

Tekum: Balanced Ternary Tapered Precision Real Arithmetic

Laslo Hunhold 

Parallel and Distributed Systems Group
University of Cologne, Cologne, Germany
`hunhold@uni-koeln.de`

Abstract In light of recent hardware advances, it is striking that real arithmetic in balanced ternary logic has received almost no attention in the literature. This is particularly surprising given ternary logic’s promising properties, which could open new avenues for energy-efficient computing and offer novel strategies for overcoming the memory wall. This paper revisits the concept of tapered precision arithmetic, as used in posit and takum formats, and introduces a new scheme for balanced ternary logic: tekum arithmetic. Several fundamental design challenges are addressed along the way. The proposed format is evaluated and shown to exhibit highly promising characteristics. In many respects, it outperforms both posits and takums. As ternary hardware matures, this work represents a crucial step toward unlocking the full potential of real-number computation in ternary systems, laying the groundwork for a new class of number formats designed from the ground up for a new category of next-generation hardware.

Keywords: tekum arithmetic · balanced ternary logic · tapered precision · real arithmetic · floating-point arithmetic · posit arithmetic · takum arithmetic

1 Introduction

Computer science is a comparatively young scientific discipline. The first programmable computers were constructed little more than eighty years ago, ZUSE’s Z3 in 1941 and the ENIAC by MAUCHLY and ECKERT in 1946. These early machines were based on the binary system, which has since shaped the foundations of computer science. The prevailing perception of computers is that, at their core, they operate through states of ‘on’ or ‘off’. Decades of research and development have refined and optimised the design of binary computers.

History, however, can narrow perspective. A lesser-known paradigm of computation is ternary logic. Instead of two states, ternary logic uses three; instead of binary digits (bits), one has ternary digits (trits), where a single trit contains $\log_2(3) \approx 1.58$ bits of information. The motivation for considering such systems lies in the concept of *radix economy*, the cost of representing a positive integer

$n \in \mathbb{N}_1$ in a given base $b \in \mathbb{R}_+$. This cost is defined as $C(b, n) := \lfloor \log_b(n) + 1 \rfloor \cdot b$, namely the number of digits required to represent n in base b , multiplied by the base b itself [11]. Intuitively, larger bases produce shorter representations but incur greater hardware complexity, whereas smaller bases ease implementation but require more digits. It can be shown analytically that base $e \approx 2.718$ minimises cost asymptotically, with 3 being the nearest integer. If one assumes that base size correlates with hardware complexity, ternary arithmetic therefore offers the most balanced trade-off between circuit complexity and representational efficiency. By contrast, binary arithmetic simplifies circuit design, but at the expense of higher representational cost and greater communication overhead.

Hardware investigations have compared binary and ternary implementations. While ternary circuits achieve comparable computational speed, they typically require more logic; for example, a ternary adder needs approximately 62% more circuitry, even accounting for the higher information density of trits [15]. This finding has often been cited as a reason to dismiss ternary logic. Yet, such dismissal warrants reconsideration in light of modern computing trends: the dominant bottleneck today is not raw computational speed, but memory bandwidth [2, 5]. Under these conditions, ternary logic acquires renewed relevance. Recent advances, such as ternary deep neural networks and carbon nanotube transistors, which operate natively in ternary, further strengthen its value proposition.

The full potential of ternary computation emerges when digits are chosen from $\{-1, 0, 1\}$ rather than $\{0, 1, 2\}$. This *balanced ternary* system is unique to odd bases and may appear unfamiliar, given the dominance of even bases such as 2, 8, 10, 16, and 64, or historically base 60. In balanced ternary, the digit -1 is commonly denoted by a special symbol, here written as T. For instance, 2 is represented as 1T ($1 \cdot 3 + (-1) \cdot 1$), and -4 as TT ($(-1) \cdot 3 + (-1) \cdot 1$). This representation has several appealing properties: integers are inherently signed, and perfectly symmetric around zero, without unlike two’s complement binary integers. Negation is achieved simply by inverting digits, and rounding is the same as truncation, eliminating the need for carries. KNUTH has described it as ‘perhaps the prettiest number system of all’ [16, p. 207].

Balanced ternary has been studied mainly in the context of integers. Real-number representations, such as floating-point formats, remain surprisingly underexplored. One might expect that properties such as rounding by truncation could mitigate issues inherent in binary floating-point arithmetic, where carries complicate rounding. Yet the literature is sparse. The only notable proposal is **ternary27** by O’HARE [21], which is heavily inspired by IEEE 754. Although it achieves a favourable dynamic range, its design is inefficient, wasting many representations. This inefficiency is more detrimental in ternary than in binary arithmetic, as each trit carries more information than a bit. While the IEEE 854-1987 standard specifies radix-independent floating-point numbers, it does not address balanced bases.

Given current hardware developments, it is undesirable for theoretical progress to lag behind practical advances. This paper therefore revisits balanced ternary real arithmetic and makes the following contributions. After introducing

the fundamentals of balanced ternary in Section 2, we identify key design challenges in Section 3 that up to the design of takums [12] were insurmountable. We then propose a new balanced ternary number format, *tekum*, in Section 4, and evaluate its properties in Section 5.1, before drawing conclusions in Section 6.

2 Balanced Ternary

In this section we introduce balanced ternary and all the tools necessary to define and analyse the subsequently introduced tekum arithmetic. We follow a formal approach given the unintuitiveness compared to standard binary logic. Additionally, this paper is more or less an inaugural paper on balanced ternary real arithmetic, and it's helpful to suggest a notation for this nascent field.

Definition 1 (balanced ternary strings). *Let $n \in \mathbb{N}_0$. The set of n -trit balanced ternary strings is defined as $\mathbb{T}_n := \{\text{T}, 0, 1\}^n$ with $\text{T} := -1$. By convention $\mathbb{T}_n \ni \mathbf{t} := (t_{n-1}, \dots, t_0) =: t_{n-1} \cdots t_0$.*

The first thing we notice is that, compared to posits and takums, we denote ternary strings with bold lowercase letters instead of uppercase letters. This is more in line with the common notation for vectors in mathematics and computer science.

Definition 2 (concatenation). *Let $m, n \in \mathbb{N}_0$, $\mathbf{t} \in \mathbb{T}_m$ and $\mathbf{u} \in \mathbb{T}_n$. The concatenation of \mathbf{t} and \mathbf{u} is defined as $\mathbf{t} \# \mathbf{u} := (t_{m-1}, \dots, t_0, u_{n-1}, \dots, u_0) \in \mathbb{T}_{m+n}$.*

This concatenation operator is distinct from the more relaxed, overloaded concatenation notation with parentheses.

Definition 3 (integer mapping). *Let $n \in \mathbb{N}_0$. The integer mapping $\text{int}_n: \mathbb{T}_n \rightarrow \{-\frac{1}{2}(3^n - 1), \dots, \frac{1}{2}(3^n - 1)\}$ is defined as $\text{int}_n(\mathbf{t}) := \sum_{i=0}^{n-1} t_i 3^i$.*

Unlike with two's complement binary integers, there is no explicit sign bit. Instead the sign is indicated by the most significant non-zero trit, and there is no concept of an unsigned integer in balanced ternary. The range of integer values follows from the finite geometric series obtained with the extremal values $\text{int}_n(\text{T} \dots \text{T})$ and $\text{int}_n(1 \dots 1)$, and is perfectly symmetric around zero (unlike two's complement integers).

Definition 4 (negation). *Let $n \in \mathbb{N}_0$ and $\mathbf{t} \in \mathbb{T}_n$. The negation of \mathbf{t} is defined as $-\mathbf{t} := (-t_{n-1}, \dots, -t_0)$.*

By construction we can see that it holds $\text{int}_n(-\mathbf{t}) = -\text{int}_n(\mathbf{t})$, making it well-defined. The negation operation outlines another distinctive difference between two's complement binary and balanced ternary integers: Whereas to negate the underlying integral value you need to negate all bits and add one with the former, which is expensive in hardware due to the carry-chain, negation in balanced ternary is a very cheap, entrywise operation.

