CS 64

FLOATING-POINT NUMBER REPRESENTATIONS: BASE CHOICE VERSUS EXPONENT RANGE

BY

PAUL RICHMAN

TECHNICAL REPORT NO. CS 64
APRIL 28, 1967

This work was supported by the
National Science Foundation and the
Office of Naval Research

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

FLOATING-POINT NUMBER REPRESENTATIONS:
BASE CHOICE VERSUS EXPONENT RANGE

by

Paul Richman

Abstract:

A digital computer whose memory words are composed of r-state devices is considered. The choice of the base, β, for the internal floating-point numbers on such a computer is discussed. Larger values of β necessitate the use of more r-state devices for the mantissa, in order to preserve some "minimum accuracy," leaving fewer r-state devices for the exponent of β. As β increases, the exponent range may increase for a short period, but it must ultimately decrease to zero. Of course, this behavior depends on what definition of accuracy is used. This behavior is analyzed for a recently proposed definition[2] of accuracy which specifies when it is to be said that the set of q-digit base β floating-point numbers is accurate to p-digits base t. The only case of practical importance today is $t = 10$ and $r = 2$; and in this case we find that $\beta = 2$ is always best. However the analysis is done to cover all cases.

Table of Contents

Rotation

Symbol	Meaning
digit_r	digit of a base r number
$(o.a_1a_2\ldots a_n)_r$	the base r number $o.a_1\ldots a_n$ (where $o \leq a_i < r$)
$(b_1b_2\ldots b_n)_r$	the base r integer $b_1\ldots b_n$ (where $o \leq b_i < r$)
gcd(i,j)	greatest common divisor of the integers i and j
[X]	greatest integer < X

1. Introduction

I. B. Goldberg recently showed that 27 bits are not enough for 8-digit$_{10}$ accuracy (under a suitable definition of accuracy), but that 28 bits are.[1] He proved that if $2^{q-1} > 10^p > 2^{q-2}$ then q bits are enough for p-digit$_{10}$ accuracy. He also gave several examples (p=1,2,8) in which q-1 bits are not enough for p-digit$_{10}$ accuracy.

Shortly after this D. W. Matula independently discovered and proved his Base Conversion Theorem.[2] Let r and t be incommensurable integers ≥ 2 (r and t are commensurable if and only if $r^i = t^j$ for some positive integers i and j). The Base Conversion Theorem essentially states that q-digits$_r$ suffice for p-digit$_t$ accuracy if and only if $r^{q-1} > t^p - 1$.

In this paper the Base Conversion Theorem is extended to commensurable bases. These results are used in a discussion of the choice of the internal representation of floating-point numbers for an r-ary computer; i.e., a digital computer whose memory cells are composed of r-state devices. This representation is specified when

1) a base for the floating-point numbers is chosen and

2) the number of r-state devices to be used for the mantissa (and hence the exponent) is chosen.

For example the IBM **7090**, the Burroughs B5500 and the IBM 360 series computers are 2-ary computers. The bases for the internal representation of floating-point numbers in these three computers are 2, 8, and 16, respectively. In the IBM **7090**, 27 bits are used for the mantissa and 8 bits for the exponent. The mantissa is stored in binary notation,

an extra bit being provided for the sign. The value of the 8 bit exponent is used as an excess 128 exponent of 2; i.e., 2 raised to the power [(the value of the 8 --bit exponent)-128] is the exponent-part of the floating-point number. In the B5500, 39 bits are used for the mantissa and 7 bits for the exponent. The mantissa is stored in octal notation, each group of three bits representing one octal digit. The 7 bit exponent is used as a signed magnitude exponent of 8. The following is a basic property of this representation: if 1 is added to the exponent of such a number then its mantissa must be shifted right three bits: $(o.a_1a_2\ldots a_{12})_8 8^n = (o.oa_1a_2\ldots a_{12})_8 x 8^{n+1}$. In the IBM 360 series, 56 bits are used for the mantissa and 7 bits for the exponent (of a long word). The mantissa is stored in hexadecimal notation? each group of four bits representing one hex digit. The value of the 7 bit exponent is used as an excess 64 exponent of 16, and again $(o.b_1b_2\ldots b_{14})_{16} x 16^n = (o.ob_1b_2\ldots b_{13})_{16} x 16^{n+1}$.

We restrict our discussion to the case in which the choice of representation for an r-ary computer is subject to the following constraints only:

(i) if base s is chosen, with $r^k > s > r^{k-1}$, then the mantissa must be made of an integral multiple of k r-state devices, i.e., fractions of digits_s are not permitted;

(ii) the mantissa must be accurate to at least p-digits_t, for given p and t (accuracy is defined in Sec. 2);

(iii) the base chosen must give the largest exponent range possible subject to (i) and (ii).

Observe that larger bases offer larger exponent ranges, but require more bits to be used for the mantissa. Thus there is a definite trade-off involved in using larger bases, and it is not obvious which base(-s) will satisfy (i)-(iii). We prove that (1) if t is a power of r then t is the only base which allows all of (i)-(iii) to be satisfied; (2) if t and r are incommensurable then r is the only base which allows all of (i)-(iii) to be satisfied; (3) if t and r are commensurable and $r > 8$, then there are cases in which r is the only such base and cases in which r^2 is the only such base.

Constraints (i)-(iii) above are discussed further in the conclusion. We will find that these constraints can be weakened somewhat without disturbing our results. Applications are also discussed there.

2. p-digit$_t$ Accuracy

Following D. W. Matula[2] let us define the set of q-digit$_r$ numbers, $S(q\text{-}D_r)$, for $r > 2$ and $q > 1$, by

(201) $S(q\text{-}D_r) \equiv (x: |x| = \sum_{i=1}^{q} a_i r^{n-i}$ for integers n, a_i, with $0 \leq a_i < r\}$.

