THE UNIVERSITY OF MICHIGAN

INDUSTRY PROGRAM OF THE COLLEGE OF ENGINEERING

AN INVESTIGATION OF THE ALGEBRAIC PROPERTIES OF THE RESIDUE NUMBER SYSTEM

Donald P. Rozenberg

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in the
University of Michigan
1961

June, 1961

IP-520

ACKNOWLEDGMENT

The author wishes to express his appreciation to the members of his committee for their assistance, criticisms, and suggestions during the course of the work.

The research upon which this dissertation is based was supported by the United States Air Force under contract AF33(6.6)-7340.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.1 Introduction to the Problem

The first large scale digital computer designed to incorporate the binary number system was the EDVAC developed at the Moore School of Electrical Engineering, University of Pennsylvania. The addition time of the EDVAC was one millisecond. Since 1949 when the design was described, the trend in digital computer technology has been toward higher speed, in particular drastically reduced addition time. One of the fastest digital computers under development is the IBM STRETCH which is to effect addition in 0.6 microseconds. This great arithmetic speed is obtained partly by using very fast components and partly by the novelties in the logical structure of the system. As yet new and unconventional number systems have not been employed to increase the speed of arithmetic operations in a large scale digital computer.

When two numbers represented in conventional number systems are added in a digital computer, carry propagation normally consumes the major portion of the computation time. The residue number system,[1,2] based on the algebra of residue classes, allows addition to be performed without carry computation. Furthermore, multiplication is as fast as addition. Cheney[3] has designed a digital correlator based on the residue number

---

1 H. L. Garner, "The Residue Number System." IRE Trans. Electronic Computers, Vol. EC-8, June 1959, pp.140-7.

2 A. Svoboda, "Rational Number Systems of Residual Classes." Stroje Na Zpracovani Informaci, Sbornik V; 1957.

3 P. W. Cheney, "A Digital Correlator Based on the Residue Number System." Technical Document LMSD-702670 Lockheed Aircraft Corporation, 1960.

system and concluded that a correlator of similar accuracy based on the binary system would have been 10 times slower if the organization were parallel and 100 times slower if serial.

The residue number system poses rather formidable problems which have prevented its adoption in the design of general purpose digital computers. These problems include (1) sign determination, (2) the prevention of additive and multiplicative overflow, (3) magnitude comparison, and (4) division. In this thesis the algebraic properties of the residue number system and related number systems are investigated and employed to provide insight into the above problems.

The remainder of this chapter will be devoted to the presentation of the residue number system as the direct sum of rings of integers. The above mentioned problems will be discussed.

Chapter II will contain the algebraic theory of the R-space, a pseudo-vector space of the residue representations. The R-space will provide the formulation of the residue number system and the associated number systems.

Carry functions which provide a basis for the addition and multiplication of the residue and related number systems are discussed in Chapter III.

The mixed base number system related to the residue number system is discussed in Chapter IV. It is shown that the mixed base number system possesses several crucial properties. These properties allow the solution of the problems of the residue number system.

Chapter V shows that there is no number system related to the residue number system which shares the special properties of the mixed base system.

This dissertation will investigate the problem of the residue number system and give solutions to the problems of (1) sign detection, (2) magnitude comparison, (3) overflow, and (4) division. These results will be obtained by considering the residue number system to be a pseudo-vector space.

1.2 The Residue Number System

As a preliminary step in the introduction of the residue number system the ring I_M of integers modulo M will be discussed.[4] The elements of the ring are the integers 0, 1, 2, ..., M-1. Ring addition and ring multiplication of two elements of I_M are performed by reducing the conventional sum and product modulo M. That is, the sum or product is divided by M and the remainder is retained as the result. Thus, for example, the ring I_6 consists of the integers 0, 1, 2, 3, 4, 5. Dividing the ordinary sum of 4 and 5 by 6, one obtains 3 for the remainder. Therefore, in I_6, $4 + 5 = 3$. Similarly $2 + 4 = 0$, $2 \cdot 2 = 4$, and $4 \cdot 2 = 2$. The proof that I_M is actually a ring follows from the fact that in the division of an integer by M, the remainder is unique. The complete proof may be found in the literature.[5] The order of I_M is M.

---

[4] N. H. McCoy, "Rings and Ideals." The Mathematical Association of America, Buffalo, New York, p. 1.

A Ring is a set of elements a, b, c, ..., a unique defined addition a + b, and a uniquely defined product ab satisfying the following five properties:

P_1 $a + (b+c) = (a+b) + c$;
P_2 $a + b = b + a$;
P_3 the equation $a + x = b$ has a solution x in R;
P_4 $a(bc) = (ab)c$;
P_5 $a(b+c) = ab + ac$, $(b+c)a = ba + ca$.

[5] Ibid., p. 64.

Consider the ring I of integers and the ring I_M of integers modulo M. An arbitrary element m of I determines a unique element $\bar{m}$ of I_M; namely, the remainder upon division by M. If one denotes the correspondence of m to $\bar{m}$ by $m \longleftrightarrow \bar{m}$, it is clear that if $m_1 \longleftrightarrow \bar{m}_1$, $m_2 \longleftrightarrow \bar{m}_2$, then $m_1 + m_2 \longleftrightarrow \overline{m_1 + m_2} = \bar{m}_1 + \bar{m}_2$ and $m_1 m_2 \longleftrightarrow \overline{m_1 m_2} = \bar{m}_1 \bar{m}_2$. Therefore, the correspondence between I and I_M is a homomorphism.

Fixed point digital computers do not operate upon the ring I of integers but rather on the ring I_M. If one thinks of the radix point as being on the right, then the largest number which can be represented is $M-1$. That the homomorphism is a many-to-one mapping is painfully apparent when overflow occurs and the most significant digit(s) are lost. Fortunately, one can detect such an overflow by sensing the carry from the most significant digit position. The ring I_M possesses an additive inverse of every element. If m is an element of I_M, $M - m$ is the additive inverse. Thus we map the integers less than $M/2$ onto the elements of I_M less than $M/2$ and the negatives greater than $-M/2$ onto the remaining elements of I_M. Thus the sign is readily determined and thereby magnitude comparisons can be performed. Since one can perform overflow detection, sign determination, and magnitude comparisons, division is possible.

Let S_1 and S_2 be two rings and consider the ordered pairs of symbols (s_1, s_2) where $s_1 \in S_1$ and $s_2 \in S_2$. If one defines addition and multiplication to be

$$(s_1, s_2) + (t_1, t_2) = (s_1 + t_1, s_2 + t_2)$$

and

$$(s_1, s_2)\,(t_1, t_2) = (s_1 t_1, s_2 t_2)$$

this set of ordered pairs becomes the ring S termed the direct sum of S_1 and S_2 and denoted $S_1 \dot{+} S_2$. In the above definition the operations $s_i + t_i$ and $s_i t_i$ are the ring operations of S_i.

An especially important theorem is the following:

Theorem 1. If a ring S has positive characteristic $n = n_1 \cdot n_2$ where n_1 and n_2 are greater than 1 and relatively prime, then S is isomorphic ($\cong$) to $S_1 \dot{+} S_2$ where S_i is a ring of characteristic n_i $(i = 1, 2)$.[6]

This theorem states, for instance, that I_6 is isomorphic to the direct sum $I_2 \dot{+} I_3$ with the following mapping

$$0 \rightarrow (0, 0) \qquad 3 \rightarrow (1, 0)$$
$$1 \rightarrow (1, 1) \qquad 4 \rightarrow (0, 1)$$
$$2 \rightarrow (0, 2) \qquad 5 \rightarrow (1, 2) .$$

Theorem 1 may be extended as follows:

Theorem 2. If the ring I_M has positive characteristic $M = m_1 m_2 \cdots m_n$ where the m_i are integers greater than 1 and pairwise relatively prime, then

$$I_M \cong I_{m_1} \dot{+} I_{m_2} \dot{+} \cdots \dot{+} I_{m_n} .$$

Proof: If $n = 2$, the result follows from Theorem 1. Assume the result for $n = K$ and consider

$$M = m_1 m_2 \cdots m_K m_{K+1} .$$

---

[6] B. L. van der Waerden, op. cit. p. 116. Theorem 1 is Theorem 28 in McCoy.

M may be written

$$M = m_1 m_2 \ldots m'_K$$

where $m'_K = m_K m_{K+1}$.

Thus we have

$$I_M \cong I_{m_1} \dot{+} I_{m_2} \dot{+} \ldots \dot{+} I_{m'_K} ,$$

but by Theorem 1,

$$I_{m'_K} \cong I_{m_K} \dot{+} I_{m_{K+1}} .$$

Therefore,

$$I_M \cong I_{m_1} \dot{+} I_{m_2} \dot{+} \ldots \dot{+} I_{m_{K+1}} .$$

Thereby, the conclusion is proved.

The direct sum of Theorem 2 defines the residue number system. Elements of I_M are mapped unto n-tuples of the direct sum (elements of the residue number system) according to the following scheme

$$x \longleftrightarrow (x \bmod m , x \bmod m , \ldots, x \bmod m).^7 \qquad (1\text{-}1)$$

If $x \longleftrightarrow (x_1, x_2, \ldots, x_n)$,

then

$$x + y \longleftrightarrow [(x_1 + y_1) \bmod m_1, \ldots, (x_n + y_n) \bmod m_n]$$

and

$$x \cdot y \longleftrightarrow [(x_1 y_1) \bmod m_1, \ldots, (x_n y_n) \bmod m_n] .$$

It is clear from both the example following Theorem 1 and Expression (1-1) that the components of the residue representations have no positional significance. Therefore, there is no direct means of

---

[7] $x \bmod m_i$ is defined to be the residue of x_i modulo m_i

comparing two elements to determine which is the image of the larger integer. Thus one cannot compare any element with the image of $M/2$ (or $\frac{M+1}{2}$ if M is odd) to learn sign. A result is that Euclidean division is not possible in the residue number system.

Consider the following examples of additions in the ring I_{210} with $m_1 = 2$, $m_2 = 3$, $m_3 = 5$, and $m_4 = 7$.

(a)
$$\begin{array}{rl} 209 & \longleftrightarrow (1,2,4,6) \\ 1 & \longleftrightarrow (1,1,1,1) \\ \hline 210 & \longleftrightarrow (0,0,0,0) \end{array}$$

(b)
$$\begin{array}{rl} 209 & \longleftrightarrow (1,2,4,6) \\ 105 & \longleftrightarrow (1,0,0,0) \\ \hline 104 & \longleftrightarrow (0,2,4,6) \end{array}$$

(c)
$$\begin{array}{rl} 105 & \longleftrightarrow (1,0,0,0) \\ 120 & \longleftrightarrow (0,0,0,1) \\ \hline 015 & \longleftrightarrow (1,0,0,1) \end{array}$$

(d)
$$\begin{array}{rl} 23 & \longleftrightarrow (1,2,3,1) \\ 162 & \longleftrightarrow (1,2,2,6) \\ \hline 185 & \longleftrightarrow (0,1,0,0) \end{array}$$

In example (a), the sum produces an overflow and each component ring indicates an overflow. The sum in (b) overflows but only one component overflows. In (c), overflow occurs but the component subrings do not indicate overflow. No overflow accompanies the addition in (d); however, overflow is present in each component. These examples indicate that overflow in the component subrings is unrelated to overflow of I_M. Similar statements may be made concerning multiplicative overflow.

