurniak, and André DeHon. Accurate parallel floating-point accumulation. In *Computer Arithmetic (ARITH), 2013 21st IEEE Symposium on*, pages 153–162. IEEE, 2013.

[19] jop devel. Java optimized processor. `https://github.com/jop-devel/jop`.

[20] Tristan Gingold. Ghdl guide. `http://ghdl.free.fr/ghdl/`, 2007.

[21] P.-M. Seidel and G. Even. On the design of fast ieee floating-point adders. In *Computer Arithmetic, 2001. Proceedings. 15th IEEE Symposium on*, pages 184–194, 2001.

[22] Nelson H. F. Beebe. Re: round32 (round64 (x)) ?= round32 (x). `http://grouper.ieee.org/groups/754/email/msg04169.html`. IEEE 754 Working Group Mail Archives.

[23] Jim Demmel. Basic issues in floating point arithmetic and error analysis. *U.C. Berkeley CS267*. Lecture Notes.

[24] N. S. Dodge. Charles babbage. *Annals of the History of Computing, IEEE*, 22(4):22–43, Oct 2000.

[25] Unknown Author. Charles babbage. *Nature*, 5(106):28–29, 1871.

[26] T.J. Dekker. A floating-point technique for extending the available precision. *Numerische Mathematik*, 18(3):224–242, 1971.

[27] Jean-Michel Muller, Nicolas Brisebarre, Florent de Dinechin, Claude-Pierre Jeannerod, Vincent Lefèvre, Guillaume Melquiond, Nathalie Revol, Damien Stehlé, and Serge Torres. *Handbook of Floating-Point Arithmetic*. Birkhäuser Boston Inc., Cambridge, MA, USA, 2010.

[28] David Goldberg. The design of floating-point data types. *ACM Lett. Program. Lang. Syst.*, 1(2):138–151, June 1992.

[29] D.M. Priest. Algorithms for arbitrary precision floating point arithmetic. In *Computer Arithmetic, 1991. Proceedings., 10th IEEE Symposium on*, pages 132–143, Jun 1991.