We will discuss the rounding and truncation conversion mappings from $S(q\text{-}D_r)$ into $S(p\text{-}D_t)$. Since the results presented in this paper are the same for both methods of conversion, we let $C: S(q\text{-}D_r) \rightarrow S(p\text{-}D_t)$ stand for either mapping. We say that $S(q\text{-}D_r)$ is accurate to p-digits$_t$ if and only if $C: S(p\text{-}D_t) \rightarrow S(q\text{-}D_r)$ is (l-l) and

$C:S(q\text{-}Dr) \rightarrow S(p\text{-}D_t)$ in onto. This means that distinctness of "input" numbers from $S(p\text{-}D_t)$ is preserved by rounding (or truncation) conversion into $S(q\text{-}D_r)$, and that all "output" numbers in $S(p\text{-}D_t)$ are attainable in the "output" conversion from $S(q\text{-}D_r)$ onto $S(p\text{-}D_t)$. This definition of accuracy is essentially* equivalent to the following due to I. B. Goldberg': for all x, if $x \in S(p\text{-}D_t)$ converts into $y \in S(q\text{-}D_r)$ which converts into $z \in S(p\text{-}D_t)$ then $S(q\text{-}D_r)$ is accurate to p-digits$_t$ if and only if $z = x$. Roughly speaking, this means that you must get out what you put in. We now state

Theorem I (The Base Conversion Theorem -- D. W. Matula[2])
Let r and t be incommensurable integers both ≥ 2. Then $C:S(p\text{-}D_t) \rightarrow S(q\text{-}D_r)$ is (l-l) if and only if $r^{q-1} > t^p - 1$ and is onto if and only if $t^{p-1} \geq r^q - 1$.

Observe that $S(q\text{-}D_r)$ is accurate to p-digits$_t$ precisely when $r^{q-1} \geq t^{p-1}$, since this inequality alone implies both the required (1-l)-ness and the required onto-ness.

Corollary I (D. W. Matula[2])
Let r and t be incommensurable integers both > 2.
Then $C:S(p\text{-}D_t) \rightarrow S(q\text{-}D_r)$ cannot be both (l-l) and onto.

*They are essentially, but not completely, equivalent.

Thus if the conversion mapping is to preserve distinctness, it cannot make use of all the numbers available in the range set, and vice versa. This corollary also applies to the commensurable case, as is shown in the appendix*

Example 2.1:

By our definition of accuracy, the sets $S(14\text{-}D_{16})$ and $S(51\text{-}D_2)$ are both accurate to 15-digits_{10}. Observe that all numbers in $S(51\text{-}D_2)$ can be represented exactly in $S(14\text{-}D_{16})$, but not vice versa. Yet $S(51\text{-}D_2)$ is just as accurate, base 10, as $S(14\text{-}D_{16})$. Of course $S(14\text{-}D_{16})$ is more accurate, base 2 or base 16, than is $S(51\text{-}D_2)$, since $S(51\text{-}D_2)$ is only accurate to 12-digits_{16}.

We are mainly interested in the case $r = 2$ and $t = 10$ since modern computers are binary and since base 10 is used both in daily life and in higher level computer languages such as FORTRAN and ALGOL. Applications to other values of t and (eventually) to other values of r are also of interest It may be, for example, that one really wants to attain 14-digit_{16} accuracy in a binary computer. Our results show that, in this case, the unique best representation (subject to (i)-(iii) in Sec. 1) is just 14-digits_{16}.*

In the next section we give an example to clarify and direct our discussion. The reader is referred to D. Matula's paper[2] for a clear, detailed discussion of this definition of accuracy and its ramifications.

*This is not as obvious as it may at first appear. It seems possible that base 32 or base 64, for example, could yield a wider exponent range than base 16 while preserving 14-digit_{16} accuracy.

3. An Example

Suppose we are given 63 bits in which to store the mantissa and exponent of a floating-point number and we wish to achieve 15-digit$_{10}$ accuracy. Which of the bases 2, 4, 8 and 16 will give us the widest exponent range (while preserving 15-digit$_{10}$ accuracy)?

The inequalities

$$
\begin{aligned}
2^{50} &\geq 10^{15} - 1 > 2^{49} \\
4^{25} &> 10^{15} - 1 > 4^{24} \\
8^{17} &\geq 10^{15} - 1 > 8^{16} \\
16^{13} &\geq 10^{15} - 1 > 16^{12}
\end{aligned}
\tag{3.1}
$$

along with D. W. Matula's Base Conversion Theorem, show that we need $[51, 26, 18, 14]$ - digits$_{[2, 4, 8, 16]}$ of mantissa, respectively, for 15-digit$_{10}$ accuracy. Thus we need $[51, 52, 54, 56,]$-bits for the mantissa, leaving $[12, 11, 9, 7]$-bits for the exponent. If the signed magnitude method of storing the exponent is used, the exponent-part ranges are $[2^{\&(2\%)}, 4^{\pm(2^{10}-1)}, 8^{\pm(2^{8}-1)}, 16^{\pm(2^{6}-1)}]$. Let $B_{\pm} = 2^{\pm(2^{11}-1)}$, the range for base 2. The equations

$$
\begin{aligned}
4^{\pm(2^{10}-1)} &= 2^{\mp 1} \; B_{\pm} \\
8^{\pm(2^{8}-1)} &= 2^{\mp 27/8} \; B_{\pm}^{3/8} \\
16^{\pm(2^{6}-1)} &= 2^{\mp 33/8} \; B_{\pm}^{1/8}
\end{aligned}
\tag{3.2}
$$

show that base 2 has the largest exponent range; for example, $4^{2^{10}-1} = .5B_+ < B_+$ and $4^{-2^{10}+1} = 2B > B$. The difference between base 2 and base 16 is a factor of 8 in the exponent; a range of $10^{\pm 616}$ versus $10^{\pm 76}$.

The excess-quantity method of storing exponents will be of principal interest here, although our results are the same for both methods. If n r-state devices are used to form the exponent, then $[.5r^n]$ is considered a zero exponent; those above are positive, those below are negative. This avoids wasting one of the possible exponent values on -o and eliminates the necessity of an exponent sign bit. With this method the exponent-part ranges for the example above are $[2_{-2^{11}}, 4_{-2^{10}}, 8^{-2^8}, 16^{-2^6}]$ to $[22^{11}-1, 4\,2^{10}-1, 8^{2^8}-1, 16^{2^6}-1]$ and again base 2 is best.

If 16-digit$_{10}$ accuracy is desired then the same sort of analysis shows that base 2 is again better than 4, 8, 16, yielding an exponent-part range of $10^{\pm 38}$ as opposed to $10^{\pm 4}$ for base 16.