CHAPTER II

THE R-SPACE

2.1 Basic Properties of the R-Space

Let any two residue representations be

$$x = (x_1, x_2, \ldots, x_n)$$

$$y = (y_1, y_2, \ldots, y_n)$$

then $x + y \triangleq [x_1 + y_1\ (m_1), \ldots, x_n + y_n\ (m_n)]$ where the m_j are pairwise relatively prime. The component x_i is said to be associated with the base modulus m_i. The residue number system representations form an additive Abelian group which is denoted by R^n. For S, a ring with identity, we select the integers and define multiplication by a scalar to be

$$ax \triangleq [ax_1(m_1), \ldots, ax_n(m_n)]$$

With these definitions it is seen that

a.) $ax \in R^n$,

b.) $a(x + y) = ax + ay$

for $a(x + y) = \{a[(x_1 + y_1)(m_1)]\ (m_1), \ldots,$

$$a[(x_n + y_n)(m_n)]\ (m_n)\}$$

$$= \{[(ax_1)(m_1) + (ay_1)(m_1)]\ (m_1), \ldots,$$

$$[(ax_n)(m_n) + (ay_n)(m_n)]\ (m_n)\}$$

$$= ax + ay,$$

c.) $(a+b)x = ax + bx$

since

$$(a + b)x = [(a + b)x_1\ (m_1), \ldots, (a + b)x_n\ (m_n)]$$

$$= \{[ax_1(m_1) + bx_1(m_1)]\ (m_1), \ldots,$$

$$[ax_n(m_n) + bx_n(m_n)]\ (m_n)\} \ ,$$

d.) $(ab)x = a(bx)$

for $(ab)x = [abx_1(m_1), \ldots, abx_n(m_n)]$

$= a[bx_1(m_1), \ldots, bx_n(m_n)]$

$= a(bx)$, and

e.) All elements of R^n are uniquely expressible as linear forms $a_1\epsilon_1 + \ldots + a_n\epsilon_n$ by means of n fixed basis elements with $0 \leq a_i < m_i$ where ϵ_j has one for its j-th component and zero for the other components.

The five properties which must be satisfied for R^n to be a n-dimensional vector space are a, b, c, d above, and:

e'.) All elements of R^n are uniquely expressible as linear forms $a_1u_1 + \ldots + a_nu_n$ by means of n fixed "basis elements" $u_1, \ldots, u_n$ and $a_i \in S$.[8]

This property is not satisfied, and thus R^n is not a true vector space. The pseudo-vector space R^n will be called the R-space and all vector space terms which follow will have an interpretation in the R-space which is analogous to the vector space definitions.

Consider the set of vectors $< \alpha_1, \ldots, \alpha_n >$ with each α_i having n components. Form a square array of the components by placing the components of α_i in the i-th row. If this array can be made triangular (specifically a lower traingular array) by reordering rows and columns,

[8] B. L. van der Waerden, "Modern Algebra." Frederick Ungar Publishing Company, New York, Vol. 1, p. 42.

the set $< \alpha_1, \ldots, \alpha_n >$ is termed <u>semi-triangular</u> and the reordered set termed <u>triangular</u>.

<u>Theorem 3.</u> If the set $\{\alpha_1, \ldots, \alpha_n\}$ is triangular and the elements on the principal diagonal are relatively prime to the associated base moduli, then any linear form $a_1\alpha_1 + a_2\alpha_2 + \ldots + a_n\alpha_n$ where the a_i are integers can be uniquely expressed as

$$\sum_{i=1}^{n} c_i \alpha_i \quad \text{where} \quad 0 \leq c_i < m_i \quad .$$

Proof: $k_{nn}c_n \equiv k_{nn}a_n \mod m_n$

where k_{ij} is the j-th component of α_i.[9]

Thus $c_n \equiv a_n \mod m_n$ since m_n is relatively prime to k_{nn}, and $0 \leq c_n < m_n$ is uniquely determined.

Assume the conclusion true for c_j for $j = m, m+1, \ldots, n$ and also that these c_j have been determined and consider the congruence

$$k_{m-1,\,m-1}\, c_{m-1} + k_{m,\,m-1}\, c_m + \ldots + k_{n,\,m-1}\, c_n \equiv$$

$$k_{m-1,\,m-1}\, a_{m-1} + k_{m,\,m-1}\, a_m + \ldots + k_{n,\,m-1}\, a_n \mod m_{m-1}.$$

By adding to both sides of the above congruence the additive inverse of $(k_{m,m-1}\, c_m + \ldots + k_{n,\,m-1}\, c_n) \mod m_{m-1}$, this congruence takes on the form $k_{m-1,\,m-1}\, c_{m-1} \equiv A \mod m_{m-1}$. Since $(k_{m-1,\,m-1},\, m_{m-1}) = 1$, c_{m-1} is uniquely determined mod m_{m-1} and

$$0 \leq c_{m-1} < m_{m-1}.$$[10]

---

[9] $a \equiv b \mod m$ (a is <u>congruent</u> to be modulo m) if and only if $m \mid a-b$ (m divides a-b or $a = b + km$).

[10] The greatest common divisor (d) of a and b is written $(a, b) = d$.

Theorem 3 is of key importance for it allows us to abandon the ring S and to concentrate our attentions on linear forms of triangular sets of vectors; namely, $\sum_{i=1}^{n} c_i \alpha_i$ where $< \alpha_1, \ldots, \alpha_n >$ is triangular and $0 \leq c_i < m_i$. Except where indicated the following discussion will be limited to sets of triangular vectors and linear combinations with restricted scalars.

Definition: The vectors $\alpha_1, \ldots, \alpha_n$ of a triangular array are <u>linearly independent</u> if and only if

$$c_1\alpha_1 + \ldots + c_n\alpha_n = 0 \quad \text{where} \quad 0 \leq c_j < m_j$$

implies $c_1 = c_2 = \ldots = c_n = 0$.

Otherwise the vectors $\alpha_1, \ldots, \alpha_n$ are termed <u>linearly dependent</u>.

<u>Theorem 4.</u> The triangular set of vectors $< \alpha_1, \alpha_2, \ldots, \alpha_n >$ is linearly independent if and only if each element on the principal diagonal is relatively prime to the associated base modulus.

Proof: Consider the equation

$$c_1\alpha_1 + \ldots + c_n\alpha_n = 0 . \tag{2-1}$$

Equation (2-1) is equivalent to the following simultaneous linear congruences

$$k_{nn}c_n \equiv 0 \bmod m_n$$

$$k_{ii}c_i + \sum_{j=i+1}^{n} k_{ji}\, c_j \equiv 0 \bmod m_i \quad \text{for} \quad i = n-1, \ldots, 1.$$

For $k_{nn}c_n \equiv 0 \bmod m_n$ to yield the unique solution $c_n = 0$, it is sufficient that $(k_{nn}, m_n) = 1$. ~~Assume~~ $(k_{\ell\ell}, m_\ell) = 1$ and $c_i = 0$ for $i = 2\ell+1, \ell+2, \ldots, n$. The ℓ-th congruence becomes $k_{\ell\ell}\, c_\ell \equiv 0 \bmod m_\ell$ and $c_\ell = 0$ is the unique solution. This proves the sufficiency of the hypothesis.

Assume r to be the least index for which $(k_{ii}, m_i) \neq 1$. Let $c_i = 0$ for $i > r$. The above congruences become

$$k_{rr}\, c_r \equiv 0 \bmod m_r \tag{2-2}$$

$$k_{ii}\, c_i + \sum_{j=i+1}^{r} k_{ji}\, c_j \equiv 0 \bmod m_i \quad \text{for} \quad i = r-1, r-2, \ldots, 1.$$

Congruence (2-2) may be solved with $c_r \neq 0$. Since $(k_{ii}, m_i) = 1$ for $i = r-1, r-2, \ldots, 1$, the remaining congruences assume the form

$$k_{ii}\, c_i \equiv A_i \bmod m_i$$

The condition $(k_{ii}, m_i) = 1$ for $i < r$ guarantees the existence of a solution for each congruence. This completes the proof.

Definition: A set of vectors $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ is said to <u>span</u> R^n if and only if there exists a set of coefficients c_i in the ranges $0 \leq c_i < m_i$ satisfying the equation

$$\sum_{i=1}^{n} c_i\, \alpha_i = r$$

for all $r \epsilon R^n$.

<u>Theorem 5.</u> For a triangular set of vectors $< \alpha_1, \ldots, \alpha_n >$ to span R^n it is necessary and sufficient that each element on the principal diagonal be relatively prime to the associated base modulus.

Proof: The existence of solutions c_i to the following congruences will be shown

$$k_{nn}\, c_n \equiv a_n \bmod m_n \tag{2-3}$$

and $$k_{ii} c_i + \sum_{j=i+1}^{n} k_{ji}\, c_j \equiv a_i \bmod m_i \quad \text{for all } a_i \text{ in the range}$$

$0 \leq a_i < m_i$. The necessary and sufficient condition for the existence of a solution to (2-3) is $(k_{nn}, m_n) | a_n$. Since a_n (an integer) will range $0 \leq a_n < m_n$, it is necessary and sufficient for $(k_{nn}, m_n) = 1$. Assume that solutions c_i exist for $i > \ell$. These solutions are substituted into the ℓ-th congruence yielding an expression of the form

$$k_{\ell\ell}\, c_\ell + D \equiv a_\ell \bmod m_\ell$$

or

$$k_{\ell\ell}\, c_\ell \equiv a_\ell + E \bmod m_\ell$$

Again $(a_\ell + E) \bmod m_\ell$ will include all integers in the range 0 through $m_\ell - 1$. Thus to guarantee solution it is necessary and sufficient that $(k_{\ell\ell}, m_\ell) = 1$. The proof is complete.

Corollary: The triangular set of vectors $< \alpha_1, \ldots, \alpha_n >$ is a spanning set of R^n if and only if it is an independent set.

Definition: A <u>basis</u> of an R-space is a linearly independent set of vectors which generate the R-space.

Theorem 5 may be rephrased in more conventional terms as follows:

<u>Theorem 6.</u> For a triangular set of vectors $< \alpha_1, \ldots, \alpha_n >$ to be a basis of R^n it is necessary and sufficient that each element on the principal diagonal of the array be relatively prime to the associated modulus.

<u>Theorem 7.</u> If $\alpha_1, \ldots, \alpha_n$ forms a basis for R^n, then every vector $\beta \in R^n$ has a unique expression

$$\beta = x_1 \alpha_1 + x_2 \alpha_2 + \ldots + x_n \alpha_n, \quad 0 \leq x_i < m_i.$$

Proof: It will be shown that the congruences which must be satisfied for the above expression to hold will provide x_j which are unique modulo m_j. Let $\beta = (b_1, b_2, \ldots, b_n)$.

The first congruence is $x_n k_{n,n} \equiv b_n \bmod m_n$. Since k_{nn} and m_n are relatively prime, x_n is uniquely specified modulo m_n. Assume $x_{m+1}, x_{m+2}, \ldots, x_n$ have been uniquely determined. Then to be considered is the congruence

$$k_{m,m} x_m + k_{m+1,m} x_{m+1} + \ldots + k_{n,m} x_n \equiv b_m \bmod m_m.$$

Add to each side the additive inverse modulo m_m of

$$(k_{m+1,m} x_{m+1} + \ldots + k_{n,m} x_n)$$

to obtain

$$k_{m,m} x_m \equiv B \bmod m_n .$$

Since $(k_{mm}, m_m) = 1$, x_m is uniquely determined modulo m_m.

Theorem 7 states that every residue number has unique coordinates relative to a given basis. Thus every basis serves to define a number system related to the residue number system. The chapters which follow will be devoted to the arithmetic properties of this class of number systems. Algorithms which permit addition, complementation, and multiplication will be developed. The relation of these number systems to the problems of the residue number system will be investigated. It has been proven by Garner and Arnold that only triangular arrays of vectors will span R^n.