Definition 5 (addition and subtraction). Let $n \in \mathbb{N}_0$ and $\mathbf{t}, \mathbf{u} \in \mathbb{T}_n$. The addition of \mathbf{t} and \mathbf{u} is defined with $s := \text{int}_n(\mathbf{t}) + \text{int}_n(\mathbf{u})$ as

$$\mathbf{t} + \mathbf{u} := \text{int}_n^{\text{inv}} \left(\begin{cases} s + \frac{1}{2}(3^n + 1) & s < -\frac{1}{2}(3^n - 1) \\ s - \frac{1}{2}(3^n + 1) & s > \frac{1}{2}(3^n - 1) \\ s & \text{otherwise} \end{cases} \right). \quad (1)$$

The subtraction of \mathbf{t} and \mathbf{u} is defined as $\mathbf{t} - \mathbf{u} := \mathbf{t} + (-\mathbf{u})$.

Despite the complex appearance, the addition is merely defined here as a fixed-size operation, taking two n -trit inputs and yielding an n -trit output, discarding any carries that might occur due to over- or underflow and using the inverse integer mapping to obtain the ternary string. If carries should be taken into account, the inputs need to be extended to $n + 1$ -trit strings before the addition. Especially notable is the definition of the subtraction: Given negation is so cheap, you can define it in terms of the addition.

Definition 6 (modulus). Let $n \in \mathbb{N}_0$ and $\mathbf{t} \in \mathbb{T}_n$. The modulus of \mathbf{t} is defined as

$$|\mathbf{t}| := \begin{cases} -\mathbf{t} & \text{int}_n(\mathbf{t}) < 0 \\ \mathbf{t} & \text{int}_n(\mathbf{t}) \geq 0. \end{cases} \quad (2)$$

With these definitions in place we can proceed with the derivation of the tekum arithmetic.

3 The Three Filters

In this section, we derive the central contribution of this paper: a new ternary real arithmetic. As outlined in the introduction, the current state of the art in ternary real arithmetic is surprisingly underdeveloped. This may be attributed in part to the relatively small number of arithmetic designers worldwide, as well as the limited training and interest in ternary hardware to date. However, a closer examination of recently established design criteria for number formats reveals three possible consecutive filters that may have led previous researchers to attempt, and ultimately abandon, the pursuit of a new format.

The term filter is borrowed from Robert HANSON's notion of the 'Great Filter', which refers to the fundamental developmental barriers a civilisation must overcome, proposed to explain the (apparent) absence of advanced extra-terrestrial life in the universe [9]. Here, we adapt the term 'filter' to describe conceptual or technical barriers that have likely hindered the development of a proper ternary real arithmetic. We outline and reflect on these filters throughout the derivation, highlighting how each is encountered and overcome in the formulation of our format.

Fundamentally, the design of a real number format involves defining a mapping between discrete representations (in this case, trit strings) and a corresponding set of real values. This set is constructed as a subset of the real numbers \mathbb{R} , often extended with additional non-number representations. The most

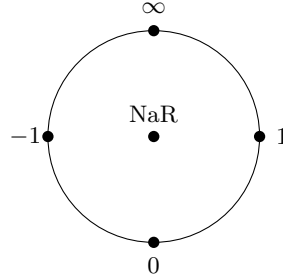


Figure 1: A visualisation of the real wheel algebra with the customary positioning of the bottom element NaR in the center as a ‘wheel hub’ that distinguishes it from the projectively extended real numbers.

common such extension is the *affinely extended real numbers*, $\mathbb{R} \cup \{-\infty, +\infty\}$. A less common alternative is the *projectively extended real numbers*, $\mathbb{R} \cup \{\infty\}$, where division by zero is well-defined.

Further extensions typically include one or more special representations to denote undefined or invalid values. These are commonly referred to as NaN (not a number) or NaR (not a real), and are used to represent non-real complex numbers or domain violations (e.g. $\ln(-1)$). While IEEE 754 floating-point numbers are based on the affinely extended real numbers and have multiple NaN representations, formats such as posits and takums instead exclude explicit representations of infinity and use a single NaR for both infinities and non-real numbers.

This was not always the case for posits. Early versions of the posit format only included a single infinity non-number representation [7, 18]. However, this was later revised in the draft standard [8], which removed the infinity representation in favour of a single NaR. The rationale for this decision is that the encoding scheme only permits a single special non-numeric value, and the practical utility of being able to propagate NaR in computations to signal error conditions was deemed to outweigh the mathematical elegance of a single infinity. The *takum* format subsequently adopted this revised approach [12].

The most mathematically complete and elegant alternative would be a set of represented values that includes both ∞ and NaR. Such a structure does exist in the form of the *wheel algebra* applied to reals, $\mathbb{R} \cup \{\infty, \text{NaR}\}$ [4]. In this framework, one defines $\infty := 1/0$ and $\text{NaR} := 0/0$, where NaR is referred to as the ‘bottom element’. The wheel algebra supports multiplication and inversion across all its elements, offering a closed and consistent model for arithmetic. Figure 1 illustrates a common visualisation of the wheel algebra: a circle representing the extended real values, with NaR at the centre, resembling a wheel with a hub. The numbers 1 and -1 are explicitly included as fixed points under inversion (reflection of the circle across the x -axis), just as 0 and ∞ serve as fixed points under negation (reflection across the y -axis).

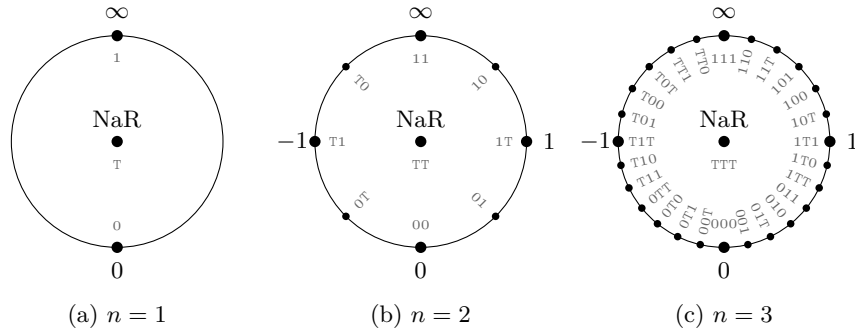


Figure 2: Mapping of ternary strings of lengths $n \in \{1, 2, 3\}$ to the real wheel algebra.

In light of the mathematical elegance of the real wheel algebra, one might naturally ask why formats such as posits and takums do not include both ∞ and NaR. If we consider discretising this set, we would ideally require an equal number of values in each quadrant of the wheel so that both negation and inversion map represented values back into the set. In other words, after assigning the five fundamental elements, NaR, -1 , 0 , 1 , and ∞ , the remaining number of unassigned states must be divisible by four.

In binary representations, this condition can never be satisfied. The total number of states is a power of two (and thus even), and subtracting the five fixed elements yields an odd number, which cannot be evenly divided by four. Ternary strings, on the other hand, have a total number of states equal to a power of three, which is always odd. Subtracting five from an odd number yields an even result, raising the question: could this make the construction viable?

We assign the bottom element, NaR, to the lexicographically smallest balanced ternary integer, $T \cdots T$, ensuring that it is also the smallest element under comparison. This approach is consistent with the total ordering predicate introduced in the 2019 revision of the IEEE 754 standard, which similarly defines NaN as smaller than all real numbers [13, §5.10]. Proceeding in increasing order, we assign the remaining values such that $0 \cdots 0$ always maps to zero, and the lexicographically largest balanced ternary integer, $1 \cdots 1$, maps to ∞ . The result of this assignment scheme for $n \in \{1, 2, 3\}$ is illustrated in Figure 2.

3.1 Filter 1: Asymmetry

The attentive reader may have already observed that while the assignment illustrated in Figure 2 works well for $n \in \{1, 2\}$, it introduces an asymmetry for $n = 3$: the upper quadrants contain five elements, whereas the lower quadrants contain six. This irregularity marks the first filter in our derivation process. It stems from the fact that ternary strings do not yield state counts as conveniently structured as in the binary case, where the total number of states is 2^n for n bits, and the number of assignable values per quadrant, namely $(2^n - 4)/4$ (after

reserving the four special elements NaR, -1 , 0 , and 1), is always an integer, specifically $2^{n-2} - 1$.