4. The General Case

We now consider the general case in which N r-state devices are used in forming the exponent and mantissa of floating-point numbers. A mantissa sign bit is to be provided separately. Given positive integers t and p, we wish to find the base, s, which gives the widest exponent range while preserving p-digit$_t$ accuracy (see Sec. 1, constraints

(i)-(iii)). We will call s a best (p-D_t, N, r)-base. If s remains a best (p-D_t, N, r)-base for all appropriate N (i.e., all N which allow p-$digit_t$ accuracy to be realized in fewer than N-$digits_r$) we will call s a best (p-D_t, *, r)-base, etc. We will call t the target base. We will find that there is always a unique best (p-D_t, *, r)-base.

Theorem II. The best (p-D_t, N, r)-bases are always of the form r^j for $j \geq 1$.

In other words, the best bases are those which make use of all the possible states of the r-state devices.

Proof: If a base s not of the form r^j is used, then k-$digits_r$ (i.e. k r-state devices) are used for each $digit_s$ of the mantissa, where $r^k > s > r^{k-1}$. If n-$digits_r$ are left for the exponent then the exponent-part range is $s^{[\pm .5(r^n-1)]}$, using the excess-quantity method. (It is $s^{\pm(r^n-1)}$ if an exponent sign bit is provided and the signed magnitude method is used.) If base r^k were used instead of base s then no more $digits_r$ would be needed and the exponent-part range would be $r^{k[\pm .5(r^n-1)]}$ ($r^{\pm k(r^n-1)}$ for signed magnitude exponents), a strictly larger range.

Q.E.D.

The following lemma gives a sufficient condition for [the exponent-part range of a representation using the smaller base r^k] to be greater than [the exponent-part range of a representation using the larger base $r^i (i > k)$] . The lemma states that if positive integers i, k, n_i and n_k satisfy $n_i \leq n_k + k-i$ and $k < i$ then the exponent-part $(r^k)^{a_1 a_2 \cdots a_{n_k}}$ has a wider range of values than the exponent-part $(r^i)^{b_1 b_2 \cdots b_{n_i}}$, where $a_1 \cdots a_{n_k}$ and $b_1 \cdots b_{n_i}$ are arbitrary n_k-digit$_r$ and n_i-digit$_r$ integers, respectively.

Lemma I. If $i < n_i + i < n_k + k$ for given integers $i > k > 1$ then

$$i(r^{n_i}-1) < k\,(r^{n_k}-1) \tag{4.1}$$

$$i[.5(r^{n_i}-1)] < k\,[.5(r^{n_k}-1)] \tag{4.2}$$

$$i[-.5(r^{n_i}-1)] \geq k\,[-.5(r^{n_k}-1)] \tag{4.3}$$

Proof: The hypothesis $n_i \leq n_k + k-i$ can be rewritten as

$$i\,r^{n_i} \leq k\,r^{n_k} \left(\frac{i}{r^i}\right)\left(\frac{r^k}{k}\right) . \tag{4.4}$$

The function $f(x) = x\,r^{-x}$ is a strictly decreasing function for $x > 2$. Further $f(2) \leq f(1)$, where equality can occur only when $r = 2$, and so equation (4.4) along with $-i < -k$ imply (4.1). The factor $r^{-(i-k)_{i/k}}$ in (4.4) essentially bounds the ratio of [the range

of the exponent for base r^i] to [the range of the exponent for base r^k] . As i-k increases, this upper bound decreases slightly less than exponentially. Inequalities (4.2) and (4.3) follow easily from (4.1).

Q.E.D.

This lemma will be used in the proof of Theorems III and V. We now state the Best Base Theorem for incommensurable bases. It is given in full generality in Sec. 6.

Theorem III. If t and r are incommensurable then base r alone is the best $(*\text{-}D_t, *, r)$-base.

Proof:

Let integers N, $p \geq 1$ and $t \geq 2$ (t and r incommensurable) be given. Let integers q_i satisfy

$$(4.5) \qquad (r^i)^{q_i-1} \geq t^p - 1 > (r^i)^{q_i-2} \quad \text{for } i = 1, 2, \ldots$$

If r^i is used as base then, according to D. Matula's Base Conversion Theorem, q_i-digits$_{r^i}$ are needed for p-digit$_t$ accuracy. Precisely iq_i-digits$_r$ are needed to hold q_i-digits$_{r^i}$ and so $n_i = N-iq_i$ r-state devices are left for the exponent of r^1. The exponent-part range is $r^{i[\pm.5(r^{n_i}-1)]}$, so the range of the exponent of r is $i[\pm.5(r^{n_i}-1)]$.

By equation (4.5) and the fact that q_1-1 is the smallest integer value of x satisfying $r^x > t^p-1$, we have

(4.6) $$i(q_i-1) \geq q_1-1$$

(4.7) $$n_1 = N - i\,q_i \leq N - (q_1-1) - i = n_1 + 1-i$$

We need consider only i for which $n_i > 0$ so that

(4.8) $$i < n_i + i \leq n_1 + 1$$

Applying Lemma I with k = 1, we find that r is the best $(p\text{-}D_t, N, r)$-base. But p and N were arbitrary, so base r is the best $(*\text{-}D_t, *, r)$-base (provided t and r are incommensurable).

Q.E.D.

5. The Base Conversion Theorem for Commensurable Bases.

We next discuss the case when r and t are commensurable, As mentioned earlier, we will find that if $t = r^j$ then r^j is the best $(*\text{-}D_t, *, r)$-base and so r is not always a best $(p\text{-}D_t, N, r)$-base. If $r = 2$ then $t = r^j$ is the only case left to be considered. But in general we must discuss the case $t^\rho = r^\tau$ in order to complete our results. First we must extend the Base Conversion Theorem to the case where t and r are commensurable.