2.2 Linear Transformations and Matrix Multiplication in the R-Space

Definition: A linear transformation $G: R^n \longrightarrow S^m$ of an R-space R^n into an R-space S^m is a transformation which satisfies

$$(\xi + \eta)\, G = \xi\, G + \eta\, G$$

$$(c\, \xi)\, G = c(\xi\, G) .$$

Theorem 8. If $\alpha_1, \ldots, \alpha_n$ is any basis of R^n and $\gamma_1, \ldots, \gamma_n$ are any vectors in S^m, then there is one and only one linear transformation $G: R^n \to S^m$ with

$$\alpha_1 G = \gamma_1$$

$$\ldots$$

$$\alpha_n G = \gamma_n .$$

This transformation is defined by

$$(x_1\alpha_1 + \ldots + x_n\alpha_n)G = x_1\gamma_1 + x_2\gamma_2 + \ldots + x_n\gamma_n .$$

Proof: Let $A = \{\alpha_1, \ldots, \alpha_n\}$ be a basis for R^n

and $B = \{\beta_1, \ldots, \beta_m\}$ be a basis for S^m,

then $\alpha_1 G = \gamma_1$

$$\alpha_2 G = \gamma_2$$

$$\ldots$$

$\alpha_n G = \gamma_n$, where $\gamma_i = a_{i1}\beta_1 + \ldots + a_{1m}\beta_m$ for $i = 1, \ldots, n$.

If

$$\xi = (x_1, \ldots, x_n)_A \in R^n ,$$

then

$\xi = x_1\alpha_1 + \ldots + x_n\alpha_n$, from which results

$$\xi G = x_1\gamma_1 + x_2\gamma_2 + \ldots + x_n\gamma_n .$$

By substituting for γ_i one obtains

$$\xi G = x_1 (a_{11}\beta_1 + \ldots + a_{1m}\beta_m)$$
$$+ x_2 (a_{21}\beta_1 + \ldots + a_{2m}\beta_m)$$
$$\ldots$$
$$+ x_n (a_{n1}\beta_1 + \ldots + a_{nm}\beta_m), \text{ and by rearranging terms,}$$

$$\begin{aligned}\xi G = {} & (x_1a_{11} + x_2a_{21} + \ldots + x_na_{n1})\beta_1 \\ & + (x_1a_{12} + x_2a_{22} + \ldots + x_na_{n2})\beta_2 \\ & \cdots \\ & + (x_1a_{1m} + x_2a_{2m} + \ldots + x_na_{nm})\beta_m \ .\end{aligned}$$

By Theorem 3 the image ξG can be uniquely expressed

$$\sum_{i=1}^{m} d_i\beta_i \quad \text{where} \quad 0 < d_i < m_i$$

and, therefore, $\xi G \in S^m$ and the transformation G is a transformation $R^n \to S^m$.

Let $\eta = (y_1, \ldots, y_n)_A$

then $\eta = y_1\alpha_1 + \ldots + y_n\alpha_n$.

Thus $\eta G = y_1\gamma_1 + \ldots + y_n\gamma_n$,

$$\xi + \eta = (x_1 + y_1)\alpha_1 + (x_n + y_n)\alpha_n$$

$$\begin{aligned}(\xi + \eta)G &= (x_1 + y_1)\gamma_1 + \ldots + (x_n + y_n)\gamma_n \\ &= x_1\gamma_1 + \ldots + x_n\gamma_n + y_1\gamma_1 + \ldots + y_n\gamma_n \\ &= \xi G + \eta G, \quad \text{and}\end{aligned}$$

$$c\xi = cx_1\alpha_1 + \ldots + cx_n\alpha_n$$

$$\begin{aligned}(c\xi)G &= cx_1\gamma_1 + \ldots + cx_n\gamma_n \\ &= c(\xi G) \ .\end{aligned}$$

Thus the transformation is linear. It is to be noted that it is not necessary that the set $\{\gamma_1, \ldots, \gamma_n\}$ be a triangular set.

Since every vector in R^n can be expressed uniquely as $x_1\alpha_1 + \ldots + x_n\alpha_n$ for $0 \leq x_i < m_i$, the transformation is single valued

and, therefore, no other transformation from R^n into S^m yields $\alpha_i G$. The proof is complete.

Let $A = \{\alpha_1, \ldots, \alpha_n\}$ be a basis for R^n and $B = \{\beta_1, \ldots, \beta_m\}$ be a basis for S^m. We then have the equations

$$\alpha_1 G = a_{11}\beta_1 + \ldots + a_{1m}\beta_m$$

$$\alpha_2 G = a_{21}\beta_1 + \ldots + a_{2m}\beta_m$$

$$\ldots$$

$$\alpha_n G = a_{n1}\beta_1 + \ldots + a_{nm}\beta_m$$

The form of these equations suggests writing the rectangular matrix

$$\mathcal{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & \cdots & & a_{2m} \\ & \cdots & & \\ a_{n1} & \cdots & & a_{nm} \end{bmatrix}$$

If R^n, S^n, and T^n are three n dimensional R-spaces, G is a transformation of R^n into S^n, and H is a linear transformation of S^n into T^n, then the product of the transformations is defined

$$\alpha(GH) = (\alpha G)H$$

Theorem 9. If the product of two transformations is defined, then it is a linear transformation.

Proof: Let G and H be linear transformations whose product exists, let c be a scalar, and let α_1, α_2 be vectors in the domain of G. Then we have

$$(\alpha_1 + \alpha_2)(GH) = [(\alpha_1 + \alpha_2)G]H$$

$$= [(\alpha_1 G) + (\alpha_2 G)]H$$

$$= (\alpha_1 G)H + (\alpha_2 G)H$$
$$= \alpha_1(GH) + \alpha_2(GH)$$

and

$$(c\alpha_1)(GH) = [(c\alpha_1)G]H$$
$$= [c(\alpha_1 G)]H$$
$$= c[(\alpha_1 G)]H$$
$$= c[\alpha_1(GH)] .$$

Multiplication of linear transformations has been defined and will now be used as a guide to defining multiplication of matrices. To this end let R^n, S^n, and T^n be three $R-$spaces of dimension n with respective bases

$$A = \{\alpha_1, \ldots, \alpha_n\}, \; B = \{\beta_1, \ldots, \beta_n\}, \text{ and } C = \{\gamma_1, \ldots, \gamma_n\} .$$

Relative to these bases G has the matrix $\mathcal{A}$ and H has the matrix $\mathcal{B}$.

Consider the matrix $P = ||p_{ij}||$ of the product transformation $J = GH$ relative to the bases for R^n and S^n. A development of the rows of P will be given in terms of a_{ij} and b_{ij}.

$$\alpha_1 G = a_{11}\beta_1 + \ldots + a_{1n}\beta_n$$
$$\alpha_2 G = a_{21}\beta_1 + \ldots + a_{2n}\beta_n$$
$$\ldots$$
$$\alpha_n G = a_{n1}\beta_1 + \ldots + a_{nn}\beta_n$$

and

$$\beta_1 H = b_{11}\gamma_1 + \ldots + b_{1n}\gamma_n$$
$$\beta_2 H = b_{21}\gamma_1 + \ldots + b_{2n}\gamma_n$$
$$\ldots$$
$$\beta_n H = b_{n1}\gamma_1 + \ldots + b_{nn}\gamma_n .$$

Thus $(\alpha_1 G)H = a_{11}(\beta_1 H) + a_{12}(\beta_2 H) + \ldots + a_{1n}(\beta_n H)$

$$= a_{11}(b_{11}\gamma_1 + b_{12}\gamma_2 + \ldots + b_{1n}\gamma_n)$$

$$+ a_{12}(b_{21}\gamma_1 + b_{22}\gamma_2 + \ldots + b_{2n}\gamma_n)$$

$$+ \ldots$$

$$a_{1n}(b_{n1}\gamma_1 + b_{n2}\gamma_2 + \ldots + b_{nn}\gamma_n) .$$

Therefore

$$(\alpha_1 GH) = \left(\sum_{k=1}^{n} a_{1k}b_{k1}, \sum_{k=1}^{n} a_{1k}b_{k2}, \ldots, \sum_{k=1}^{n} a_{1k}b_{kn}\right).$$

Similarly

$$(\alpha_2 GH) = \left(\sum_{k=1}^{n} a_{2k}b_{k1}, \sum_{k=1}^{n} a_{2k}b_{k2}, \ldots, \sum_{k=1}^{n} a_{2k}b_{kn}\right).$$

Thus

$$(\alpha_n GH) = \left(\sum_{k=1}^{n} a_{nk}b_{k1}, \sum_{k=1}^{n} a_{nk}b_{k2}, \ldots, \sum_{k=1}^{n} a_{nk}b_{kn}\right).$$

The above equations are preliminary results in the determination of the rows of P. Each of the linear forms indicated above must be expressed as linear forms with restricted coefficients. The linear form with restricted coefficients corresponding to $(\alpha_i GH)$ constitutes the i-th row of P.

The justification for restricting the discussion to the multiplication of matrices which correspond to linear transformations from n dimensional R spaces into n dimensional R spaces is the projected application of such multiplication. The principal application will be in the change of basis operation (conversion from one number system into another). In this application the linear transformation is an automorphism from R^n onto R^n.

In the interest of completeness, the result for the most general case will be given. Let R^V, S^W, and T^Z be three R-spaces with respective bases

$$\mathcal{A} = \{\alpha_1, \ldots, \alpha_v\}, \mathcal{B} = \{\beta_1, \ldots, \beta_w\}, \text{ and } \mathcal{C} = \{\gamma_1, \ldots, \gamma_z\} .$$

The linear transformations G and H are defined by

$$\alpha_i G = \sum_{\nu=1}^{w} a_{i\nu}\beta_\nu \quad (i = 1, 2, \ldots, v),$$

and

$$\beta_i H = \sum_{\mu=1}^{z} b_{i\mu}\gamma_\mu \quad (i = 1, 2, \ldots, w) .$$

Therefore,

$$(\alpha_i GH) = \Big(\sum_{\nu=1}^{w} a_{i\nu}\beta_\nu\Big)H = \sum_{\nu=1}^{w} a_{i\nu}(\beta_\nu H) = \sum_{\nu=1}^{w} a_{i\nu} \sum_{\mu=1}^{z} b_{\nu\mu}\gamma_\mu$$

$$= \sum_{\mu=1}^{z} \sum_{\nu=1}^{w} a_{i\nu}b_{\nu\mu}\gamma_\mu = \Big(\sum_{k=1}^{w} a_{ik}b_{k1}, \sum_{k=1}^{w} a_{ik}b_{k2}, \ldots, \sum_{k=1}^{w} a_{ik}b_{kz}\Big).$$

The above linear form when expressed with restricted coefficients constitutes the i-th row of the product matrix.

If $A = ||a_{ij}||$, $B = ||b_{ij}||$, and $C = ||c_{ij}||$ are n x n matrices relative to the natural basis, the product (AB)C is developed in the following manner:

$AB = E = ||e_{ij}||$ relative to the natural basis and the i-th row of E is obtained from the linear form

$$\Big(\sum_{k=1}^{n} a_{ik}b_{k1}, \sum_{k=1}^{n} a_{ik}b_{k2}, \ldots, \sum_{k=1}^{n} a_{ik}b_{kn}\Big) .$$

Since this linear form is to be reduced relative to the natural basis, the i-th row of E is

$$[(\sum_{k=1}^{n} a_{ik}b_{k1})(m_1), (\sum_{k=1}^{n} a_{ik}b_{k2})(m_2), \ldots, (\sum_{k=1}^{n} a_{ik}b_{kn})(m_n)] .$$

The product $(AB)C = F = ||f_{ij}||$ is similarly developed and the i-th row of F is

$$[(\sum_{k=1}^{n} e_{ik}c_{k1})(m_1), \ldots, (\sum_{k=1}^{n} e_{ik}c_{kn})(m_n)]$$

$$= \{\sum_{k=1}^{n} [(\sum_{r=1}^{n} a_{ir}b_{bk})(m_k)] c_{k1}\} (m_1), \ldots,$$

$$\{\sum_{k=1}^{n} [(\sum_{r=1}^{n} a_{ir}b_{rk})(m_k)] c_{kn}\} (m_n) . \qquad (2\text{-}4)$$

Consideration of the product BC gives rise to the following linear form with restricted coefficients.

$$[(\sum_{k=1}^{n} b_{ik}c_{k1})(m_1), \ldots, (\sum_{k=1}^{n} b_{ik}c_{kn})(m_n)] .$$

Also

$A(BC) = D = ||d_{ij}||$ where the i-th row of D is

$$[(\sum_{r=1}^{n} a_{ir} \sum_{k=1}^{n} b_{rk}c_{k1})(m_1), \ldots, (\sum_{r=1}^{n} a_{ir} \sum_{k=1}^{n} b_{rk}c_{kn})(m_n)] . \qquad (2\text{-}5)$$

As shown by Equation (2-4) and (2-5), multiplication of matrices associated with the residue system is not associative, for the ranges of the components in (2-4) and (2-5) differ.