In contrast, no such regularity emerges for $3^n - 5$ (or $3^n - 4$ if we omit ∞). Nevertheless, a single viable path is shown in the following

Proposition 1. *Let $n \in \mathbb{N}_0$. It holds $4 \mid (3^{2n} - 5)$ and $4 \nmid (3^n - 4)$.*

Proof. With $a \equiv b \pmod{c} \rightarrow a^n \equiv b^n \pmod{c}$ for $a, b, c \in \mathbb{Z}$ via [1, Lemma 5.2c] it holds

$$4 \mid (3^{2n} - 5) \Leftrightarrow 3^{2n} - 5 \equiv 0 \pmod{4} \quad (3)$$

$$\Leftrightarrow 9^n \equiv 5 \pmod{4} \quad (4)$$

$$\Leftrightarrow 9^n \equiv 1 \pmod{4} \quad (5)$$

$$\Leftrightarrow 9^n \equiv 1^n \pmod{4} \quad (6)$$

$$\Leftrightarrow 9 \equiv 1 \pmod{4}. \quad (7)$$

We can also show

$$4 \mid (3^n - 4) \Leftrightarrow 3^n - 4 \equiv 0 \pmod{4} \Leftrightarrow 3^n \equiv 0 \pmod{4} \Leftrightarrow 4 \mid 3^n, \quad (8)$$

however 3^n is never divisible by four as it's odd. \square

This proposition demonstrates two key points. First, it confirms that a meaningful ternary discretisation of the real wheel algebra is indeed possible, provided that n is even. Second, it highlights that reserving only four special elements, the approach taken by posits and takums, can never yield a symmetric distribution, and thus is fundamentally unsuitable for this purpose. We therefore proceed under the assumption that n is even and use the assignment as outlined earlier.

3.2 Filter 2: Misfit Tool

Despite having overcome the first filter, and having settled on the general approach to restrict n to even numbers, we have yet to define the mapping rule for the values within each quadrant beyond the fundamental elements. This mapping should ideally follow the tapered precision paradigm. However, here we encounter the second filter: the prefix strings used in posit arithmetic cannot be translated directly into ternary logic, rendering them an ill-suited tool for ternary arithmetic. Similarly, takum's fixed-size regime offers no straightforward adaptation to the ternary case.

Let us first consider a positive trit string $t \in \{0 \cdots 01, \dots, 1 \cdots 10\}$. Our aim is to process this string in such a way that it yields a meaningful mapping to the positive real numbers. The initial observation is that, given the extreme cases $0 \cdots 0$ (representing 0) and $1 \cdots 1$ (representing ∞), and knowing that all quadrants contain the same number of elements, the value 1 must lie exactly in the middle. The corresponding ternary sequence for 1 is $1T \cdots 1T$, since $1T + 1T = 11$ holds, a property that extends to any trit string of even length. Given that

1 corresponds to exponent zero, a natural approach is to subtract $1T \cdots 1T$ from \mathbf{t} . This yields zero when $\mathbf{t} = 1T \cdots 1T$, a positive value when \mathbf{t} lies in the upper quadrant $(1, \infty)$, and a negative value when \mathbf{t} is in the lower quadrant $(0, 1)$.

By construction, the expression $(\mathbf{t} - 1T \cdots 1T)$ lies within the integer set $\{T1 \cdots T10T, \dots, 1T \cdots 1T01\}$. We can now employ the takum approach, using a fixed number of trits to encode the so-called ‘regime’ value, which serves as the parameter for a variable-length exponent. Although this may seem counter-intuitive given, after all, the regime values appear to be evenly spaced, the key insight is that each additional trit used for the exponent reduces the number of trits available for encoding the fraction. As a result, the number density decreases by a factor of three for each trit allocated to the exponent, thus realising tapered precision.

The number of trits allocated to the regime represents a trade-off: fewer regime trits allow for higher maximum precision, but at the expense of the number of representable regime values. We now explore this design choice in more detail. If we assign the first two trits of $\mathbf{t} - 1T \cdots 1T$ to represent the regime, the regime trits fall within the range $\{T1, \dots, 1T\}$, corresponding to the integer interval $\{-2, \dots, 2\}$. In contrast, assigning the first four trits as regime results in a range of $\{T1T1, \dots, 1T1T\}$, mapping to the integer interval $\{-20, \dots, 20\}$, a range far exceeding the requirements for encoding exponent lengths. A compromise is to use three regime trits, yielding a regime range of $\{T1T, \dots, 1T1\}$, corresponding to the integer interval $\{-7, \dots, 7\}$. This choice offers a balanced trade-off between exponent flexibility and precision. The precise role of the regime value in variable-length exponent encoding will be discussed later.

The final aspect to address is the treatment of negative trit strings $\mathbf{t} \in \{T \cdots T0, \dots, 0 \cdots 0T\}$. A straightforward approach is to compute the modulus $|\mathbf{t}|$ and process the result. This is justified by the symmetry of the representation: both upper quadrants $(-\infty, -1)$ and $(1, \infty)$ correspond to the same positive exponents, while the lower quadrants $(-1, 0)$ and $(0, 1)$ share the same negative exponents. The only difference lies in the sign. Thus, for all trit strings $\mathbf{t} \in \mathbb{T}_n$, we shall compute $|\mathbf{t}| - 1T \cdots 1T$ and assign the first three trits to be the regime trits \mathbf{r} . As we ‘anchor’ the trits onto a workable representation, we give the following corresponding

Definition 7 (anchor function). *Let $n \in 2\mathbb{N}_0$. The anchor function $\text{anc}_n: \mathbb{T}_n \rightarrow \mathbb{T}_n$ is defined as $\text{anc}_n(\mathbf{t}) = |\mathbf{t}| - 1T \cdots 1T$.*

An illustration of the proposed mapping for $n = 4$ is provided in Figure 3. At this point, we can confidently conclude that we have found a way to make the tool, namely the takum approach, fit again, overcoming the second filter in the process.

3.3 Filter 3: Excess

Having set the number of regime trits to three, after excluding the alternative choices of two and four, the next stage is to actually investigate the format we obtain from this approach. As mentioned before, a straightforward approach is

Table 1: Overview of mappings from regime values to exponent trit counts.

	r	0	1	2	3	4	5	6	7
$c(r)$	$\text{anc}(t)$	000T...T	001T...T	01TT...T	010T...T	011T...T	1TTT...T	1T0T...T	1T1T...T
		0001...1	0011...1	01T1...1	0101...1	0111...1	1TT1...1	1T01...1	1T1T...1T01
$ r $	c	0	1	2	3	4	5	6	7
	e_{\min}		T	TT	TTT	TTTT	TTTTT	TTTTTT	TTTTTTT
	e_{\max}		1	11	111	1111	11111	111111	T1T1T1T
	$\text{int}(e)$	0..0	-1..1	-4..4	-13..13	-40..40	-121..121	-364..364	-1093..-547
	b	0	2	8	26	80	242	728	2186
	e	0..0	1..3	4..12	13..39	40..120	121..363	364..1092	1093..1639
	$\lg(3^e)$	0..0	0.5..1.4	1.9..5.7	6.2..19	19..57	58..173	174..521	521..782
$\max(0, r - 1)$	c	0	0	1	2	3	4	5	6
	e_{\min}			T	TT	TTT	TTTT	TTTTT	TTTTTT
	e_{\max}			1	11	111	1111	11111	T1T1T1
	$\text{int}(e)$	0..0	0..0	-1..1	-4..4	-13..13	-40..40	-121..121	-364..-182
	b	0	1	3	9	27	81	243	729
	e	0..0	1..1	2..4	5..13	14..40	41..121	122..364	365..547
	$\lg(3^e)$	0..0	0.5..0.5	1.0..1.9	2.4..6.2	6.7..19	20..58	58..174	174..261
α	c	0	0	1	1	2	3	4	5
	e_{\min}			T	T	TT	TTT	TTTT	TTTTT
	e_{\max}			1	1	11	111	1111	T1T1T
	$\text{int}(e)$	0..0	0..0	-1..1	-1..1	-4..4	-13..13	-40..40	-121..-61
	b	0	1	3	6	12	30	84	246
	e	0..0	1..1	2..4	5..7	8..16	17..43	44..124	125..185
	$\lg(3^e)$	0..0	0.5..0.5	1.0..1.9	2.4..3.3	3.8..7.6	8.1..21	21..59	60..88
$\max(0, r - 2)$	c	0	0	0	1	2	3	4	5
	e_{\min}				T	TT	TTT	TTTT	TTTTT
	e_{\max}				1	11	111	1111	T1T1T
	$\text{int}(e)$	0..0	0..0	0..0	-1..1	-4..4	-13..13	-40..40	-121..-61
	b	0	1	2	4	10	28	82	244
	e	0..0	1..1	2..2	3..5	6..14	15..41	42..122	123..183
	$\lg(3^e)$	0..0	0.5..0.5	1.0..1.0	1.4..2.4	2.9..6.7	7.1..20	20..58	59..87
β	c	0	0	0	1	2	3	4	
	e_{\min}				T	T	TT	TTT	TTTT
	e_{\max}				1	1	11	111	T1T1
	$\text{int}(e)$	0..0	0..0	0..0	-1..1	-1..1	-4..4	-13..13	-40..-20
	b	0	1	2	4	7	13	31	85
	e	0..0	1..1	2..2	3..5	6..8	9..17	18..44	45..65
	$\lg(3^e)$	0..0	0.5..0.5	1.0..1.0	1.4..2.4	2.9..3.8	4.3..8.1	8.6..21	21..31
$\max(0, r - 3)$	c	0	0	0	0	1	2	3	4
	e_{\min}					T	TT	TTT	TTTT
	e_{\max}					1	11	111	T1T1
	$\text{int}(e)$	0..0	0..0	0..0	0..0	-1..1	-4..4	-13..13	-40..-20
	b	0	1	2	3	5	11	29	83
	e	0..0	1..1	2..2	3..3	4..6	7..15	16..42	43..63
	$\lg(3^e)$	0..0	0.5..0.5	1.0..1.0	1.4..1.4	1.9..2.9	3.3..7.2	7.6..20	21..30
γ	c	0	0	0	0	1	1	2	3
	e_{\min}					T	T	TT	TTT
	e_{\max}					1	1	11	T1T
	$\text{int}(e)$	0..0	0..0	0..0	0..0	-1..1	-1..1	-4..4	-13..-7
	b	0	1	2	3	5	8	14	32
	e	0..0	1..1	2..2	3..3	4..6	7..9	10..18	19..25
	$\lg(3^e)$	0..0	0.5..0.5	1.0..1.0	1.4..1.4	1.9..2.9	3.3..4.3	4.8..8.6	9.1..12