Example 5.1: Consider the conversion from $S(2\text{-}D_{16})$ into $S(4\text{-}D_8)$. 16 and 8 are commensurable since $16^3 = 8^4$. The mapping $C: S(2\text{-}D_{16}) \to S(4\text{-}D_8)$ is (l-l) since

$$(o.a_1a_2)_{16} \times 16^{3n-2} = (o.b_1b_2\cdots b_8)_2 \times 2^{12n-8}$$

$$= (o.oob_1b_2\cdots b_8)_2 \times 8^{4n-2} \tag{5.1}$$

$$= (o.c_1c_2c_3c_4)_8 \times 8^{4n-2}$$

$$(o.a_1a_2)_{16} \times 16^{3n-1} = (o.b_1b_2\cdots b_8)_2 \times 2^{12n-4}$$

$$= (o.ob_1b_2\cdots b_8)_2 \times 8^{4n-1} \tag{5.2}$$

$$= (o.c_1c_2c_3)_8 \times 8^{4n-1}$$

$$(o.a_1a_2)_{16} \times 16^{3n} = (o.b_1b_2\cdots b_8)_2 \times 2^{12n}$$

$$= (o.c_1c_2c_3)_8 \times 8^{4n} \tag{5.3}$$

Since any number in $S(2\text{-}D_{16})$ can be written in the form $(o.a_1a_2)_{16} \times 16^{3n-k}$ for k=0, 1 or 2, the above shows that any element of $S(2\text{-}D_{16})$ is exactly expressible in $S(4\text{-}D_8)$, i.e. that $S(2\text{-}D_{16}) \subset S(4\text{-}D_8)$. Further, (5.1) shows that q = 4 is the smallest value of q for which $S(2\text{-}D_{16}) \subset S(q\text{-}D_8)$. The proof (given in the appendix) of the following theorem is nothing more than a generalization of the methods of this example.

Theorem IV. Suppose $t^{\rho} = r^{\tau}$ for some relatively prime positive integers ρ and τ. Let

$$\tau = c\,\rho + d \qquad \text{with} \quad 0 \le d < \rho \tag{594}$$

$$\tau\, p = x\,\rho + y \qquad \text{with} \quad 0 \le d < \rho \,. \tag{5.5}$$

The conversion mapping $C{:}S(p\text{-}D_t) \to S(q\text{-}D_r)$ is (l-l) if and only if

$$q \ge x + 6(d) + \delta(y\text{-}1) \tag{5.6}$$

where

$$\delta(n) = \begin{cases} 0 & \text{if } n < 0 \\ 1 & \text{if } n < 0 \end{cases} \tag{5.7}$$

Corollary III in the appendix shows that $C{:}S(q\text{-}D_r) \to S(p\text{-}D_t)$ is onto precisely when $C{:}S(p\text{-}D_t) \to S(q\text{-}D_r)$ is (1-l). Thus $S(q\text{-}D_r)$ is accurate to $p\text{-digits}_t$ precisely when (5.6) is satisfied.

6. The Best Base Theorem,

Theorem V. Base r^{τ} is the best $(*\text{-}D_{r^{\tau}}, *, r)$-base, for $\tau = 1, 2, \ldots$.

Base r^2 is the best $(p\text{-}D_t, *, r)$-base if and only if the following three conditions all hold:

(1) $t^{\rho} = r^{\tau}$ for some relatively prime integers ρ and τ

(2) ρ is odd, $\rho \ge 3$ and $\tau > 2$

(3) $\tau\, p = x\,\rho + 2$ for some integer x.

In these cases base r^2 affords twice the exponent range as does r. Otherwise (when at least one of (1)-(3) does not hold) r is the best (p-D_t, *, r)-base. If either (1) or (2) does not hold then r is the best (*-D_t, *, r)-base.

Example 6.1: An example in which r^2 is a better (p-D_t, N, r)-base than r should clarify matters. Such is the case when $r = 8, t = 16$ and $p = 2$. In this example we wish to decide which of 8 and 64 is the better base for the internal representation of floating-point numbers in an 8-ary computer with N-digits$_8$ per memory word. The constraints on this decision are (1) achieving the widest exponent range while (2) preserving 2-digit$_{16}$ accuracy (see Sec. 1). If base 8 were used then k-digits$_8$ would be used for the mantissa, since $q = 4$ is the smallest value of q for which $S(q\text{-}D_8)$ is accurate to 2-digits$_{16}$ (see Example 5.1). This leaves n_1=N-4 digits$_8$ for the exponent, affording an exponent-part range of $8^{[\pm.5(8^{n_1}-1)]}$ (or $8^{\pm(8^{n_1}-1)}$ for signed magnitude exponents). If base 64 were used, then, by Theorem IV, 2-digits$_{64}$ or equivalently, k-digits$_8$ would be needed for the mantissa, again leaving n_1-digits$_8$ for the exponent. This allows an exponent-part range of $64^{[\pm.5(8^{n_1}-1)]}$ (or $64^{\pm(8^{n_1}-1)}$ for signed magnitude exponents). Hence base 8^2 is better than base 8 here. In general, base r^2 is better than r precisely when $n_2 = n_1$

The method of proof for this theorem is essentially the same as that used for Theorem III. It is a more involved proof because some of the inequalities to be proved are more delicate.

Proof of Theorem V: See appendix.

7. Conclusion.

The Best Base Theorem shows that, in many cases, the choice of the base under constraints (i)-(iii) of Sec. 1 is independent of the variable p in constraint (ii). For example, this is the case when binary computers are under consideration (r=2) or when the target base is ten (t=10). In these cases constraints (ii) and (iii) can be replaced by

(ii)' if representations A and B have the same accuracy base t, for a given t, then A must be chosen over B if A gives a larger exponent range than B

without disturbing our results. And the Best Base Theorem states that the representation chosen will use (1) base r, if t is not a power of r, or (2) base t if t is a power of r. In these cases one need not know in advance how many digits$_t$ of accuracy are desired from floating-point representation in order to choose a base. One must know only that accuracy is to be measured with respect to base t.

When r=2 there are cases in which base 4 may be preferable to, but not better-than (in our formal sense) base 2. This occurs when the exponent range for base 4 is only slightly less than that for base 2; in the notation of Theorem III, this occurs when $n_2 = n_1 - 1$ and $2q_2 = q_1 + 1$, and n_1 is not small (say $n_1 \geq 5$). This was true in

the examples of Sec. 3. In these cases the exponent range for base 4 is $\left[4^{-2^{n_1}} \text{ to } 4^{2^{n_1}-1}\right]$ and that for base 2 is $\left[4^{-2^{n_1}} \text{ to } 4^{2^{n_1}-.5}\right]$. The transformation of representation from base 2 to base r in these cases is affected by transferring one bit from the exponent to the mantissa. And so the base 4 representation is just as accurate, base 2, as the base 2 representation, and more accurate, base 4, than the base 2 representation. This gain in accuracy more than makes up for the negligible loss in exponent range.*

Of course the choice of any base, r^j $(j \geq 1)$, for an r-ary computer can be justified via (i) and (ii)' by simply asserting that accuracy is to be measured with respect to base r^j. In general the author does not agree with such reasoning, because today's computer user is interested in base 10 accuracy. Of course base 10 is not sacrosanct, but the existence of a standard base is most valuable. It facilitates comparision of new results with old results and standardizes the form in which results are to be documented.