CHAPTER III

CARRY FUNCTIONS IN RESIDUE NUMBER SYSTEM AND RELATED NUMBER SYSTEMS

3.1 The Carry Algorithm

In the second chapter many of the results depended upon the existence of a linear form with restricted coefficients which was equivalent to a given linear expression. This chapter will discuss an algorithm (the Carry Algorithm) for finding the linear form with restricted coefficients when any linear combination of the basis vectors $\{\alpha_1, \ldots, \alpha_n\}$ is given. The algorithm involves the notion of carries from some components of the representations into other components of the representation.

If the linear form to be reduced is $a_1\alpha_1 + a_2\alpha_2 + \ldots + a_n\alpha_n$, one proceeds by expressing $a_n\alpha_n$ as $b_1\alpha_1 + b_2\alpha_2 + \ldots + b_n\alpha_n$ where $0 \leq b_i < m_i$ and making the substitution to obtain $(a_1 + b_1)\alpha_1 + \ldots + b_n\alpha_n$. The process is then repeated focusing attention on $(b_{n-1} + a_{n-1})\alpha_{n-1}$ and continued until one has the desired result.

Dividing a_j by m_j, one obtains $a_j = m_jq + r$, $0 \leq r < m_j$. Any multiple of $m_j\alpha_j$ will yield a vector in the subspace $(x_1, x_2, \ldots, x_{j-1}, 0, \ldots, 0)$. Since the set of vectors $\{\alpha_i\}$ is a basis, the set $\{\alpha_1, \ldots, \alpha_j\}$ is a basis of the mentioned subspace by Theorem 4. Thus the term m_j will not affect the multiplier of α_j in the reduced expression; that coefficient must be r. The product $m_j\alpha_j$ can be expressed as a linear combination of α_1 through α_{j-1} and $qm_j\alpha_j$ is merely q times each term of that combination. The linear combination for $(q\, m_j)\alpha_j$ is added to the original linear combination and $a_j\alpha_j$ is replaced by

$r\alpha_j$. Thus one applies the above procedure to first a_n, then to the new coefficient of α_{n-1}, and so on until the desired result if obtained.

The Carry Algorithm may be stated as follows:

1. Express $m_j\alpha_j$ as a linear combination of $\alpha_1, \ldots, \alpha_{j-1}$ denoted $b_{j1}\alpha_1 + b_{j2}\alpha_2 + \ldots + b_{jj-1}\alpha_{j-1}$ for $j = 2,3,\ldots,n$. Beginning with $j = n$ and $a_n' = a_n$
2. Divide a_j' by m_j and obtain $a_j' = q_j m_j + r_j$ with $0 \leq r_j < m_j$.
3. Replace $a_j'\alpha_j$ with $r_j\alpha_j$ and a_i' with $(a_i' + q_i b_{ji})$ for $i = 1, 2, \ldots, j-1$.
4. Repeat steps 2 and 3 substituting $j-1$ for j. Stop after executing steps 2 and 3 with $j = 1$.

Theorem 10. The linear form produced by the application of the Carry Algorithm expresses the same residue number as the original linear combination.

Proof: The proof will be the demonstration that the coefficients of the resulting linear form satisfy the congruence indicated in the proof of Theorem 3.

Consider $c_n \equiv a_n \bmod m_n$ with $0 \leq c_n < m_n$.

By the uniqueness of division r_n is the residue of a_n modulo m_n and $r_n = c_n$.

Assume that $r_j = c_j$ for $j = m, m+1, \ldots, n$. The congruence which then must be satisfied is

$$k_{m-1,\,m-1}\,c_{m-1} + k_{m,\,m-1}\,c_m + \ldots + k_{n,\,m-1}\,c_n \equiv k_{m-1,\,m-1}\,a_{m-1} + k_{m,\,m-1}\,a_m + \ldots + k_{n,\,m-1}\,a_{m-1} \bmod m_{m-1}. \tag{3-1}$$

The quantity a'_j which enters the division in step 2 of the Carry Algorithm is

$$a'_j = a_j + q_n b_{nj} + q_{n-1} b_{n-1j} + \cdots + q_{j+1} b_{j+1,\,j}$$

thus

$$a'_n = a_n$$

$$a'_{n-1} = a_{n-1} + q_n b_{n,\,n-1}$$

$$a'_{n-2} = a_{n-2} + q_n b_{n,\,n-2} + q_{n-1}\, b_{n-1,\,n-2}$$

$$\cdots$$

$$a'_{m-1} = a_{m-1} + q_n b_{n,\,m-1} + \cdots + q_m b_{m,\,m-1}\,.$$

In expressing

$$p_j \alpha_j = b_{j1}\alpha_1 + b_{j2}\alpha_2 + \cdots + b_{jj-1}\alpha_{j-1}\,,$$

one solved the congruences

$$k_{j-1j-1}\, b_{jj-1} \equiv m_j k_{jj-1} \bmod m_{j-1}$$

$$k_{j-2j-2}\, b_{jj-2} + k_{j-1j-2}\, b_{jj-1} \equiv m_j\, k_{jj-2} \bmod m_{j-2}$$

$$\cdots$$

$$k_{m-1,m-1}\, b_{jm-1} + k_{m-2,m-1}\, b_{jm-2} + \cdots + k_{j-1,m-1}\, b_{jn-1}$$

$$\equiv m_j\, k_{j,m-1} \bmod m_{m-1}$$

$$\cdots$$

$$k_{11} b_{j1} + k_{21} b_{j2} + \cdots + k_{j-1,1}\, b_{jj-1} \equiv m_j\, k_{j1} \bmod m_1\,.$$

Using the induction assumption, relation (3-1) can be written

$$k_{m-1,m-1} c_{m-1} + k_{m,m-1} r_m + \cdots + k_{n,m-1} r_n$$

$$\equiv k_{m-1,m-1} a_{m-1} + k_{m,m-1} a_m + \cdots + k_{n,m-1} a_n \bmod m_{m-1}$$

which can be manipulated to yield

$$k_{m-1,m-1} c_{m-1} + k_{m,m-1} r_m + \cdots + k_{n-1,m-1} r_{n-1}$$

$$\equiv k_{m-1,m-1} a_{m-1} + k_{m,m-1} a_m + \cdots + k_{n,m-1}\,(a_n - r_n) \bmod m_{m-1}\,. \tag{3-2}$$

Since $a_n = a'_n$, $a_n - r_n = q_n m_n$, (3-2) becomes

$$\begin{aligned} & k_{m-1,m-1}c_{m-1} + k_{m,m-1}r_m + \cdots + k_{n-1,m-1}r_{n-1} \\ & \equiv k_{m-1,m-1}\,(a_{m-1} + q_n b_{n,m-1}) + k_{m,m-1}\,(a_m + q_n b_{nm}) \\ & + \cdots + k_{n-1,m-1}\,(a_{n-1} + q_n b_{n,n-1}) \bmod m_{m-1} \end{aligned} \tag{3-3}$$

upon the substitution

$$\begin{aligned} & q_n\,(k_{m-1,m-1}b_{n,m-1} + k_{m,m-1}b_{nm} + \cdots + k_{n-1,m-1}b_{n,n-1}) \\ & \equiv q_n m_n k_{n,m-1} \pmod{m_{m-1}}. \end{aligned}$$

Once again we transpose (add the inverse) $k_{n-1,m-1}r_{n-1}$ and recognize that

$$a_{n-1} + q_n b_{nn-1} - r_{n-1} = a'_{n-1} - r_{n-1} = q_{n-1}m_{n-1}\,.$$

Making the following substitution:

$$\begin{aligned} & q_{n-1}\,(k_{m-1,m-1}b_{n-1,m-1} + k_{m,m-1}b_{n-1,m} + \cdots + k_{n-1,m-1}b_{n-1,n-2} \\ & \equiv q_{n-1}m_{n-1}k_{n-1,m-1} \bmod m_{m-1}, \end{aligned}$$

we obtain

$$\begin{aligned} & k_{m-1,m-1}c_{m-1} + k_{m,m-1}r_m + \cdots + k_{n-2,m-1}r_{n-2} \\ & \equiv k_{m-1,m-1}\,(a_{m-1} + q_n b_{n,m-1} + q_{n-1}\,b_{n-1,m-1}) \\ & + k_{m,m-1}\,(a_m + q_n b_{n,m} + q_{n-1}b_{n-1,m}) + \cdots \\ & + k_{n-2,m-1}\,(a_{n-2} + q_n b_{n,n-2} + q_{n-1}b_{n-1,n-2}) \bmod m_{m-1}. \end{aligned}$$

Again we identify the last quantity in parenthesis as a'_{n-2}, transpose and substitute. By repeating these manipulations, one finally obtains

$$k_{m-1,m-1}c_{m-1} \equiv k_{m-1,m-1}\,(a_{m-1} + q_n b_{n,m-1} + \cdots + q_m b_{m,m-1}) \bmod m_{m-1}$$

or

$$c_{m-1} \equiv a'_{m-1} \bmod m_{m-1}$$

which yields the desired result $c_{m-1} = r_{m-1}$.

By showing that $r_m = c_m$, we have shown that the linear form produced by the Carry Algorithm is identical to the linear form of the conclusion of Theorem 3. Therefore the proof is completed.

Without the Carry Algorithm, one can perform operations such as addition, multiplication, and matrix multiplication by resorting to the defined operations of addition and scalar multiplication of residue numbers. The only alternative is to solve a set of simultaneous linear congruences every time one wishes to express a linear form with restricted coefficients.

With the Carry Algorithm, it is necessary to solve only n sets of simultaneous linear congruences, and further those sets of congruences may be solved immediately upon the selection of the base moduli and the basis vectors. Thereafter, any linear form may be reduced to one with restricted coefficients by the application of steps 2, 3, and 4 of the Carry Algorithm. It is thus possible to select a set of basis vectors, perform step 1 of the reduction algorithm (i.e., determine the carry functions), and thereafter perform addition of two vectors by addition of the components followed by the application of the Carry Algorithm. Scalar multiplication is effected by multiplying each component of the vector by the scalar and then applying the Carry Algorithm. The Carry Algorithm will prove significant in the multiplication of representations for it will provide a means of combining partial products.

3.2 The Borrow Algorithm and Complementation

The question arises: Given two representation X and Y of positive integers, how does one obtain the remainder $X - Y$? This question will now be considered in some detail.