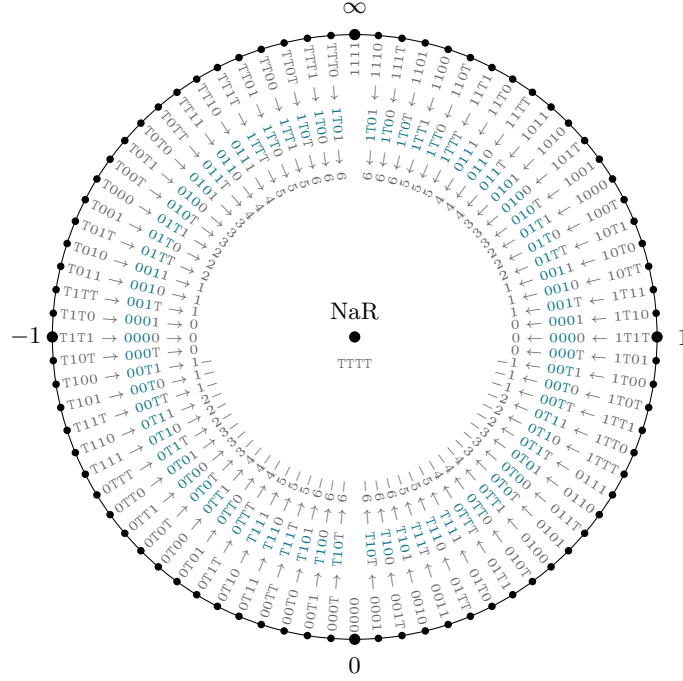


Figure 3: Mapping of ternary strings \mathbf{t} of length 4 to the real wheel algebra. The values of $\text{anc}_4(\mathbf{t})$ are given, with the first three trits, designated as the regime trits \mathbf{r} , highlighted accordingly. The corresponding regime values r are also indicated, partially with suffixed signs for better visual consistency.

to take the modulus of the regime value $r \in \{-7, \dots, 7\}$ and use it as the count c of exponent trits \mathbf{e} following after the regime bits. With a careful selection of biases b added to the value $\text{int}(\mathbf{e})$ represented by the exponent trits we obtain a consecutive sequence of exponent values e . Please refer to the first row block of Table 1 to observe the derivation of biases and outcome in terms of exponent values for each regime.

We notice a big issue: This approach yields an obviously excessive dynamic range of $10^{\pm 782}$. But what is a good general purpose dynamic range? The first work setting a rough bound is by QUEVEDO, who remarks ‘No he señalado límite al valor del exponente; pero es evidente que en todos los cálculos usuales será menor de ciento’ (I have not set a limit on the value of the [base-10] exponent, but it is clear that in all usual calculations it will be less than one hundred; translation by the author) [23, pp. 582 sq.]. A more thorough investigation in [12, Section 1.2] makes the case for a dynamic range of $10^{\pm 55}$. While one can go into the multitude of application-specific arithmetic, for which there is almost no lower bound on dynamic range, the author would like to make the case that

if someone builds general-purpose, dedicated hardware, the arithmetic shall be as well.

To tame the excessive dynamic range yielded by choosing $c(r) = |r|$, we consider the parametric liberties we have in this case. The mapping $c(r)$ shall be non-decreasing and start at zero. An additional justifiable constraint is that consecutive counts should only differ by zero or one. These constraints limit the choices of $c(r)$, and the first six are given in Table 1: Three mappings are straightforward shifts, where we only ‘start counting’ after a certain threshold is exceeded. Until then the regimes only represent a single exponent value. Given are also three more special mappings called ‘alpha’, ‘beta’ and ‘gamma’, where after the skip the counting is not linear, but actually tapered itself by repeating the count one for two times.

Overall we can see, in terms of dynamic ranges, a wide selection, going all the way down to $10^{\pm 12}$. Given we explored all possible, reasonable choices of $c(r)$ within this range we can say with confidence that this parametric space has been exhausted. While $\max(0, |r| - 1)$ still has an excessive dynamic range of $10^{\pm 261}$, the choices beta, $\max(0, |r| - 3)$ and gamma have insufficient general purpose dynamic ranges $10^{\pm 31}$, $10^{\pm 30}$ and $10^{\pm 12}$ respectively. Only remaining as candidates are alpha and $\max(0, |r| - 2)$ with similar dynamic ranges $10^{\pm 88}$ and $10^{\pm 87}$ respectively, both satisfying both the QUEVEDO limit of 10^{100} and the lower bound 10^{55} derived in [12]. The final choice falls on $\max(0, |r| - 2)$, as it has higher precision for smaller values, which should always take precedence over the rarer ‘outer’ numbers.

4 Tekum Definition

In this section, we introduce the tekum format, building upon the observations made in the previous section. The term ‘tekum’ is derived from a combination of ‘ternary’ and ‘takum’, reflecting the two foundational pillars of its design. First, the format is based on a balanced ternary representation. Second, it follows the design principles of takum, most notably its limited dynamic range. The name ‘takum’ itself originates from the Icelandic phrase ‘takmarkað umfang’, which translates to ‘limited range’. On this basis, we obtain the following format:

Definition 8 (tekum encoding). *Let $n \in 2\mathbb{N}_1$ with $n \geq 8$. Any $\mathbf{t} \in \mathbb{T}_n$ with $\mathbf{r} \# \mathbf{e} \# \mathbf{f} := \text{anc}_n(\mathbf{t})$ of the form*

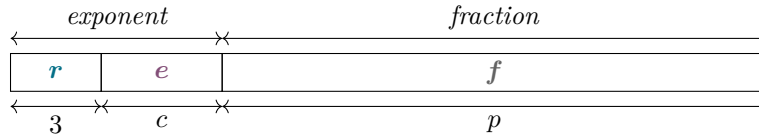





Table 2: Overview of the tekum arithmetic colour scheme.