In practice, constraints other than (i) and (ii)' can arise. Internal data paths may make particular length exponents and mantissas or a particular base advantageous. For example, suppose that (due to some other considerations) an eight bit data path is selected for a proposed binary computer. Then it would be advantageous to have the exponent and mantissa each occupy a multiple of 8 bits (the mantissa

*The author is grateful to I. B. Goldberg for bringing this phenomenon to his attention. (It should be added that this does not occur if the normalization bit of the base 2 representation is made implicit.)

sign bit being included with either the exponent or the mantissa). Also fast shift instructions which shift the contents of a register four bits at a time (right or left) may be available on such a computer. If base 16 is used for the internal floating-point numbers, then normalization (and unnormalization done when the exponents of two numbers to be added are made equal) can be done quickly by these shift instructions. Also, such normalizations would be needed less often.[3] Such constraints would change our results considerably.

One could argue that constraint (i) has prejudiced us against larger bases. Certainly this is true. However, we would (mildly) argue against the use of fractions of digits $_{r}{}^{i}$ purely for aesthetics. Nevertheless, our analysis could be redone without this constraint by generalizing Mutala's results as follows. Let $r = s^{\rho}$, $t = w^{\tau}$ and define

$$(7.1)\quad S(\tfrac{q}{\rho}\text{-}D_r) = (X: |X| = \sum_{j=1}^{q} \alpha_j s^{-j} \times r^{n} \quad \text{where} \quad 0 \le \alpha_j < s \} .$$

We will say that $S(\frac{q}{\rho}\text{-}D_r)$ is accurate to $\frac{p}{\tau}\text{-digits}_t$ if and only if $C{:}S(\frac{p}{\tau}\text{-}D_t) \to S(\frac{q}{\rho}\text{-}D_r)$ is (1-1) and $C{:}S(\frac{q}{\rho}\text{-}D_r) \to S(\frac{p}{\tau}\text{-}D_t)$ is onto. We conjecture that the corresponding theorem in the incommensurable case is the following: $S(\frac{q}{\rho}\text{-}D_r)$ is accurate to $\frac{p}{\tau}\text{-digits}_t$ if and only if $s^{q-\rho} \ge w^{p}-1$. When ρ divides q and τ divides p, this reduces to the previous results.

We conclude with the following observation on the generality of our results: the Best Base Theorem and its proof, as given here, are valid for any definition of accuracy of the form "$S(q\text{-}D_r)$ is accurate to $p\text{-digits}_t$ if and only if $r^{q-1} \geq f(p,t)$," where f is an arbitrary function.

8. Appendix

Proofs of Theorems IV and V are given here.

Proof of Theorem IV: Equality between [the prime factorization of $t]^{\rho}$ and [the prime factorization of $r]^{\tau}$ implies the existence of an integer $s \geq 2$ satisfying

$$t = s \text{ and } r = s^{\rho} \tag{8.1}$$

We are discussing conversions from $S(p\text{-}D_{s^{\tau}})$ to $S(q\text{-}D_{s^{\rho}})$. We will consider the $p\text{-digit}_t$ numbers

$$z_n = (o.\alpha_1\alpha_2\cdots\alpha_p)_t \times t^n \qquad \text{with } \alpha_1 \neq 0 \tag{8.2}$$

$$= (o.\beta_1\beta_2\cdots\beta_{\tau p})_s \times s^{\tau n}$$

The proof is divided into two cases.

Case I. $\rho = 1$

In this case $d = y = 0$ and $r = s$ so z_n expressed in base r is

$$z_n = (o.\beta_1\beta_2\cdots\beta_{\tau p})_r \times r^{\tau n} . \tag{8.3}$$

Thus the conversion mapping $C: S(p\text{-}D_t) \to S(q\text{-}D_r)$ is (l-l) if and only if $q \geq \tau p = x = x + \delta(d) + \delta(y\text{-}1)$.

Observe that in this case, as in Example 5.1, C is (l-l) if and only if every element in $S(p\text{-}D_t)$ is expressible exactly in $S(q\text{-}D_r)$,

i.e. if and only if $S(p\text{-}D_t) \subset S(q\text{-}D_r)$. This means that C is (l-l) if and only if it is the identity map.

Case II $\rho > 1$

Let u_n, v_n, a_n and b_n be integers satisfying

(8.4) $$\tau n = u_n\rho - v_n \quad \text{with } 0 < v_n < \rho$$

(8.5) $$y + v_n = a_n\rho + b_n \quad \text{with } 0 < b_n < \rho \quad .$$

In this case

(8.6) $$z_n = (0.\beta_1\beta_2\cdots\beta_{x\rho+y})_s \times s^{u_n\rho - v_n}$$

$$= (\overbrace{0.0\ldots0}^{v_n}\ \beta_1\cdots\beta_{x\rho+y})_s \times r^{u_n}$$

where at least one of $\beta_1, \beta_2, \ldots, \beta_{c\rho+d}$ is non-zero. When z_n is converted to base r the first digit_r will be $\gamma_1 = (\beta_1 \cdot \bullet\ {}^*B_{\rho - v_n})_s$.

If $x = a_n = 0$ then we are done. Otherwise there are $x\rho+y - (p\text{-}v,) = (x+a_n\text{-}l)\rho+b_n$ digits_s left to be converted. Each of the next $(x+a_n\text{-}l)$ groups of $\rho\text{-digits}_s$ convert into $\gamma_i = (\beta_{j+1}\beta_{j+2}\cdots\beta_{j+\rho})_s$, where $j = (i\text{-}l)\rho - v_n$, for $i = 2, 3, \ldots, x + {}_n a$. If $b_n = 0$ then $\gamma_{x+a_n+1} = 0$. Otherwise $\gamma_{x+a_n+1} = (\beta_{x\rho+y-b_n+1}\cdots\beta_{x\rho+y})_s \times s^{\rho - b_n}$.