Let X be represented by $(x_1, x_2, \ldots, x_n)\alpha$ and Y by $(y_1, y_2, \ldots, y_n)\alpha$. Initially one forms $(x_1 - y_1, x_2 - y_2, \ldots, x_n - y_n)$. This last expression denotes $(x_1 - y_1)\alpha_1 + (x_2 - y_2)\alpha_2 + \ldots + (x_n - y_n)\alpha_n$. Consider the j-th component of an expression to be negative. Since

$$m_j\alpha_j = b_{j1}\alpha_1 + b_{j2}\alpha_2 + \ldots + b_{j,j-1}\alpha_{j-1},$$

one may add to the above expression zero in the form

$$d_j\,[m_j\alpha_j - (b_{j1}\alpha_1 + b_{j2}\alpha_2 + b_{j,j-1}\alpha_{j-1})] = 0,$$

where d_j is the smallest multiple of m_j which is larger than the magnitude of the j-th component. This addition yields

$$(x_1 - y_1 - d_n b_{n1},\ x_2 - y_2 - d_n b_{n2},\ \ldots,\ x_{n-1} - y_{n-2} - d_n b_{n,n-1},\ x_n - y_n + d_n m_n)$$

for $j = n$,

$$(x_1 - y_1 - d_n b_{n1} - d_{n-1} b_{n-1,1},\ x_2 - y_2 - d_n b_{n2} - d_{n-1} b_{n-1,2},\ \ldots,\ x_{n-1} - y_{n-1} - d_n b_{n,n-1} + d_{n-1} m_{n-1},\ x_n - y_n + d_n m_n)$$

for $j = n-1$, and finally

$$(x_1 - y_1 - d_n b_{n1} - d_{n-1} b_{n-1,1} - \ldots - d_2 b_{21} + d_1 m_1,\ x_2 - y_2 - d_n b_{n2} - d_{n-1} b_{n-1,2} - \ldots - d_3 b_{32} + d_2 m_2,\ \ldots,\ x_{n-1} - y_{n-1} - d_n b_{n,n-1} + d_{n-1} m_{n-1},\ x_n - y_n + d_n m_n)$$ for $j = 1$.

This final expression will be a linear form with restricted coefficients which is the representation of $X - Y$ if $X \geq Y$. If $X < Y$ the above procedure yields a representation of $-(Y - X)$. Thus complementation quite naturally enters the picture.

By dividing the range of integers which can be represented by residue numbers into those integers less than $M/2$ and those integers larger than $M/2$, one can designate the first group of vectors to be representations of positive integers and the second group, codings for the

complements of the elements of the first group. If M is even, $M/2$ is self complement.

To find the complement of a given representation Z, one generates the remainder $(0 - Z)$ as indicated above.

The procedure for performing subtraction and finding complements indicated above suggests the statement of a Borrow Algorithm as follows:

1. Express $m_j\alpha_j$ as a linear combination of $\alpha_1, \ldots, \alpha_{j-1}$ denoted $b_{j1}\alpha_1 + b_{j2}\alpha_2 + \ldots + b_{j,j-1}\alpha_{j-1}$ for $j = 2, 3, \ldots, r$. Beginning with $j = n$ and $a'_j = a_n$.
2. Perform steps 3 and 4 if $a'_j < 0$, otherwise skip to step 5.
3. Determine d_j such that d_j is the smallest multiple of m_j which is larger than $|a'_j|$.
4. Add to the linear form the expression
$$(-d_jb_{j1}, -d_jb_{j2}, \ldots, -d_jb_{jj-1}, +d_jm_j, 0, \ldots, 0).$$
5. Repeat steps 2, 3, and 4 substituting $j-1$ for j. Stop after executing steps 2, 3, and 4 for $j = 1$.

Example: With the basis $< (1,0,0,0), (1,2,0,0), (1,1,2,0), (1,1,1,1) >$ where the moduli are $m_1 = 2$, $m_2 = 3$, $m_3 = 5$ and $m_4 = 7$, the carry functions are $3\alpha_2 = \alpha_1$, $5\alpha_3 = \alpha_2$, and $7\alpha_4 = \alpha_3$. An example of subtraction using the Borrow Algorithm directly will be given as well as subtraction by using the complement of the substrahend.

Subtraction using the borrow algorithm:

$$
\begin{array}{rl}
 & (\ 1,\ 2,\ 2,\ 1) \longleftrightarrow 190 \\
- & (\ 0,\ 2,\ 4,\ 6) \longleftrightarrow -104 \\
\hline
 & (\ 1,\ 0,-2,-5) \\
+ & (\ 0,\ 0,-1,\ 7) \\
\hline
 & (\ 1,\ 0,-3,\ 2) \\
+ & (\ 0,-1,\ 5,\ 0) \\
\hline
 & (\ 1,-1,\ 2,\ 2) \\
+ & (-1,\ 3,\ 0,\ 0) \\
\hline
 & (\ 0,\ 2,\ 2,\ 2) \longleftrightarrow 86
\end{array}
$$

Complement of $(0,2,4,6)$:

$$
\begin{array}{rl}
 & (\ 0,-2,-4,-6) \\
+ & (\ 0,\ 0,-1,\ 7) \\
\hline
 & (\ 0,-2,-5,\ 1) \\
+ & (\ 0,-1,\ 5,\ 0) \\
\hline
 & (\ 0,-3,\ 0,\ 1) \\
+ & (-1,-3,\ 0,\ 0) \\
\hline
 & (-1,\ 0,\ 0,\ 1) \\
+ & (\ 2,\ 0,\ 0,\ 0) \\
\hline
 & (\ 1,\ 0,\ 0,\ 1)
\end{array}
$$

Subtraction utilizing the complement:

$$
\begin{array}{rl}
 & (\ 1,\ 2,\ 2,\ 1) \longleftrightarrow 190 \\
+ & (\ 1,\ 0,\ 0,\ 1) \longleftrightarrow -\ 104 \\
\hline
 & (\ 0,\ 2,\ 2,\ 2) \longleftrightarrow 86
\end{array}
$$

3.3 Change of Basis

One quite important use of the carry algorithm and matrix multiplication is the change of basis operation. Quite often it is desirable to convert from one number system to another, i.e., express a vector in coordinates relative to a different set of basis vectors. One might wish to make the conversion because different number systems are more advantageous for particular operations than others. More will be said concerning this later.

Let a vector X be represented by $(x_1, x_2, \ldots, x_n)$ relative to the basis $< \alpha_1, \alpha_2, \ldots, \alpha_n >$. It is desired to find coordinates $(y_1, y_2, \ldots, y_n)$, relative to the basis $< \beta_1, \beta_2, \ldots, \beta_n >$. Each vector α_1 of the old basis can be expressed as a linear combination of the vectors of the new basis in the form

$$\alpha_i = q_{i1}\beta_1 + q_{i2}\beta_2 + \ldots + q_{in}\beta_n \; . \qquad (3\text{-}4)$$

However, since both bases are triangular, $q_{ik} = 0$ for $k > i$. The vector X with coordinates $(x_1, x_2, \ldots, x_n)$ relative to the basis

$$< \alpha_1, \alpha_2, \ldots, \alpha_n > \text{ is } x_1\alpha_1 + x_2\alpha_2 + \ldots + x_n\alpha_n \; .$$

Substituting from above, one obtains for X

$$x_1q_{11}\beta_1 + x_2 [q_{21}\beta_1 + q_{22}\beta_2] + \ldots + x_n [q_{n1}\beta_1 + \ldots + q_{nn}\beta_n]$$

which can be written

$$(x_1q_{11} + x_2q_{21} + \ldots + x_nq_{n1})\beta_1$$
$$+ (x_2q_{22} + \ldots + x_nq_{n2})\beta_2 + \ldots + x_nq_{nn}\beta_n \; .$$

The carry algorithm is then applied to linear form (3-5) and the result is $X = y_1\beta_1 + y_2\beta_2 + \ldots + y_n\beta_n$.

The y_i are the coordinates of X relative to the basis $< \beta_1, \beta_2, \ldots, \beta_n >$.

If the zero coefficients are retained in Expression (3-4), Expression (3-5) becomes

$$(x_1q_{11} + x_2q_{21} + \ldots + x_nq_{n1})\beta_1 + (x_1q_{12} + x_2q_{22} + \ldots + x_nq_{n2})\beta_2 + \ldots + (x_1q_{1n} + x_2q_{2n} + \ldots + x_nq_{nn})\beta_n .$$

This expression is an interpretation of

$$(x_1q_{11} + x_2q_{21} + \ldots + x_nq_{n1}),\ (x_1q_{12} + x_2q_{22} + \ldots + x_nq_{n2}),\ \ldots\ (x_1q_{1n} + \ldots + x_nq_{nn})$$

which is the matrix product

$$[x_1, x_2, \ldots, x_n] \cdot \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \cdots & & & \\ q_{n1} & q_{n2} & \cdots & q_{nn} \end{bmatrix} = [y_1, y_2, \ldots, y_n]$$

denoted $X \cdot Q = Y$.

The above procedure constitutes an effective procedure for executing change of basis.

The vectors β_i of the new basis can be written as linear combinations of the old vectors,

$$\beta_i = \sum_{j=1}^{n} p_{ij}\alpha_j \qquad (3\text{-}6)$$

Thus for a change of basis from $< \beta_1, \ldots, \beta_n >$ to $< \alpha_1, \ldots, \alpha_n >$, the appropriate matrix product is $Y \cdot P = X$ where $P = ||P_{ij}||$.

Substituting Equation (3-6) into Equation (3-4) one obtains

$$\alpha_i = q_{i1} \sum_{j=1}^{n} p_{1j}\alpha_j + q_{i2} \sum_{j=1}^{n} p_{2j}\alpha_j + \cdots + q_{in} \sum_{j=1}^{n} p_{nj}\alpha_j$$

$$= \sum_{j=1}^{n} q_{i1}p_{1j}\alpha_j + \sum_{j=1}^{n} q_{i2}p_{2j}\alpha_j + \cdots + \sum_{j=1}^{n} q_{in}p_{nj}\alpha_j$$

$$= \sum_{k=1}^{n} \sum_{j=1}^{n} q_{ik}p_{kj}\alpha_j \quad .$$

One may interchange summations to obtain

$$\alpha_i = \sum_{j=1}^{n} \sum_{k=1}^{n} q_{ik}p_{kj}\alpha_j$$

$$= \sum_{k=1}^{n} q_{ik}p_{kj}\alpha_1 + \sum_{k=1}^{n} q_{ik}p_{k2}\alpha_2 + \cdots + \sum_{k=1}^{n} q_{ik}p_{kn}\alpha_n \quad . \qquad (3\text{-}7)$$

Equation (3-7) written in n-tuple form with restricted coefficients is the i-th row of the matrix product QP. Since the α_i constitute a basis, the reduced form of Equation (3-7) must be $\alpha_i = \alpha_i$. As a consequence, it is seen that

$$Q \cdot P = I. \qquad (3\text{-}8)$$

By advancing a dual argument, one deduces

$$P \cdot Q = I. \qquad (3\text{-}9)$$

Equations (3-8) and (3-9) can be used as a check on the determination of the P and Q matrices. These equations are necessary but not sufficient conditions for the accuracy of P and Q.

3.4 Multiplication

To multiply two elements of a number system related to the residue number system, one forms partial products, one for each component of

the multiplier, and adds them together producing a linear combination of the basis vectors which is then reduced by means of the Carry Algorithm. The algorithm for determining the form of the partial products will be called the Multiplication Algorithm.

The Multiplication Algorithm may be stated as follows:

Consider the most general multiplicand $(y_1, y_2, \ldots, y_n)$ and multiplier $(x_1, x_2, \ldots, x_n)$.

1. Write the multiplicand $(y_1, y_2, \ldots, y_n)$ as the vector sum $y_1\alpha_1 + y_2\alpha_2 + \ldots + y_n\alpha_n$.

Beginning with $j = n$

2. Write the partial multiplier $(0, \ldots, 0, x_j, 0, \ldots, 0)$ as the vector $x_j\alpha_j$.
3. Multiply, component by component, $x_j\alpha_j \cdot y_i\alpha_i$ $i = 0, \ldots, n$.
4. Express the vector $x_j\alpha_j \cdot y_i\alpha_i$ as the linear combination $Z_i\alpha_1 + \ldots + Z_i\alpha_i$, $i = 0, \ldots, n$.
5. Sum the linear forms produced in Step 4 to determine the j-th partial sum.
6. Reduce j by one and repeat Steps 2 through 5. Terminate the procedure after doing the above steps with $j = 1$.