colour	identifier	OKLCH	CIELab	HEX (sRGB)
	regime	(50%, 0.09, 220)	(42.44, -20.42, -21.18)	#046F87
	exponent	(50%, 0.09, 335)	(40.72, 27.23, -13.9)	#834F78
	fraction	(50%, 0.00, 0)	(42.00, 0.00, 0.00)	#636363

with regime trits \mathbf{r} , exponent trits \mathbf{e} , fraction trits \mathbf{f} , and

$$s := \text{sign}(\text{int}_n(\mathbf{t})) \quad : \text{sign} \quad (9)$$

$$r := \text{int}_3(\mathbf{r}) \in \{-7, \dots, 7\} \quad : \text{regime value} \quad (10)$$

$$c := \max(0, |r| - 2) \in \{0, \dots, 5\} \quad : \text{exponent trit count} \quad (11)$$

$$p := n - c - 3 \in \{n - 8, \dots, n - 3\} \quad : \text{fraction trit count} \quad (12)$$

$$\begin{aligned} b &:= \text{sign}(r) \cdot \left\lfloor 3^{|r|-2} + 1 \right\rfloor \\ &= \text{sign}(r) \cdot (0, 1, 2, 4, 10, 28, 82, 244)_{|r|} \quad : \text{exponent bias} \end{aligned} \quad (13)$$

$$e := \text{int}_c(\mathbf{e}) + b \in \{-183, \dots, 183\} \quad : \text{exponent value} \quad (14)$$

$$f := 3^{-p} \text{int}_p(\mathbf{f}) \in (-0.5, 0.5) \quad : \text{fraction value} \quad (15)$$

encodes the tekum value

$$\theta_n(\mathbf{t}) := \begin{cases} \text{NaR} & \mathbf{r} \# \mathbf{e} \# \mathbf{f} = \text{T} \dots \text{T} \\ 0 & \mathbf{r} \# \mathbf{e} \# \mathbf{f} = 0 \dots 0 \\ \infty & \mathbf{r} \# \mathbf{e} \# \mathbf{f} = 1 \dots 1 \\ s \cdot (1 + f) \cdot 3^e & \text{otherwise} \end{cases} \quad (16)$$

with $\theta: \mathbb{T}_n \mapsto \{\text{NaR}, 0, \infty\} \cup \pm (0.5 \cdot 3^{-183}, 1.5 \cdot 3^{183})$. Without loss of generality, any trit string shorter than 8 trits with an even, positive number of trits is also included in the definition by matching the special cases (NaR, 0, and ∞) respectively and expanding $\mathbf{r} \# \mathbf{e} \# \mathbf{f}$ with zeros in non-special cases. The case $n = 1$, which only covers the special cases, is also trivially included. By convention, NaR and ∞ are defined to be smaller and larger than any other represented value, respectively.

We also introduce a tekum colour scheme, prioritising uniformity in both lightness and chroma within the perceptually uniform OKLCH colour space [17]. Detailed colour definitions are delineated in Table 2.

Just as with the posit and takum formats, the tekum format can also be expressed as a logarithmic number system by renaming the fraction trits as *mantissa trits*, denoted by \mathbf{m} , and correspondingly the fraction value as m , the *mantissa value*. In this case, the fourth expression in (16) becomes $s \cdot 3^{e+m}$. Since the transition from binary to ternary is already a sufficiently radical step, we

refrain from introducing a different base for the logarithmic format and retain base 3. In this way, both the linear and logarithmic tekum formats share the same overall numerical properties, with the logarithmic form serving merely as an implementation detail. Within the scope of this work we do not pursue logarithmic tekums further, but simply note the possibility.

Particular attention may be drawn to the floating-point representation $(1 + f) \cdot 3^e$. Although it may appear counterintuitive, it is indeed correct. This becomes evident when considering its value range, $(0.5 \cdot 3^e, 1.5 \cdot 3^e)$. Notably, the upper bound $1.5 \cdot 3^e$ is equal to $0.5 \cdot 3^{e+1}$, thereby connecting seamlessly to the adjacent range $(0.5 \cdot 3^{e+1}, 1.5 \cdot 3^{e+1})$.

As an illustrative example, Table 3 presents the decoding of all positive 4-trit tekums, including a complete account of the intermediate quantities. Even at this limited size, tekums already exhibit a substantial dynamic range, approaching the recommended general-purpose range of $10^{\pm 55}$ derived in [12]. Nevertheless, the format warrants further evaluation with respect to its numerical properties. This forms the focus of the following section.

5 Evaluation

Before evaluating tekums against other formats, it is first necessary to consider the general principles required for fair comparisons between binary and ternary systems.

5.1 Comparing Binary with Ternary

As noted in the introduction, a trit contains approximately 1.58 bits of information. This difference must be accounted for when comparing binary and ternary formats. For example, when displaying a quantity relative to a bit count n , the ternary dataset must be scaled accordingly.

A particular challenge arises when conducting benchmarks between formats, where such simple rescaling is not possible. The standard binary widths of number formats are 8, 16, 32, and 64 bits. A direct comparison between an 8-bit format and an 8-trit format would be misleading, as the latter possesses $3^8 = 6561$ representations in contrast to only $2^8 = 256$. By dividing the binary bit widths 8, 16, 32, and 64 by $\log_2(3) \approx 1.58$, one obtains approximately 5.0, 10.1, 20.2, and 40.4, which are naturally matched by the integers 5, 10, 20, and 40. These represent the trit counts of ternary strings with approximately equivalent information content.

The difficulty arises at 8 bits: tekums are defined only for even trit counts, meaning a 5-trit tekum cannot be used. One possible approach is to evaluate both 4-trit and 6-trit variants and present both results, noting that an ‘imaginary’ 5-trit tekum would fall somewhere in between. For higher precisions, however, suitable matches can be found at the even trit counts of 10, 20, and 40, thereby enabling more direct comparisons.

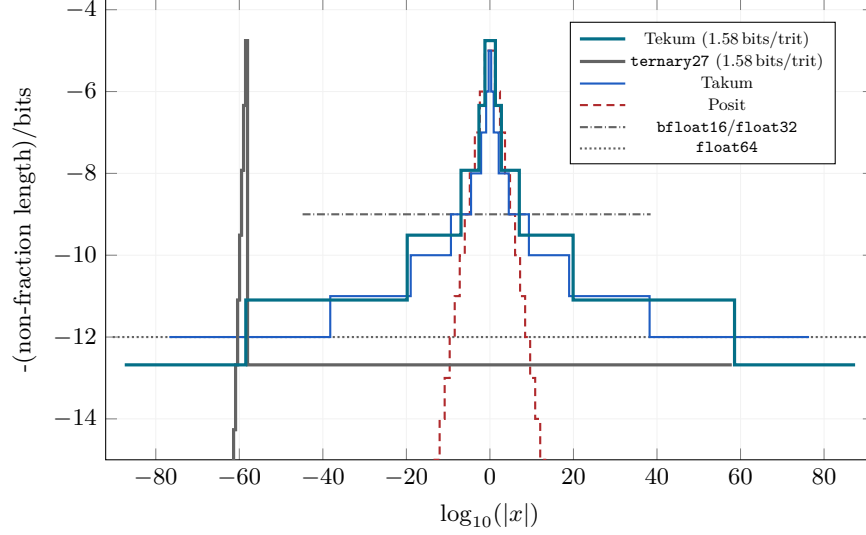


Figure 4: The number of non-fraction bits, which can be considered as overhead, relative to the represented value x in a selection of floating-point formats. The y-axis is inverted, thus meaning that higher values mean less overhead.

5.2 Format

Posits consist of five bit fields (sign bit, regime bits, regime terminator bit, exponent bits, and fraction bits). Takums are structurally similar, also requiring five fields (sign bit, direction bit, regime bits, characteristic bits, and fraction bits). In contrast, tekums are considerably simpler, comprising only three trit fields: regime trits, exponent trits, and fraction trits. This simplicity may not only facilitate formal analysis but could also prove advantageous for hardware implementation.

5.3 Accuracy

The first aspect of analysis concerns the accuracy of the tekum format. Following the approach taken in [12], one may consider the number of bits required to encode the sign and exponent of a given number $x > 0$, which can be interpreted as the overhead of encoding. The larger this overhead, the fewer fraction bits remain available. To enable comparison between binary and ternary formats, the overhead in trits for ternary formats is multiplied by $\log_2(3)$.

The results are shown in Figure 4. It is immediately evident that **ternary27** performs poorly, owing to its excessive waste of representations. The spike on the left is the result of a mechanism akin to subnormals; however, the effect is limited and occurs in an irrelevant region of numbers close to 10^{-60} .

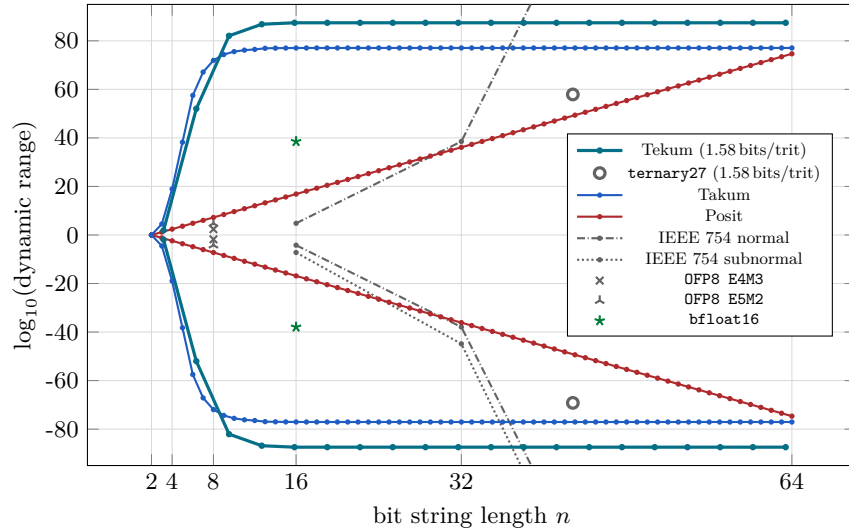


Figure 5: Dynamic range relative to the bit string length n for tekum, (linear) takum, posit and a selection of floating-point formats.