Thus z_n expressed in base r is

$$z_n = (0.\gamma_1 \cdots \gamma_{x+a_n+1})_r \times r^n \quad . \tag{8.7}$$

At most $[x + a_n + \delta(b_n)]$-digits$_r$ are needed to express z_n . When, say $\alpha_1 = t - 1$ for $i = 1,\ldots,p$ in equation (8.2), precisely $[x + a_n + \delta(b_n)]$-digits$_r$ are needed for z_n . If $y = 0$ or $y = 1$ then the exact conversions from $S(p\text{-}D_t)$ to base r requiring the most digits$_r$ occur when $[a_n = 0$ and $b_n \neq 0]$ or $[a_n = 1$ and $b_n = 0]$ (see (8.4) and (8.5)), since v_n can be made to take on any of the values 0, 1..., p-1 by varying n . (If $v_n = v_m$ then $\tau(m-n) = (u_m - u_n)\rho$ and so $m-n = k\rho$ for some $k \neq 0$. Thus $v_{n+1},\ldots,v_{n+\rho}$ are ρ distinct integers living between 0 and $\rho-1$.) It follows that the mapping $C:S(p\text{-}D_t) \to S(q\text{-}D_r)$ is (1-1) if and only if $q \geq x + 1 = x + 6(d) + \delta(y-1)$, when $y=0$ or $y=1$.

If $y \geq 2$ then conversions requiring the most digits$_r$ occur when $a_n=1$ and $b_n \neq 0$. Again $C:S(p\text{-}D_t) \to S(q\text{-}D_r)$ is (1-1) if and only if $q \geq x + 2 = x + 6(d) + \delta(y-1)$.

Q.E.D.

Corollary II:. If r and t are commensurable then $C:S(p\text{-}D_t) \to S(q\text{-}D_r)$ is (1-1) if and only if C is the identity mapping. Further, $C:S(p\text{-}D_t) \to S(q\text{-}D_r)$ is the identity mapping if and only if $S(q\text{-}D_r) \supset S(p\text{-}D_t)$.

Corollary III. Let t and r be as in Theorem III and let

(8.8) $\rho = c' \tau + d'$ with $o \leq d' < \tau$

(8.9) $\rho\, q = x' \tau + y'$ with $o \leq y' < \tau$

The conversion mapping $C:S(p\text{-}D_t) \to S(q\text{-}D_r)$ is onto if and only if

(8.10) $$p \geq x' + \delta(d') + \delta(y'-1) \ .$$

Proof:

$\Rightarrow$: If $C:S(p\text{-}D_t) \to S(q\text{-}D_r)$ is onto then $S(q\text{-}D_r) \subset S(p\text{-}D_t)$ (this is evident from the discussion of the conversion of the z_n in the proof of Theorem IV). This implies $C:S(q\text{-}D_r) \to S(p\text{-}D_t)$ is $(1\text{-}1)$.

$\Leftarrow$: If $C:S(q\text{-}D_r) \to S(p\text{-}D_t)$ is $(1\text{-}1)$ then $S(q\text{-}D_r) \subset S(p\text{-}D_t)$ and so $C:S(p\text{-}D_t) \to S(q\text{-}D_r)$ maps the subset $S(q\text{-}D_r)$ of $S(p\text{-}D_t)$ onto itself.

Q.E.D.

Corollary IV. For any t and r both > 2 the conversion $C:S(p\text{-}D_t) \to S(q\text{-}D_r)$ is both onto and $(1\text{-}1)$ if and only if $t = r$ and $p = q$.

Proof:

The case when t and r are incommensurable is covered by Corollary I. Suppose $t^\rho = r^\tau$ for some relatively prime positive integers $\rho = \tau$.

Case I $\rho = \tau = 1$

Here $t = r$ and so $p = q$ yields $S(p\text{-}D_t) = S(q\text{-}D_r)$.

$C: S(p\text{-}D_t) \to S(p\text{-}D_t)$ is both (1-1) and onto,

Case II $\rho \neq 1$ or $\tau \neq 1$ or both

The conversion is both onto and (1-1) if and only if

(8.11) $$p \geq x' + \delta(d') + \delta(y'-1)$$

(8.12) $$q \geq x + \delta(d) + \delta(y-1) \ .$$

which can be rewritten as

(8.13) $$\frac{p}{\rho} \geq \frac{q}{\tau} + \frac{1}{\rho}\left(\delta(d') + \delta(y'-1) - \frac{y'}{\tau}\right)$$

(8.14) $$\frac{q}{\tau} \geq \frac{p}{\rho} + \frac{1}{\tau}\left(\delta(d) + \delta(y-1) - \frac{y}{\rho}\right).$$

These imply

(8.15) $$0 \geq \frac{1}{\rho}\left(\delta(d') + \delta(y'-1) - \frac{y'}{\tau}\right) + \frac{1}{\tau}\left(\delta(d) + \delta(y-1) - \frac{y}{\rho}\right) ,$$

But the right side of (8.15) is positive since

i) $d \neq 0$ or $d' \neq 0$ or both;

ii) if $d = 0$ then $\rho = 1$ and $y = 0$ and so $\delta(d) + \delta(y-1) - \frac{y}{\rho} = 0$;

iii) if $d' = 0$ then $\tau = 1$ and $y' = 0$ and so $\delta(d') + \delta(y'-1) - \frac{y'}{\tau} = 0$;

iv) if $d \neq 0$ then $\delta(d) + \delta(y-1) - \frac{y}{\rho} > 0$;

v) if $d' \neq 0$ then $\delta(d') + \delta(y'-1) - \frac{y'}{\tau} > 0$.

Q.E.D.