Example: Consider the multiplication

$$(y_1, y_2, \ldots, y_4)_M \cdot (x_1, x_2, \ldots, x_4)_M$$

in the mixed base number system where $m_1 = 2$, $m_2 = 3$, $m_3 = 5$, $m_4 = 7$. Here the basis is $< (1,0,0,0), (1,2,0,0), (1,1,2,0), (1,1,1,1) >$.

The following steps are numbered to correspond to the statement of the Multiplication Algorithm.

1. $(y_1, 0,0,0) + (y_2,2y_2,0,0) + (y_3,y_3,2y_3,0) + (y_4,y_4,y_4,y_4)$

2. (x_4,x_4,x_4,x_4)

3. $(y_1x_4,0,0,0)$

   $(y_2x_4,2y_2,x_4,0,0)$

   $(y_3x_4,y_3x_4,2y_3x_4,0)$

   $(y_4x_4,y_4x_4,y_4x_4,y_4x_4)$

4. $(y_1x_4,0,0,0) = (y_1x_4,0,0,0)_M$

   $(y_2x_4,2y_2x_4,0,0) = (0,y_2x_4,0,0)_M$

   $(y_3x_4,y_3x_4,2y_3x_4,0) = (0,0,y_3x_4,0)_M$

   $(y_4x_4,y_4x_4,y_4x_4,y_4x_4) = (0,0,0,y_4x_4)_M$

5. $(y_1,y_2,y_3,y_4)_M \cdot (0,0,0,x_4)_M = (x_4,y_1,x_4y_2,x_4y_3,x_4y_4)_M$

2. $(x_3,x_3,2x_3,0)$

3. $(y_1x_3,0,0,0) = A$

   $(y_2x_3,2y_2x_3,0,0) = B$

   $(y_3x_3,y_3x_3,2(2y_3x_3),0) = C$

   $(y_4x_3,y_4x_3,2y_4x_3,0) = D$

4. $A = (y_1x_3,0,0,0)_M$

   $B = (0,y_2x_3,0,0)_M$

   $C = (0,0,2y_3x_3,0)_M + (0,x_3y_3,0,0)_M$

   since $(0,0,2x_3y_3,0)_M + (0,x_3y_3,0,0)_M$

   $= (2x_3y_3,2x_3y_3,4x_3y_3,0) + (x_3y_3,2x_3y_3,0,0)$

   $= (y_3x_3,y_3x_3,2(2y_3x_3),0)$

   $D = (0,0,x_3y_4,0)_M$.

5. $(y_1,y_2,y_3,y_4)_M \cdot (0,0,x_3,0)_M$

$= (y_1x_3, y_2x_3 + y_3x_3, 2x_3y_3 + x_3y_4, 0)_M$

2. $(x_2, 2x_2, 0, 0)$

3. $(x_2y_1, 0, 0, 0) = A'$

$(x_2y_2, 4x_2y_2, 0, 0) = B'$

$(x_2y_3, 2x_2y_3, 0, 0) = C'$

$(x_2y_4, 2x_2y_4, 0, 0) = D'$

4. $A' = (y_1x_2, 0, 0, 0)_M$

$B' = 2(0, x_2y_2, 0, 0)_M + (x_2y_2, 0, 0, 0)_M$

for $(x_2y_2, 4x_2y_2, 0, 0)_M$

$= (x_2y_2, 2x_2y_2, 0, 0) + (0, 2x_2y_2, 0, 0)$

$= (x_2y_2, 2x_2y_2, 0, 0) + (x_2y_2, 2x_2y_2, 0, 0)$

$+ (x_2y_2, 0, 0, 0)$

$C' = (0, x_2y_3, 0, 0)_M$

$D' = (0, x_2y_4, 0, 0)_M$

5. $(y_1,y_2,y_3,y_4)_M \cdot (0,x_2,0,0)_M$

$= (x_2y_1 + x_2y_2, 2x_2y_2 + x_2y_2 + x_2y_4, 0, 0)_M$

2. $(x_1, 0, 0, 0)$

3. $(x_1y_1, 0, 0, 0)$

$(x_1y_2, 0, 0, 0)$

$(x_1y_3, 0, 0, 0)$

$(x_1y_4, 0, 0, 0)$

5. $(y_1,y_2,y_3,y_4)_M \cdot (x_1,0,0,0)_M = (x_1y_1 + x_1y_2 + x_1y_3 + x_1y_4, 0, 0, 0)_M$

The results of the Multiplication Algorithm in this example are the following for rules for the formation of partial products:

$$(y_1,y_2,y_3,y_4)_M \cdot (x_1,0,0,0)_M = (x_1y_1 + x_1y_2 + x_1y_3 + x_1y_4,0,0,0)_M$$

$$(y_1,y_2,y_3,y_4)_M \cdot (0,x_2,0,0)_M = (x_2y_1 + x_2y_2,2x_2y_2 + x_2y_3 + x_2y_4,0,0)_M$$

$$(y_1,y_2,y_3,y_4)_M \cdot (0,0,x_3,0)_M = (x_3y_1,x_3y_2 + x_3y_3,2x_3y_3 + x_3y_4,0)_M$$

$$(y_1,y_2,y_3,y_4)_M \cdot (0,0,0,x_4)_M = (x_4y_1,x_4y_2,x_4y_3,x_4y_4)_M .$$

To multiply two mixed base elements $(y_1,y_2,y_3,y_4)_M$ and $(x_1,x_2,x_3,x_4)_M$ one procedes as follows:

1. Form the above partial products.
2. Reduce each partial product by employing the Carry Algorithm.
3. Sum the partial products again employing the Carry Algorithm.

As an example, consider the product $(1,2,3,4)_M \cdot (0,2,4,6)_M$

The partial products are

$$(1,2,3,4)_M \cdot (0,0,0,6)_M = (6,12,18,24) = (1,1,1,3)_M \text{ Mod } M$$

$$(1,2,3,4)_M \cdot (0,0,4,0)_M = (4,28,40,0) = (1,1,0,0)_M \text{ Mod } M$$

$$(1,2,3,4)_M \cdot (0,2,0,0)_M = (6,22,0,0) = (1,1,0,0)_M \text{ Mod } M .$$

Therefore,

$$(1,2,3,4)_M \cdot (0,2,4,6)_M = (0,0,1,3)_M \text{Mod } M$$

$$(0,2,4,6)_M \longleftrightarrow 104$$

and

$$(1,2,3,4)_M \longleftrightarrow 200$$

$$104 \cdot 200 = 20800 \equiv 10 \text{ Mod } M$$

$$(0,0,1,3)_M \longleftrightarrow 10 .$$

Let us look again at the question of associativity of matrix multiplication. Consider the three matrices $A = ||a_{ij}||$, $B = ||b_{ij}||$, and $C = ||c_{ij}||$. Let the basis of the vector spaces be $U, < \alpha_1, \alpha_2, \ldots, \alpha_n >$; $V, < \beta_1, \beta_2, \ldots, \beta_n >$; $W, < \gamma_1, \gamma_2, \ldots, \gamma_n >$; and $Z, < \delta_1, \delta_2, \ldots, \delta_n >$. T_A is a transformation from $U \rightarrow V$, $T_B : V \rightarrow W$, and $T_C : W \rightarrow Z$.

The i-th row of the product $A \cdot B$ is

$$\left(\sum_{k=1}^{n} a_{ik}b_{k1},\ \sum_{k=1}^{n} a_{ik}b_{k2},\ \ldots,\ \sum_{k=1}^{n} a_{ik}b_{kn}\right)$$

reduced relative to the basis $< \gamma_1, \ldots, \gamma_n >$. Designating the reduced form obtained $(e_{i1}, e_{i2}, \ldots, e_{in})$, we obtain

$$e_{in} = \left\{ \frac{\sum_{k=1}^{n} a_{ik}b_{kn}}{m_n} \right\}$$

Define

$$f_{in} = \left[\frac{\sum_{k=1}^{n} a_{ik}b_{kn}}{m_n} \right]$$

$$e_{i,n-1} = \left\{ \frac{\sum_{k=1}^{n} a_{ik}b_{k,n-1} + f_{in}g_{n,n-1}}{m_{n-1}} \right\}$$

$$f_{i,n-1} = \left[\frac{\sum_{k=1}^{n} a_{ik}b_{k,n-1} + f_{in}g_{n,n-1}}{m_{n-1}} \right]$$

. . .

$$e_{i1} = \left\{ \frac{\sum_{k=1}^{n} a_{ik}b_{k1} + f_{in}g_{n,1} + f_{i,n-1}g_{n-1} + \ldots + f_{i,2}g_{21}}{m_1} \right\}$$

where

$$m_j \cdot \gamma_j = g_{j1}\gamma_1 + g_{j2}\gamma_2 + \ldots + g_{jj-1}\gamma_{j-1}.$$[11]

---

[11] $\left[\frac{a}{b}\right]$ denotes the integral part of the division $\frac{a}{b}$.

$\left\{\frac{a}{b}\right\}$ denotes the fractional part of the division $\frac{a}{b}$.

The i-th row of the product $(A \cdot B)C$ is

$$\left(\sum_{k=1}^{n} e_{ik}c_{k1},\ \sum_{k=1}^{n} e_{ik}c_{k2},\ \ldots,\ \sum_{k=1}^{n} e_{ik}c_{kn}\right)$$

reduced relative to the basis $< \gamma_1, \gamma_2, \ldots, \gamma_n >$.

This gives rise to the reduced form

$$(h_{i1}, h_{i2}, \ldots, h_{in})$$

where

$$h_{in} = \left\{ \frac{\sum_{k=1}^{n} e_{ik}c_{kn}}{m_n} \right\}$$

$$\ell_{in} = \left[\frac{\sum_{k=1}^{n} e_{ik}c_{kn}}{m_n} \right]$$

$$h_{i,n-1} = \left\{ \frac{\sum_{k=1}^{n} e_{ik}c_{k,n-1} + \ell_{in}r_{n,n-1}}{m_{n-1}} \right\}$$

and

$$\ell_{i,n-1} = \left[\frac{\sum_{k=1}^{n} e_{ik}c_{k,n-1} + \ell_{in}r_{n,n-1}}{m_{n-1}} \right]$$

$$h_{i1} = \left\{ \frac{\sum_{k=1}^{n} e_{ik}c_{k1} + \ell_{in}r_{n,1} + \ell_{i,n-1}r_{n-1,1} + \ldots + f_{i2}r_{21}}{m_1} \right\}$$

where

$$m_j\delta_j = r_{j1}\delta_1 + \ldots + r_{j,j-1}\delta_{j-1} .$$

Similarly the i-th row of the product $B \cdot C$ is

$$\left(\sum_{k=1}^{n} b_{ik}c_{k1},\ \sum_{k=1}^{n} b_{ik}c_{k2},\ \ldots,\ \sum_{k=1}^{n} b_{ik}c_{kn}\right)$$

reduced relative to the basis $< \delta_1, \ldots, \delta_n >$.

The reduced form is $(s_{i1}, s_{i2}, \ldots, s_{in})$

where

$$s_{in} = \left\{ \frac{\sum_{k=1}^{n} b_{ik}c_{kn}}{m_n} \right\}$$

$$t_{in} = \left[\frac{\sum_{k=1}^{n} b_{ik}c_{kn}}{m_n} \right]$$

$$s_{i,n-1} = \left\{ \frac{\sum_{k=1}^{n} b_{ik}c_{k,n-1} + t_{in}r_{n,n-1}}{m_{n-1}} \right\}$$

$$t_{i,n-1} = \left[\frac{\sum_{k=1}^{n} b_{ik}c_{k,n-1} + t_{1n}r_{n,n-1}}{m_{n-1}} \right]$$

. . .

$$s_{i1} = \left\{ \frac{\sum_{k=1}^{n} b_{ik}c_{k1} + t_{in}r_{n1} + t_{i,n-1}r_{n-1,1} + \cdots + t_{i2}r_{21}}{m_1} \right\}$$

The i-th row of $A(BC)$ is

$$\left(\sum_{k=1}^{n} a_{ik}s_{k1}, \sum_{k=1}^{n} a_{ik}s_{k2}, \ldots, \sum_{k=1}^{n} a_{ik}s_{kn} \right)$$

reduced relative to the basis $< \delta_1, \ldots, \delta_n >$.