In contrast, tekums perform favourably. Compared to both posits and (linear) takums, tekums exhibit a region of highest accuracy of similar size to posits. This addresses a weakness of takums, which display a sharp drop in precision around the centre. Notably, tekums also achieve a significantly broader region of equal or superior precision compared with `bfloat16` and `float32`. In addition, tekums retain the same desirable logarithmic tapering property as takums.

5.4 Dynamic Range

The dynamic range results, shown in Figure 5, largely confirm expectations, as the tekum dynamic range was explicitly designed to be $10^{\pm 87}$. Although this range exceeds the previously identified minimum recommended general-purpose dynamic range of $10^{\pm 55}$, it remains comfortably within the bound suggested by QUEVEDO. Owing to the tapered precision property, only a small fraction of values extend beyond $10^{\pm 55}$.

It is also noteworthy that, as with takums, the dynamic range of tekums increases rapidly at low precisions, quickly reaching its maximum, which is a desirable feature. The `ternary27` format also achieves a general-purpose dynamic range, but this range is asymmetric due to its inclusion of subnormal representations.

5.5 Formal Analysis

Building upon the discussion of the format's properties, we now turn to the formal analysis and validation of the tekum format in terms of its fundamental characteristics, following the same methodology as applied to the *takum* format in [12, Section 4.5]. We begin by showing that there exist no redundant encodings, as stated in the following proposition.

Proposition 2 (Uniqueness). *Let $n \in 2\mathbb{N}_1$ and $\mathbf{t}, \mathbf{u} \in \mathbb{T}_n$. It holds*

$$\theta_n(\mathbf{t}) = \theta_n(\mathbf{u}) \Rightarrow \mathbf{t} = \mathbf{u}. \quad (17)$$

Proof. Assume $\theta_n(\mathbf{t}) = \theta_n(\mathbf{u})$. By construction, it follows that $\mathbf{t} = \mathbf{u}$ for the special cases $\theta_n(\mathbf{t}), \theta_n(\mathbf{u}) \in \{0, \infty, \text{NaR}\}$. Hence, it remains to consider the non-special case, for which we define $\theta_n(\mathbf{t}) := s \cdot (1 + f) \cdot 3^e$ and $\theta_n(\mathbf{u}) := \tilde{s} \cdot (1 + \tilde{f}) \cdot 3^{\tilde{e}}$.

Given $\theta_n(\mathbf{t}) = \theta_n(\mathbf{u})$, we can immediately deduce $s = \tilde{s}$, since $(1 + f), (1 + \tilde{f}), 3^{\tilde{e}}$, and 3^e are all positive. Moreover, as $(1 + f), (1 + \tilde{f}) \in (0.5, 1.5)$, both $f = \tilde{f}$ and $e = \tilde{e}$ are uniquely determined. By construction, $(s, e, f) = (\tilde{s}, \tilde{e}, \tilde{f})$ implies $\mathbf{t} = \mathbf{u}$. \square

Although this is a straightforward result, it highlights that, by design, no ternary representations are wasted on redundant encodings. We now proceed to show that ternary negation of a tekum corresponds to negating its numerical value.

Proposition 3 (Negation). *Let $n \in 2\mathbb{N}_1$ and $\mathbf{t} \in \mathbb{T}_n$ with $\theta_n(\mathbf{t}) \notin \{\text{NaR}, \infty\}$. It holds*

$$\theta_n(-\mathbf{t}) = -\theta_n(\mathbf{t}). \quad (18)$$

Proof. The only special case, namely $\mathbf{t} = 0 \cdots 0$, is trivial, given $\theta_n(-0 \cdots 0) = \theta_n(0 \cdots 0) = 0$ holds. Furthermore, it holds that

$$\text{anc}_n(-\mathbf{t}) = |-\mathbf{t}| - 1\text{T} \cdots 1\text{T} = |\mathbf{t}| - 1\text{T} \cdots 1\text{T} = \text{anc}_n(\mathbf{t}), \quad (19)$$

which implies that \mathbf{t} and $-\mathbf{t}$ share the same \mathbf{r} , \mathbf{e} , and \mathbf{f} , and thus the same fraction value f and exponent value e . The only distinction between them lies in their signs s , which are opposite, yielding the desired result. \square

Another result concerns the ternary monotonicity of tekums, which is formalised as follows.

Proposition 4 (Monotonicity). *Let $n \in 2\mathbb{N}_1$ and $\mathbf{t}, \mathbf{u} \in \mathbb{T}_n$. It holds*

$$\text{int}_n(\mathbf{t}) < \text{int}_n(\mathbf{u}) \Rightarrow \theta_n(\mathbf{t}) < \theta_n(\mathbf{u}). \quad (20)$$

Proof. We first consider the special cases. By convention, NaR is smaller than any other represented tekum value, which means that $\text{NaR} < \theta_n(\mathbf{u})$ holds. It also holds $\text{int}_n(\text{T} \cdots \text{T}) < \text{int}_n(\mathbf{u})$, and $\text{T} \cdots \text{T}$ is the ternary representation of NaR. Hence, monotonicity is satisfied for NaR. An analogous argument applies to ∞ , which by convention is larger than any other represented tekum value and has the ternary representation $1 \cdots 1$, thereby establishing monotonicity for ∞ .

We now turn to the non-special cases. By Proposition 3, it suffices to prove the result for \mathbf{t} and \mathbf{u} with $\text{int}_n(\mathbf{t}) > 0$ and $\text{int}_n(\mathbf{u}) > 0$, as $\text{int}_n(0 \cdots 0) = 0$ and $\theta_n(0 \cdots 0) = 0 < \theta_n(\mathbf{t}), \theta_n(\mathbf{u})$. The objective can be further reduced to showing that, for any $\mathbf{t} \in \{0 \cdots 01, \dots, 1 \cdots 01\}$,

$$\theta_n(\mathbf{t}) < \theta_n(\mathbf{t} + 0 \cdots 01), \quad (21)$$

i.e. the successive monotonicity of the positive non-special cases.

We denote $\text{anc}_n(\mathbf{t}) := \mathbf{r} \# \mathbf{e} \# \mathbf{f}$ and $\text{anc}_n(\mathbf{t} + 0 \cdots 01) := \tilde{\mathbf{r}} \# \tilde{\mathbf{e}} \# \tilde{\mathbf{f}}$, where tildes denote the components of the incremented ternary vector. For $\text{int}_n(\mathbf{t}) > 0$, we have $|\mathbf{t}| = \mathbf{t}$, and in particular, provided there is no overflow,

$$\text{anc}_n(\mathbf{t} + 0 \cdots 01) = \text{anc}_n(\mathbf{t}) + 0 \cdots 01. \quad (22)$$

Hence, the field $\mathbf{r} \# \mathbf{e} \# \mathbf{f}$ is incremented, and to establish monotonicity, we must examine how this affects \mathbf{r} , \mathbf{e} and \mathbf{f} .

Although we work in balanced ternary, carries may still occur, analogous to the binary case. We therefore distinguish cases depending on whether the increment affects only \mathbf{f} , or whether a carry propagates into \mathbf{e} . The latter occurs precisely when $\mathbf{f} = 1 \cdots 1$, in which case we have $\tilde{\mathbf{f}} = \mathbf{T} \cdots \mathbf{T}$, leading to an additional case distinction on \mathbf{e} . In all cases, $\tilde{s} = s$.