Proof of Theorem V:

Theorem III takes care of the case when t and r are incommensurable. Here we consider the conversions. $C:S(p\text{-}D_t) \rightarrow S(q_i\text{-}D_{r^i})$ for r and t satisfying $t^\rho = r^\tau$. For i = 1, 2,..., let

$$g_i = \gcd(\tau, i) \quad , \quad \rho_i = \frac{i\,\rho}{g_i} \quad , \quad \tau_i = \frac{\tau}{g_i} \tag{8.16}$$

$$\tau_1 = c_i\rho_i + d_i \qquad \text{with} \quad 0 \leq d_i < \rho_i \tag{8.17}$$

$$\tau_i p = x_i\rho_i + y_i \qquad \text{with} \quad 0 \leq y_i < \rho_i \ . \tag{8.18}$$

Since $t^{i\rho} = (r^i)^\tau$, Theorem IV and its corollaries imply that the smallest value of q_i for which $S(q_i\text{-}D_{r^i})$ is accurate to p-digits$_t$ is

$$q_i = x_i + \delta(d_i) + \delta(y_i\text{-}1) \ . \tag{8.19}$$

Let us define

$$n_1 = N - i\,q_i \qquad \text{for } i = 1,2,\ldots \ . \tag{8.20}$$

We need consider only those i for which $n_i > 0$ and we assume that N is larger than $\min_{i \geq 1} i\,q_i$ so that there is some i for which $n_i > 0$. Equation (8.19) written in the form of (8.22) will prove useful:

$$q_i = \frac{\tau p - y_i g_i}{i\,\rho} + \delta(d_i) + \delta(y_i\text{-}1) \tag{8.21}$$

(8.22) $i \quad q_i = \frac{\tau p}{\rho} + i \left\{ \delta(d_i) + \delta(y_i - 1) - \frac{y_i}{\rho_i} \right\} \geq \frac{\tau p}{\rho}$.

Case I $\rho = 1$

In this case we prove that r^τ is the best $(*\text{-}D_{r^\tau}, *, r)$-base.

Here $d_k = 0$ if and only if $\frac{\tau}{g_k} = \tau_k = c_k \rho_k = \frac{c_k k}{g_k}$ which occurs if and only if k divides τ . Also $d_k = 0$ implies $y_k = 0$. Thus

(8.23) $kq_k = \tau p$ when k divides τ .

In particular, $n_\tau = N - \tau q_\tau = N - \tau p = \min_{i \geq 1} n_i > 0$ by (8.22). Thus r^τ is the best $(P\text{-}D_t, N, r)$-base among $r, r^2, \ldots, r^\tau$. If $i > \tau$ then $\rho_i = \frac{i}{g_i} > \frac{\tau}{g_i} = \tau_i$ and so $\tau_i = d_i \neq 0$ and

(8.24) $$iq_i \geq \tau p + i\left(1 - \frac{1}{\rho_i}\right) = \tau q_\tau + i - g_i$$

(8.25) $$-iq_i + i \leq -\tau q_r + g_i \leq -\tau q_\tau + \tau$$

(8.26) $i < n_i + i < n_\tau + \tau$ for i with $i > \tau$ and $n_i > 0$.

Applying Lemma I with $k = \tau$ shows that the exponent-part range for r^τ is strictly greater than that for r^i when $i > \tau$. This completes the proof that r^τ is the best $(*\text{-}D_{r^\tau}, *, r)$-base.

This also completes the proof for $r = 2$ (and for any other r which is not an integral power of an integer).

Case II $\rho > 1$

In this case $d_i \neq 0$ for $i = 1, 2, \ldots$ since

$d_i = 0$ implies $\frac{i\rho}{g_i} c_i = c_i \rho_i = \tau_i = \frac{\tau}{g_i}$ which implies $\tau = i c_i \rho$,

the last equation being impossible because $\gcd(\tau, \rho) = 1$. Also,

$r = s^\rho > 2^\rho > 4$. Equation (8.22) becomes

$$(8.27) \qquad i\, q_i = \frac{\tau p}{\rho} + i\left\{ + \delta(y_i - 1) - \frac{y_{ii}}{\rho_i} \right\} .$$

For brevity, let us define

$$(8.28) \qquad h_i = 1 + \delta(y_i - 1) - \frac{y_i}{\rho_i} \qquad \text{for } i = 1, 2, \ldots .$$

For a fixed i, h_i takes its minimum when $y_i = 1$ and its maximum when $y_i = 2$ and so

$$(8.29) \qquad 2\left(\frac{\rho_i - 1}{\rho_i}\right) \geq h_i \geq \left(\frac{\rho_i - 1}{\rho_i}\right) \qquad \text{for } i = 1, 2, \ldots .$$

Equations (8.27) to (8.29) imply

$$(8.30) \qquad n_i + i\, h_i = n_j + j\, h_j \qquad \text{for all } i, j$$

$$(8.31) \qquad n_i + i\left(\frac{\rho_i - 1}{\rho_i}\right) \leq n_1 + 2\left(\frac{\rho - 1}{\rho}\right) \qquad \text{for } i = 2, 3, \ldots .$$

But $\rho_i = \frac{i\rho}{g_i} \geq \rho$ and so $\left(\frac{\rho_i - 1}{\rho_i}\right) \geq \frac{\rho - 1}{\rho}$ and

(8.32) $$n_i \le n_1 - (i-2)(\rho-1)/\rho$$

(8.33) $$i\, r^{n_i} < r^{n_1}(i\, s^{-(i-2)(\rho-1)}) \quad \text{where } r = \frac{\rho}{s} .$$

The function $f(x) = x\, s^{-(x-2)(\rho-1)}$ is strictly decreasing for $x > 2$. Further $f(4) < 1$ and so

(8.34) $$i\, r^{n_i} < r^{n_1} \quad \text{for } i > 4 .$$

As shown in the proof of Lemma I, (8.34) implies that the exponent-part range for $r^{a_1 \cdots a_{n_1}}$ is strictly greater than that for $(r^i)^{b_1 \cdots b_{n_i}}$, for $i > 4$, where $a_1 \cdots a_{n_1}$ and $b_1 \cdots b_{n_i}$ are aribitrary n_1 and n_i-digit$_r$ integers, respectively. Thus the best $(p\text{-}D_t, N, r)$-bases are among r, r^2, r^3. Further if $\rho > 2$ then $f(3) < 3\, s^{-2} < 1$ and r^3 cannot be a best base.