The reduced form is

$$(v_{i1}, v_{i2}, \ldots, v_{in})$$

where

$$v_{in} = \left\{ \frac{\sum_{k=1}^{n} a_{ik}s_{kn}}{m_n} \right\}$$

$$v_{in} = \left\{ \frac{\sum_{k=1}^{n} a_{ik}s_{kn}}{m_n} \right\}$$

$$v_{i,n-1} = \left\{ \frac{\sum_{k=1}^{n} a_{ik}s_{k,n-1} + u_{in}r_{n,n-1}}{m_{n-1}} \right\}$$

$$u_{i,n-1} = \left[\frac{\sum_{k=1}^{n} a_{ik}s_{kn-1} + u_{in}r_{n,n-1}}{m_{n-1}} \right]$$

. . .

$$v_{il} = \left\{ \frac{\sum_{k=1}^{n} a_{ik}s_{kl} + u_{in}r_{n,1} + u_{i,n-1}r_{n-1,1} + \cdots + u_{i2}r_{21}}{m_1} \right\}$$

Selecting particular components for comparison

$$h_{in} = \left(\sum_{k=1}^{n} e_{ik}c_{kn}\right) \bmod m_n$$

$$= \left[\left(\sum_{k=1}^{n} a_{ik}b_{kn}\right) \bmod m_n \cdot c_n\right.$$

$$+ \left(\sum_{k=1}^{n} a_{ik}b_{k,n-1} + f_{in}g_{n,n-1}\right) \bmod m_{n-1} \cdot c_{n-1,n} + \cdots$$

$$+ \left(\sum_{k=1}^{n} a_{ik}b_{kl} + f_{in}g_{nl} + f_{i,n-1}g_{n-1,1} + \cdots\right.$$

$$\left. + f_{i2}g_{21}\right) \bmod m_1 \cdot c_{1,n}] \bmod m_n$$

or

$$h_{in} = \sum_{r=1}^{n} \left(\sum_{k=1}^{n} a_{ik}b_{kr} \bmod m_r\right)c_{rn} \bmod m_n$$

$$+ \sum_{\ell=1}^{n-1} \left(\sum_{j=\ell}^{n} f_{ij}g_j \bmod m_\ell\right) \cdot c_{\ell n} \bmod m_n .$$

Removing one term and changing indicies, one finds

$$h_{in} = \sum_{k=1}^{n} a_{ik}b_{kn}c_{nn} \bmod m_n$$

$$+ \left[\sum_{r=1}^{n-1} \left(\sum_{k=1}^{n} a_{ik}b_{kr} \bmod m_r \right) \cdot c_{rn} \right] \bmod m_n$$

$$+ \sum_{r=1}^{n-1} \left(\sum_{j=r}^{n} f_{ij}g_{jr} \bmod m_r \right) c_{rn} \bmod m_n$$

$$= \sum_{k=1}^{n} a_{ik}b_{kn}c_{nn} \bmod m_n$$

$$+ \sum_{r=1}^{n-1} \left[\sum_{k=1}^{n} a_{ik}b_{kr} + \sum_{j=r}^{n} f_{ij}g_{jr} \right] \bmod m_r \cdot c_{rn} \bmod m_n \, .$$

Also

$$v_{in} = \left(\sum_{k=1}^{n} a_{ik}s_{kn} \right) \bmod m_n$$

$$= \left[a_{i1} \left(\sum_{k=1}^{n} b_{ik}c_{kn} \bmod m_n \right) + a_{i2} \left(\sum_{k=1}^{n} b_{2k}c_{kn} \bmod m_n \right) \right.$$

$$\left. + \ldots 2\, a_{in} \left(\sum_{k=1}^{n} b_{nk}c_{kn} \bmod m_n \right) \right] \bmod m_n \quad \text{or}$$

$$v_{in} = \sum_{k=1}^{n} \sum_{r=1}^{n} a_{ir}b_{rk}c_{kn} \bmod m_n$$

$$= \sum_{r=1}^{n} a_{ir}b_{rn}c_{nn} \bmod m_n + \sum_{k=1}^{n-1} \sum_{r=1}^{n} a_{ir}b_{rk}c_{kn} \bmod m_n \, .$$

Changing indicies for clarity, one obtains

$$v_{in} = \sum_{k=1}^{n} a_{ik}b_{kn}c_{nn} \bmod m_n + \sum_{r=1}^{n-1} \sum_{k=1}^{n} a_{ik}b_{kr}c_{rn} \bmod m_n \, .$$

Since the first terms of h_{in} and v_{in} are the same, it is sufficient to look at

$$[\sum_{r=1}^{n} a_{ir}b_{rk} + \sum_{j=k}^{n} f_{ij}g_{ik}] \bmod m_k \qquad (3\text{-}10)$$

and

$$[\sum_{r=1}^{n} a_{ir}b_{rk}] \bmod m_n \ . \qquad (3\text{-}11)$$

It is seen that the above expressions are not in general equal, for the range of Expression (3-10) is the positive integers less than m_k whereas the range of (3-11) is the positive integers less than m_n. Since the m_j are relatively prime, one is led to conclude that the i-th row of $(AB)C$ is not equal in general to the corresponding row of $A(BC)$. This was shown independent of the selection of the basis for the various image spaces involved. Therefore, no selection of basis will guarantee associativity of matrix multiplication.

CHAPTER IV

THE MIXED BASE NUMBER SYSTEM

4.1 The Mixed Base Number System

When one allows complementation as a means of subtraction, the representations are partitioned into two classes, those representations termed positive and those termed negative (or complements of positives). For sign determination and magnitude comparison, it is necessary to identify the classification of each and every representation. As is well understood, the elements of the residue number system contain insufficient information for making such an identification. The structure of the mixed base system makes the identification immediate. In addition, the mixed base system allows one to handle the problems of additive and multiplicative overflow, and division. We shall see that the mixed base system extracts payment in the form of carries for these advantages.

The basis vectors for the mixed base system are generated in the following manner:

1. Order the prime $m_1, m_2, \ldots, m_n$.
2. Set α_n equal to the vector consisting of all 1's.
   Beginning with $j = n$
3. Obtain $\alpha_{j-1} = m_j \alpha_j$
4. Repeat step 3 with j replaced by $j-1$. Stop after performing 3 with $j = 1$.

Theorem 11. The vectors $\{\alpha_1, \ldots, \alpha_n\}$ produced by the above scheme constitute a basis.

In order to prove the theorem Lemma 1 will be proven.

Lemma 1: If $(a, b) = 1$, then $(a \bmod b, b) = 1$

Proof: $a = bg + r$, $0 \leq r < b$ and by definition r is the residue of $a \bmod b$

$$a \equiv r \bmod b$$

or

$$a \equiv (a \bmod b) \bmod b .$$

Since $x \equiv y \bmod z \Rightarrow (x,z) = (y,z)$, $(a \bmod b, b) = (a,b)$. The conclusion follows.

Proof of Theorem 11: α_n is a representation of the integer 1. The vector α_ℓ is the residue representation of $A_\ell = \prod_{i=\ell+1}^{n} m$. Each m_i for $i \leq \ell$ is relatively prime to A_ℓ. Thus it is seen that

$$k_{\ell i} = 0 \quad \text{for} \quad i > \ell$$

and $(k_{\ell i}, m_i) = 1$ for $i \leq \ell$. By Theorem 4 the set $< \alpha_1, \alpha_2, \ldots, \alpha_n >$ is a basis of R^n.

Theorem 12. An element $(x_1, x_2, \ldots, x_n)$ of the mixed base number system is a representation of an integer X in the range $C \frac{M}{m_1} \leq X < (C+1) \frac{M}{m_1}$ if and only if $x_1 = C$.

Proof: The element $X = (x_1, x_2, \ldots, x_n)$ is a coding of the integer

$$x_1 \frac{M}{m_1} + x_2 \frac{M}{m_1 m_2} + x_3 \frac{M}{m_1 m_2 m_3} + \ldots + x_n \text{ modulo } M. \qquad (4\text{-}1)$$

Consider the quantity

$$x_1 \frac{M}{m_1} + x_2 \frac{M}{m_1 m_2} + x_3 \frac{M}{m_1 m_2 m_3} + \ldots + x_n. \qquad (4\text{-}2)$$

The largest value Expression (4-2) can assume is attained when

$$x_i = m_i - 1 .$$

Upon substitution one obtains

$$(m_1 - 1)\frac{M}{m_1} + (m_2 - 1)\frac{M}{m_1 m_2} + (m_3 - 1)\frac{M}{m_1 m_2 m_3} + \ldots + (m_n - 1)$$

This is then rewritten as

$$(m_n - 1) + m_n (m_{n-1} - 1) + m_n m_{n-1}(m_{n-2} - 1) + \ldots + \frac{M}{m_1 m_2}(m_2 - 1) + \frac{M}{m_1}$$

remembering the scheme for generating the base vectors. All but the following terms add out:

$$-1 + \frac{M}{m_1} + \frac{M}{m_1} = M - 1 .$$

This means that Expression (4-2) is equivalent to Expression (4-1). Consider next the quantity:

$$(m_2 - 1)\frac{M}{m_1 m_2} + (m_3 - 1)\frac{M}{m_1 m_2 m_3} + \ldots + (m_n - 1)$$

which is equal to

$$(m_n - 1) + m_n (m_{n-1} - 1) + m_n m_{n-1} (m_{n-2} - 1) + \ldots + (m_2 - 1)\frac{M}{m_1 m_2} .$$

The above expression reduces to

$$-1 + \frac{M}{m_1} .$$

From this the conclusion is clear.

To determine the sign of a mixed base number it is necessary to have a partitioning of elements representing consecutive integers into two classes. The first coordinate gives such a partitioning if m_1 is even, i.e., $x_1 < m_1/2 \equiv 0 \leq X < M/2$.[12]

[12] The relevance of the mixed base number system to the problems of sign detection, magnitude comparison, and additive overflow has been cited by Garner and Svoboda.

When applying the Carry Algorithm to the reduction of the expression

$$a_1\alpha_1 + a_2\alpha_2 + \ldots + a_n\alpha_n$$

(α_i is a base vector of the mixed base system) one must increase by one a'_j for every multiple of m_{j+1} contained in a'_{j+1} (The notation here is consistent with that contained in the discussion of the Carry Algorithm). This indicates that a carry may be propagated from the n-th coordinate to the first. Therefore, when adding two elements of the mixed base system, one unit of time is required to produce the unreduced sum and up to n-1 units of time may be required to propagate carries. Borrowing is accomplished by reducing a'_j by 1 and increasing a'_{j+1} by m_{j+1}. Again up to n-1 units of time may be required to perform subtraction or complementation.

Multiplication of two elements of the mixed base system was discussed and an example given in Chapter III.

Theorem 13. When two mixed base numbers are added, the carries are binary and a position which produces a carry cannot also propagate a carry.

Proof: The largest j-th component of the unreduced sum occurs when the j-th components of the addend and the augend are maximum, i.e., equal to $m_j - 1$. The maximum sum will be $m_j - 2$.

Let j=n. The maximum unreduced n-th component will be $2m_n - 2$; therefore, the carry can only be zero or one. Consider $j = n-1$. The component $2m_{n-1} - 2$ will produce a carry. If a carry was generated by the n-th position $m_{n-1} - 2 + 1 = m_{n-1} - 1 < m_{n-1}$; therefore, no carry is both propagated and generated.