Following this scheme, as the first case we assume $\mathbf{f} \neq 1 \cdots 1$. It then holds $\tilde{\mathbf{r}} = \mathbf{r}$, $\tilde{\mathbf{e}} = \mathbf{e}$ and $\tilde{\mathbf{f}} = \mathbf{f} + 3^{-p} > \mathbf{f}$, in particular $\tilde{e} = e$. It follows that

$$\theta_n(\mathbf{t} + 0 \cdots 01) = \tilde{s} \cdot (1 + \tilde{\mathbf{f}})3^{\tilde{e}} = s \cdot (1 + \tilde{\mathbf{f}})3^e > s \cdot (1 + \mathbf{f})3^e = \theta_n(\mathbf{t}), \quad (23)$$

which was to be shown.

The second case is $\mathbf{f} = 1 \cdots 1$ and $\mathbf{e} \neq 1 \cdots 1$. Then $\tilde{\mathbf{f}} = \mathbf{T} \cdots \mathbf{T}$ and $\tilde{\mathbf{r}} = \mathbf{r}$, which implies $\tilde{r} = r$, $\tilde{c} = c$, $\tilde{p} = p$, $\tilde{b} = b$, and $\tilde{e} = e + 1$ since the exponent field increments without overflowing. As $\tilde{\mathbf{f}} = -\mathbf{f}$, we have $\tilde{\mathbf{f}} = -\mathbf{f}$, and thus

$$\theta_n(\mathbf{t} + 0 \cdots 01) = \tilde{s} \cdot (1 + \tilde{\mathbf{f}})3^{\tilde{e}} = s \cdot (1 - \mathbf{f})3^{e+1} > s \cdot (1 + \mathbf{f})3^e = \theta_n(\mathbf{t}), \quad (24)$$

as required.

The third case is $\mathbf{r} \neq 1 \cdots 1$, $\mathbf{e} = 1 \cdots 1$ and $\mathbf{f} = 1 \cdots 1$. Then $\tilde{\mathbf{f}} = \mathbf{T} \cdots \mathbf{T}$ (so $\tilde{\mathbf{f}} = -\mathbf{f}$) and $\tilde{\mathbf{e}} = \mathbf{T} \cdots \mathbf{T}$. In particular, $\tilde{r} = r + 1$, hence $\tilde{b} > b$. By construction, the bias is chosen such that each exponent range is distinct for every regime. We can therefore deduce $\tilde{e} > e$, and it follows that

$$\theta_n(\mathbf{t} + 0 \cdots 01) = s \cdot (1 - \mathbf{f})3^{\tilde{e}} \geq s \cdot (1 - \mathbf{f})3^{e+1} > s \cdot (1 + \mathbf{f})3^e = \theta_n(\mathbf{t}), \quad (25)$$

as required.

There is no fourth case where all trit fields are $1 \cdots 1$, as \mathbf{t} is explicitly at most $1 \cdots 01$. Thus, all possibilities have been exhausted, completing the proof. \square

The significance of this result lies in the fact that there exists a direct correspondence between the ordering of ternary integer and tekum representations. This property, also present for posits and takums, offers a notable implementation advantage: the same logic used for integer comparison can be employed

for floating-point comparison, even eliminating the need for explicit special-case handling of NaR and ∞ .

A particularly noteworthy aspect in the context of tekums is that both NaR and ∞ integrate seamlessly into this framework. Since tekums constitute the first balanced ternary real number system, the introduction of a second special value, ∞ , in addition to NaR, posed a potential risk of breaking monotonicity. Remarkably, however, the structure remains fully consistent, an elegant and non-trivial property of the format.

We now turn to the next aspect, namely rounding. Balanced ternary exhibits the elegant property that rounding an integer is equivalent to simply truncating it. This can be seen by considering an arbitrary real number expressed in balanced ternary as the sum of an integral and a fractional part,

$$\sum_{i=0}^{\infty} a_i 3^i + \sum_{i=0}^{\infty} b_i 3^{-(i+1)}, \quad (26)$$

with $\mathbf{a}, \mathbf{b} \in \mathbb{T}_{\infty}$, which we wish to round to the nearest integer. The outcome depends solely on the fractional part, which in the binary case lies in the interval $[0, 1)$, necessitating a distinction between the intervals $[0, 0.5)$, exactly 0.5, and $(0.5, 1)$.

In the balanced ternary case, however, the fractional part lies in the interval $(-0.5, 0.5)$, meaning the nearest integer is always the integral part $\sum_{i=0}^{\infty} a_i 3^i$. In other words, rounding amounts to unconditionally truncating the fractional part. The question is how this property translates to tekums, which would benefit from such stable rounding behaviour. For this, we provide the following:

Proposition 5 (Truncation is Rounding). *Let $n \in 2\mathbb{N}_3$, $\mathbf{t} \in \mathbb{T}_n$ with $\theta(\mathbf{t}) \notin \{\text{NaR}, \infty\}$ and $\mathbf{a} := \text{anc}_n(\mathbf{t}) \in \mathbb{T}_n$. It holds*

$$\text{anc}_n^{\text{inv}}(\mathbf{a}_{n-1} \cdots \mathbf{a}_2) = \arg \min_{\mathbf{u} \in \mathbb{T}_{n-2}} (|\theta(\mathbf{t}) - \theta(\mathbf{u})|). \quad (27)$$

In other words, for a given n -trit tekum, the closest $(n-2)$ -trit tekum is obtained by truncating its anchor by two trits and converting it back to a tekum (using \mathbf{t} 's sign). Since the conversion between a tekum and its anchor is lossless, this property generalises to any truncation of the anchor, provided the anchor remains at least four trits long to avoid truncating the regime trits.

Proof. Within the scope of the anchor, we are not restricted to even trit counts, and may therefore consider the effect of truncating a single trit. Let

$$\mathbf{a} = \mathbf{r} \# \mathbf{e} \# \mathbf{f}, \quad (28)$$

and denote all truncated quantities with a tilde. Since truncation never occurs within the regime trits, we only distinguish between truncating an exponent trit and truncating a fraction trit. In both cases, the sign s remains unchanged.

In the first case, truncating a trit from the fraction yields $\tilde{\mathbf{f}}$, which represents the closest integer to \mathbf{f} . Consequently, $\tilde{\mathbf{f}}$ is the closest representable value to \mathbf{f} ,

and since e is unchanged, the tekum value

$$s \cdot (1 + \tilde{f}) \cdot 3^{\tilde{e}} = s \cdot (1 + \tilde{f}) \cdot 3^e \quad (29)$$

is optimally close to $\theta(\mathbf{t})$, as required.

In the second case, we truncate a trit from the exponent trits. It holds $f = \tilde{f} = 0$. As the regime trits remain unchanged, both the bias and the number of exponent trits are preserved ($b = \tilde{b}$, $c = \tilde{c}$). The truncated exponent

$$\tilde{e} = \text{int}_c(\tilde{e}) + b \quad (30)$$

is the closest representable integer to e , and the tekum value $s \cdot 3^{\tilde{e}}$ is therefore the optimal approximation of $\theta(\mathbf{t})$. \square

To give an example, consider rounding $\mathbf{t} = 01\text{TTT}1\text{TT}$ ($\text{anc}_n(\mathbf{t}) = 001\text{T}1110$, $s = 1$, $r = 1$, $c = 0$, $p = 5$, $b = 1$, $e = 1$, $f = 3^{-5} \cdot (-42) \approx -0.173$, $\theta(\mathbf{t}) \approx 1.654$) to four trits. We truncate the anchor to the four trits 001T and obtain the tekum $1\text{T}11$, which represents the value 2.0 and is the closest representation (see Table 3).

This property has far-reaching implications despite the small computational overhead of determining the anchor. Conversion between precisions is essentially cost-free, requiring only the truncation or extension of the anchor. Well-known issues in binary floating-point arithmetic, such as inaccuracies due to ‘double rounding’, do not arise here: Truncation can be performed in any number of steps with identical results.

5.6 Hardware Implementation

Ternary logic hardware remains in its infancy and is largely orthogonal to contemporary chip design. The theoretical advantages of ternary logic are substantial, offering higher information density, owing to its superior radix economy, which is particularly relevant in the context of the memory wall that constrains computer performance, and consequently reduced circuit and interconnect complexity, lower power dissipation, and potentially faster operational speeds.

Historically, apart from binary computers that merely emulate trits using two bits each, such as the Setun and Setun 70 computers [3], proposals for genuine ternary computers have relied on alternative technologies. These include JOSEPHSON junctions [20] and optical computing, where dark represents zero and two orthogonal polarisations of light encode T and 1 [14]. However, such approaches never progressed beyond experimental prototypes.