Subcase IIa base r^3 versus base r for $\rho = 2$

We compare y and y_3 given by

(8.35) $$\tau p = 2X + y \qquad 0 \le y \le 1$$

(8.36) $$\frac{\tau}{g_3} p = \frac{6X_3}{g_3} + y_3 \qquad 0 \le y_3 < \frac{6}{g_3} .$$

From these we find that

(8.37) $$2(X - 3X_3) = g_3 y_3 - y .$$

Since $g_3 = \gcd(\tau,3)$ is odd, equation (8.37) implies that

(8.38) $[y=0 \Leftrightarrow y_3$ is even] and so $[h_1=1 \Leftrightarrow h_3 > 1]$

(8.39) $[y=1 \Leftrightarrow y_3$ is odd] and so $[h_1=\frac{1}{2} \Leftrightarrow h_3 \geq \frac{1}{2}]$.

The corresponding bounds for n_3 from (8.30) are

(8.40) $$n_3 \leq n_1 - 2 \quad \text{for (8.38)}$$

(8.41) $$n_3 \leq n_1 - 1 \quad \text{for (8.39)}$$

and these equations imply

(8.42) $$3r^{n_3} < r^{n_1}\frac{3}{r^2} < r^{n_1} \quad \text{for (8.38)}$$

(8.43) $$3r^{n_3} < r^{n_1}\frac{3}{r} < r^{n_1} \quad \text{for (8.39)}$$

and so base r^3 is never a best $(p\text{-}D_t, N, r)$-base.

Subcase IIb. base r^2 versus base r for odd τ

We must compare y and y_2 in

(8.44) $$\tau p = \chi \rho + y \qquad 0 \leq y < \rho$$

(8.45) $$\tau_2 p = \tau p = 2\chi_2\rho + Y2 \qquad 0 \leq y_2 < 2\rho \ .$$

The last equation is valid since τ is odd and so $g_2 = \gcd(\tau,2) = 1$.

From these equations we find that

(8.46) $$\rho(\chi - 2\chi_2) = y_2 - y = kp \quad \text{where } k = 0, \pm 1$$

and so ρ divides $y_2 - y$. Thus

(8.47) $[y=0 \Leftrightarrow y_2=0,\rho]$ and so $[h_1=1 \Leftrightarrow h_2=1,3/2]$

(8.48) $[y=1 \Leftrightarrow y_2=1,\rho+1]$ and so $[h_1=1-\frac{1}{\rho} \Leftrightarrow h_2=1-\frac{1}{2P},2-\frac{\rho+1}{2P}]$

(8.49) $[y>1 \Leftrightarrow Y2 \neq 0,1,\rho,\rho+1]$ and so $[h_1 \leq 2-\frac{2}{\rho} \Leftrightarrow h_2 \geq 2-\frac{2\rho-1}{2\rho}]$

(8.50) $[y \leq 1 \Leftrightarrow y_2=0,1,\rho,\rho+1]$ and so $[h_1 \leq 1 \Leftrightarrow h_2 \geq 3/4]$

where the last equation summarizes the first two. Corresponding bounds on n_2 from (8.30) are

(8.51) $$n_2 \leq n_1 - \frac{3}{\rho} \qquad \text{for (8.49)}$$

(8.52) $$n_2 \leq n_1 - \frac{1}{2} \qquad \text{for (8.50)}$$

and these equations imply

(8.53) $$2r^{n_2} \leq r^{n_1} \frac{2}{r^{3/\rho}} = r^{n_1} \frac{2}{s^3} < r^{n_1} \qquad \text{for (8.49)}$$

(8.54) $$2r^{n_2} < r^{n_1} \frac{2}{r^{1/2}} \leq r^{n_1} \qquad \text{for (8.50).}$$

So when ρ is even, base r is strictly better than base r^2 .

Subcase IIc base r^2 versus base r for odd $\rho > 3$ and even τ

We now characterize the situations in which r^2 is the best (p-D_t, N, r)-base. As in the other subcases, we compare y and y_2 in

(8.55) $\tau p = \chi\rho + y \qquad 0 \leq y < \rho$

(8.56) $\tau_2 p = \frac{\tau p}{2} = \chi_2\rho + Y2 \qquad 0 \leq y_2 < \rho$.

The last equation is valid since τ is even and so $g_2 = 2$. From these we derive

(8.57) $\rho(\chi-2\chi_2)=2y_2 - y = kp \qquad$ for $k = 0, \pm 1$.

Thus

(8.58) $[y=0 \Leftrightarrow y_2=0]$ and so $[h_1=1 \Leftrightarrow h_2=1]$

(8.59) $[y=1 \Leftrightarrow y_2= \frac{\rho+1}{2}]$ and so $[h_1=1- \frac{1}{\rho} \Leftrightarrow h_2=2- \frac{\rho+1}{2\rho}]$

(8.60) $[y=2 \Leftrightarrow y_2=1]$ and so $[h_1=2- \frac{2}{\rho} \Leftrightarrow h_2=1- \frac{1}{\rho}]$

(8.61) $[y>2 \Leftrightarrow y_2 \neq 0,1 \frac{\rho+1}{2}]$ and so $[h_1 \leq 2- \frac{3}{\rho} \Leftrightarrow h_2 \geq 1]$.

The corresponding relations between n_1 and n_2 are

(8.62) $n_2 = n_1 - 1 \qquad$ for (8.58)

(8.63) $n_2 = n_1 - 2 \qquad$ for (8.59)

(8.64) $n_2 = n_1 \qquad$ for (8.60)

(8.65) $n_2 \leq n_1 - \frac{3}{\rho} \qquad$ for (8.61) .

As shown in the previous subcases, all these conditions except (8.64) imply that base r is the best $(p\text{-}D_t, N, r)$-base; (8.64) implies that r^2 is the best $(p\text{-}D_t, N, r)$-base.

Q.E.D.

ACKNOWLEDGEMENT

This work is dedicated to my wife, Jackie, without whose patience and encouragement it would never have been completed. I would also like to thank Professor D. Gries for devising the notation $S(q\text{-}D_r)$ as well as for finding several blunders in the original draft.

References

[1] Goldberg, I. B., "27 Bits Are Not Enough for 8-digit Accuracy', Communications of the ACM, Vol. 10, No, 2, February, 1967, pp. 105-6.

[2] Matula, D. W., "Base Conversion Mappings", to appear in the Proceedings of the Fall Joint Computer Conference, 1967.

[3] Sweeney, D. W., "An Analysis of Floating Point Addition", IBM Systems Journal, Vol. 4, No. 1, 1965, pp. 31-42.

[4] Buchholz, W., Planning A Computer System, McGraw-Hill Book Co., Inc., 1962, pp. 42-59.