Ternary computing has recently attracted renewed interest for two main reasons: advances in ternary large-language models (LLMs) and developments in Carbon Nanotube Field-Effect Transistors (CNTFETs). In the former case, single trits are used to represent weights within deep neural networks, with notable examples including BitNet [19, 24] and proposals for ternary hardware implementations of neural networks [25]. Regarding CNTFETs, hundreds of publications in recent years have addressed advances ranging from general logic gates [6, 22] to arithmetic units such as adders [10].

In the context of implementing takums, a particular significance is attributed to an adder dedicated to adding $T1 \cdots T1$ for computing the anchor function anc_n , which is central to decoding and rounding tekums. This capability could enable more efficient circuit designs tailored to the tekum format.

It is noteworthy that the recent AI revolution may not only have catalysed the exploration of novel arithmetic formats, but could also indirectly foster a transition from binary to ternary computing.

6 Conclusion

This paper has introduced *tekum*, a novel balanced ternary real arithmetic format, derived from three fundamental design challenges and evaluated against rigorous criteria for fair comparison between binary and ternary representations.

Our analysis has demonstrated that tekums possess highly favourable numerical properties, not only in terms of precision and dynamic range, but also in offering features unique to the ternary domain. Unlike their binary counterparts, posits and takums, tekums simultaneously accommodate both ∞ and NaR, while retaining the simplicity of negation by flipping the underlying trit string. Perhaps most strikingly, tekums enable rounding by truncation, a property that eradicates at a stroke some notorious problems of rounding in binary arithmetic: double rounding errors, cascading carries in hardware, and the attendant inefficiencies.

Taken together, these properties position tekums not merely as a curiosity, but as a radical and compelling alternative to established binary real number formats. They demonstrate that balanced ternary arithmetic is not only viable, but potentially transformative. While substantial work remains, tekums stand as a bold step towards a new numerical foundation for future ternary computing, a foundation that could redefine the very architecture of computation itself.

References

- [1] Tom M. Apostol. *Introduction to Analytic Number Theory*. 1st ed. Undergraduate Texts in Mathematics. New York, NY, USA: Springer New York, 1976. DOI: 10.1007/978-1-4757-5579-4.
- [2] Peter A. Boncz, Stefan Manegold and Martin L. Kersten. ‘Database Architecture Optimized for the New Bottleneck: Memory Access’. In: *Proceedings of the 25th International Conference on Very Large Data Bases*. VLDB ’99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Sept. 1999, pp. 54–65. ISBN: 1558606157. DOI: 10.5555/645925.671364.
- [3] Nikolay Petrovich Brusentsov and José Ramil Alvarez. ‘Ternary Computers: The Setun and the Setun 70’. In: *Perspectives on Soviet and Russian Computing*. Ed. by John Impagliazzo and Eduard Proydakov. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, 2011, pp. 74–80. DOI: 10.1007/978-3-642-22816-2_10.

- [4] Jesper Carlström. ‘Wheels – on Division by Zero’. In: *Mathematical Structures in Computer Science* 14 (Jan. 2004), pp. 143–184. DOI: 10.1017/S0960129503004110.
- [5] Amir Gholami et al. ‘AI and Memory Wall’. In: *IEEE Micro* 44.3 (2024), pp. 33–39. DOI: 10.1109/MM.2024.3373763.
- [6] Namineni Gireesh, Shaik Javid Basha and Ahmed Elbarbary. ‘CNTFET-based digital arithmetic circuit designs in ternary logic with improved performance’. In: *e-Prime - Advances in Electrical Engineering, Electronics and Energy* 7 (2024). DOI: <https://doi.org/10.1016/j.prime.2024.100427>.
- [7] John L. Gustafson and Isaac Yonemoto. ‘Beating Floating Point at Its Own Game: Posit Arithmetic’. In: *Supercomputing Frontiers and Innovations* 4.2 (June 2017), pp. 71–86. DOI: 10.14529/jsfi170206.
- [8] John L. Gustafson et al. ‘Standard for Posit™ Arithmetic (2022)’. Mar. 2022. URL: https://web.archive.org/web/20220603115338/https://posithub.org/docs/posit_standard-2.pdf.
- [9] Robin Hanson. ‘The Great Filter – Are We Almost Past It?’ Fairfax, VA, USA, Sept. 1998. URL: <https://web.archive.org/web/20250906181317/https://mason.gmu.edu/~rhanson/greatfilter.html>.
- [10] Marzieh Hashemipour, Reza Faghih Mirzaee and Keivan Navi. ‘21T Ternary Full Adder Based on Capacitive Threshold Logic and Carbon Nanotube FETs’. In: *IEEE Transactions on Nanotechnology* 23 (2024), pp. 338–345. DOI: 10.1109/TNANO.2024.3386825.
- [11] Brian Hayes. ‘Third Base’. In: *American Scientist* 89.6 (Dec. 2001), pp. 490–494. DOI: 10.1511/2001.40.490.
- [12] Laslo Hunhold. ‘Beating Posits at Their Own Game: Takum Arithmetic’. In: *Next Generation Arithmetic. 5th International Conference, CoNGA 2024, Sydney, NSW, Australia, February 20–21, 2024, Proceedings*. Vol. 14666. Lecture Notes in Computer Science. Sydney, NSW, Australia: Springer Nature Switzerland, Oct. 2024. DOI: 10.1007/978-3-031-72709-2_1.
- [13] *IEEE Standard for Floating-Point Arithmetic*. July 2019. DOI: 10.1109/IEEESTD.2019.8766229.
- [14] Yi Jin, Huacan He and Yangtian Lü. ‘Ternary Optical Computer Principle’. In: *Science in China Series F: Information Sciences* 46.2 (Apr. 2003), pp. 145–150. ISSN: 1862-2836. DOI: 10.1360/03yf9012. URL: <https://link.springer.com/article/10.1360/03yf9012>.
- [15] Douglas W. Jones. ‘The Ternary Manifesto’. Apr. 2012. URL: <https://web.archive.org/web/20250815213618/https://homepage.cs.uiowa.edu/~dwjones/ternary/>.
- [16] Donald Ervin Knuth. *Seminumerical Algorithms*. 3rd ed. Vol. 2. The Art of Computer Programming. Boston, MA, USA: Addison Wesley Longman, 1997. ISBN: 978-0-201-89684-8. URL: <https://dl.acm.org/doi/10.5555/270146>.

- [17] Chris Lilley. ‘Color on the Web’. In: *Fundamentals and Applications of Colour Engineering*. John Wiley & Sons, Ltd, 2023. Chap. 16, pp. 271–291. ISBN: 9781119827214. DOI: 10.1002/9781119827214.ch16.
- [18] Peter Lindstrom, Scott Lloyd and Jeffrey Hittinger. ‘Universal Coding of the Reals: Alternatives to IEEE Floating Point’. In: CoNGA ’18 (Mar. 2018). DOI: 10.1145/3190339.3190344.
- [19] Shuming Ma et al. ‘BitNet b1.58 2B4T Technical Report’. In: (Apr. 2025), pp. 1–14. arXiv: 2504.12285 [cs.CL].
- [20] M. Morisue et al. ‘A Josephson Ternary Memory Circuit’. In: *Proceedings. 1998 28th IEEE International Symposium on Multiple- Valued Logic (Cat. No.98CB36138)*. 1998, pp. 19–24. DOI: 10.1109/ISMVL.1998.679270.
- [21] Rory O’Hare. ‘Ternary27: A Balanced Ternary Floating Point Format’. Sept. 2021. DOI: 10.5281/zenodo.16575889.
- [22] Kumar R. Santhosh and Munibyarappa Roopa. ‘The Designing of Carbon Nanotube FET Based Encoder Using Ternary Logic’. In: *2024 Global Conference on Communications and Information Technologies (GCCIT)*. 2024, pp. 1–6. DOI: 10.1109/GCCIT63234.2024.10862332.
- [23] Leonardo Torres Quevedo. ‘Automática: Complemento de la Teoría de las Máquinas’. In: *Revista de Obras Públicas* 2043 (Nov. 1914), pp. 575–583. URL: https://quickclick.es/rop/pdf/publico/1914/1914_tomoI_2043_01.pdf.
- [24] Hongyu Wang et al. ‘BitNet: Scaling 1-bit Transformers for Large Language Models’. In: (Oct. 2023), pp. 1–14. arXiv: 2310.11453 [cs.CL].
- [25] Xuehao Zhu et al. ‘High-Performance Ternary Logic Circuits and Neural Networks Based on Carbon Nanotube Source-Gating Transistors’. In: *Science Advances* 11.2 (2025). DOI: 10.1126/sciadv.adt1909.