Assume the results true for $j=m$. In this case $2m_{m-1} - 2$ will generate a carry of one, and $2m_{m-1} - 2 + 1$ will also give rise to a carry of one.

4.2 Overflow

A problem of the residue number system which can be solved with the mixed base system is overflow. In a mixed base number system where $m_1 = 2$, the integers less than $M/2$ are represented in the system. Therefore, overflow is defined to be the condition where the true arithmetic result is an integer larger than or equal to $M/2$. First additive overflow will be discussed and then various conditions for the absence of multiplicative overflow will be demonstrated. (Due to sign detection and overflow conditions it will be convenient to assume $m_1 = 2$ for the remainder of this chapter. In this chapter all n-tuples will be elements of the mixed base number system.)

Theorem 14. If the sum of two positive elements of the mixed base system is $(z_1, \ldots, z_n)$ additive overflow occurs if and only if $z_1 = 1$.

Proof: $(z_1, z_2, \ldots, z_n)$ represents the integer

$$z_1 \frac{M}{2} + z_2 \frac{M}{2m_2} + \ldots + z_n \bmod M \qquad (4\text{-}3)$$

and overflow occurs when

$$z_1 \frac{M}{2} + z_2 \frac{M}{2m_2} + \ldots + z_n \geq \frac{M}{2} \quad .$$

This will clearly be the case if $z_1 = 1$.

The largest value that $z_2 \frac{M}{2m_2} + \ldots + z_n$ can attain has been shown to be $\frac{M}{2} - 1$. The largest sum possible is $\frac{M}{2} - 1 + \frac{M}{2} - 1 = M - 2 < M$. Thus Expression (4-3) becomes $z_1 \frac{M}{2} + z_2 \frac{M}{2m_2} + \ldots + z_n$ and the other conclusion follows.

To insure that multiplicative overflow will not occur, conditions will be given which will insure that no multiplicative overflow will occur as the partial products are formed. It is also necessary that overflow does not occur when the partial products are summed. This also will be treated.

Consider $(y_1, y_2, \ldots, y_n) \cdot (x_1, 0, \ldots, 0)$. Since $(x_1, 0, \ldots, 0) \longleftrightarrow x_1 \frac{M}{2}$, overflow will occur unless $y_1 = y_2 = \ldots = y_n = 0$

Consider next $(y_1, y_2, \ldots, y_n) \cdot (0, x_2, 0, \ldots, 0)$ and note that $(0, x_2, 0, \ldots, 0) \longleftrightarrow x_2 \frac{M}{2m_2}$. Therefore, the condition which is necessary and sufficient for the absence of multiplicative overflow in this partial product is:

$$y_1 = y_2 = \ldots y_{n-1} = 0 \quad \text{and} \quad x_2 \cdot y_n < m_2 .$$

Since $(0,0,x_3,0,\ldots,0)$ represents $x_3 \frac{M}{2m_2m_3}$ and $(y_1,y_2,\ldots,y_n)$ represents $Y = y_n + y_{n-1}m_n + y_{n-2}m_nm_{n-1} + \ldots + y_1 \frac{M}{m_1}$, a necessary and sufficient condition to prevent overflow is

$$x_3 \frac{M}{2m_2m_3} \; Y < \frac{M}{2} \qquad (4\text{-}4)$$

or

$$x_3 \, Y < m_2m_3 .$$

Thus it is seen that condition (4-4) becomes

$$x_3(y_n + y_{n-1}m_n) < m_2m_3 \tag{4-5}$$

and

$$y_1 = y_2 = \cdots = y_{n-2} = 0 .$$

Since $y_n < m_n$ we may substitute for Expression (4-5)

$$x_3m_n(y_{n-1} + 1) < m_2m_3$$

to obtain a sufficient condition.

Consider now the general partial product $(y_1, y_2, \ldots, y_n) \cdot (0, \ldots, 0, x_j, 0, \ldots, 0)$. In this case $(0, \ldots, 0, x_j, 0, \ldots, 0)$ represents $x_j \frac{M}{2m_2 \cdots m_j}$. To guarantee that no overflow will occur we must satisfy the inequality $x_jY < m_2 \cdots m_j$ or

$$x_j(y_n + y_{n-1}\, m_n + \cdots + y_{n-(j-2)}m_n \cdots m_{n-(j-3)}$$

$$+ y_{n-j+1}\, m_nm_{n-1} \cdots m_{n-j-2} + \cdots + y_1 \tfrac{M}{2}) < m_2m_3m_j . \tag{4-6}$$

The condition becomes

$$y_{n-j+1} = y_{n-j} = \cdots = y_1 = 0 \quad \text{and}$$

$$x_j(y_n + y_{n-1}m_n + \cdots + y_{n-(j-2)}m_n \cdots m_{n-(j-3)}) < m_2m_3m_j. \tag{4-7}$$

This constitutes a necessary and sufficient condition that this partial product does not produce an overflow. Again there are simpler inequalities the validity of which will imply the validity of (4-7). These

inequalities are

$$x_j\,[(y_{n-1}+1)\,m_n + \ldots + y_{n-(j-2)}m_n\cdots m_{n-(j-3)}] < m_2m_3\cdots m_j,$$

$$x_j\,[(y_{n-2}+1)\,m_nm_{n-1} + \ldots + y_{n-(j-2)}m_n\cdots m_{n-(j-3)}] < m_2m_3\cdots m_j,$$

$$\cdots$$

$$x_j\,[(y_{n-(j-2)}+1)\,m_n\cdots m_{n-(j-3)}] < m_2m_3\cdots m_j\ .$$

The sufficiency of the above inequalities stems from Theorem 15.

Theorem 15. If $x_i < m_i$, then

$$x_1 + x_2m_1 + x_3m_1m_2 + \ldots + x_{k-1}m_1m_2\cdots m_{k-2} < m_1m_2\cdots m_{k-1}\ .$$

Proof: The maximum value that

$$x_1 + x_2m_1 + x_3m_1m_2 + \ldots + x_{k-1}m_1m_2\cdots m_{k-2}$$

can attain is

$$(m_1-1) + (m_2-1)\,m_1 + (m_3-1)\,m_1m_2 + \ldots + (m_{k-1}-1)\,m_1m_2\cdots m_{k-2}$$

$$= m_1m_2\cdots m_{k-2}\,m_{k-1} - 1 < m_1m_2\cdots m_{k-1}\ .$$

Even when none of the partial products of a multiplication involve multiplicative overflow, overflow may result when the partial products are summed. Such overflow will be avoided only if the first coefficient of the unreduced sum is zero and no carry is propagated into that position.

4.3 Division in the Mixed Base System

Utilizing the overflow rules given in this chapter, one may now state rules for performing division in the mixed base system. The conditions

for the absence of multiplicative overflow and rules for forming partial products provide a means for estimating trial divisors. The subtraction rules then allow one to determine the new dividend. The method will be demonstrated before the formal rules are stated.

Example: Divide 95 by 14 using the mixed base system where $m_1 = 2$, $m_2 = 3$, $m_3 = 5$, and $m_4 = 7$.

$$95 \longleftrightarrow (0,2,3,4)$$
$$14 \longleftrightarrow (0,0,2,0)$$

Step 1: Determine first divisor

$$0,0,2,0 \overline{)\ 0,2,3,4}^{\ \alpha}$$

It is seen that it is necessary for $\alpha = 0$ to avoid overflow.

Step 2: Determine second divisor

$$0,0,2,0 \overline{)\ 0,2,3,4}^{\ 0,\beta}$$

Again to avoid overflow $\beta = 0$.

Step 3: Determine third divisor

$$0,0,2,0 \overline{)\ 0,2,3,4}^{\ 0,0,\gamma}$$

From overflow considerations it is seen that $\gamma = 1$ is satisfactory.

$$0,0,2,0 \overline{)\ 0,2,3,4}^{\ 0,0,1}$$
$$\underline{0,2,4,0}$$

It is seen that $\gamma = 1$ is too large; therefore, $\gamma = 0$.

Step 4: Determine fourth divisor

$$0,0,2,0 \overline{)\ 0,2,3,4}^{\ 0,0,0,\xi}$$

From reference to overflow rules and multiplication the estimate is $\xi = 6$.

$$
\begin{array}{r}
0,0,0,6 \\
0,0,2,0 \;\overline{)\;0,2,3,4} \\
\underline{0,2,2,0} \\
0,0,1,4
\end{array}
$$

The division is completed giving $95 = \underline{6} \cdot 14 + \underline{\underline{11}}$.

The procedure for determining the trial divisor is as follows:

1. Consult the conditions for the absence of multiplicative overflow to determine the possible range of trial divisors.
2. Use the rules for forming partial products to determine a trial divisor which will yield the required zero(s) in the most significant place(s) and a non-zero component in the most significant non-zero position (k-th) of the divident. (The k-th component must be less than or equal to the k-th component of the divident.)
3. Subtract the partial product from the divident and if necessary revise the quotient so that the least possible non-negative result is achieved.

4.4 Digit Fill-In

An addition problem which can be solved by using the mixed base system is the problem of digit fill in. If one is given the coordinates of a vector representing the integer relative to the basis $A = \{\alpha_1,\alpha_2,\ldots,\alpha_n\}$ and base moduli $m_1,m_2,\ldots,m_n$, the problem is to express the integer s in terms of coordinates relative to be basis $\{\beta_1,\beta_2,\ldots,\beta_n,\beta_{n+1},\ldots,\beta_{n+m}\} = B$ and base moduli $m_1,m_2,\ldots,m_n,m_{n+1},\ldots,m_{n+m}$. The m_i are pairwise relatively prime.

Since one can represent s as a vector relative to the base moduli $m_1, m_2, \ldots, m_n$, s is less than $\prod_{i=1}^{n} m$. Therefore, if s is expressed in the mixed base system relative to the moduli $m_{n+1}, m_{n+2}, \ldots, m_{n+m}, m_1, m_2, \ldots, m_n$, the coordinates with weights greater than or equal to $\prod_{i=1}^{n} m_i$ must be zeros. The weights of the last n components of the mixed system will be, in reverse order, $1, m_n, m_n m_{n-1}, \ldots, m_n m_{n-1} \ldots m_2$. These weights are the same as the weights of the components in the mixed base system relative to the primes $m_1, m_2, \ldots, m_n$. Therefore, digit fill-in is accomplished as follows:

1. Perform the change of basis operation from the basis A to the mixed base system.
2. Prefix the m zeros to produce the correct representation in the extended mixed base system.
3. Perform the change of basis operation, this time from the extended mixed base system to the basis B.

CHAPTER V

OTHER NUMBER SYSTEMS RELATED TO THE RESIDUE NUMBER SYSTEM

5.1 Partitioning Properties

In Chapter IV it was noted that the mixed base number system representations corresponding to consecutive integers are partitioned by the first coordinate. It was this property which lead to the solution of the outstanding problems of the residue number system. A question arises whether number systems exist which possess the desirable partitioning while having simpler rules of arithmetic manipulation. It will be shown that the only related number system to achieve a partitioning of elements representing consecutive integers is the mixed base system.

Lemma 2: For any integer k within the range $1 < k < p$, there exists an integer ℓ such that

$$p \leq \ell k < 2p \quad \text{where} \quad \ell < p \quad \text{and} \quad p > 2 .$$

Proof: It is evident that k cannot lie in the range

$$1 \leq k < \frac{p}{p-1} ,$$

for

$$p > 2, \quad p-2 > 0$$

$$2p-2 > p$$

$$2(p-1) > p ;$$

therefore,

$$\frac{p}{p-1} < 2 .$$

Thus there exists an integer ℓ such that

$$\frac{p}{\ell} \leq k < \frac{p}{\ell-1} \quad \text{with} \quad p > \ell$$

and one concludes $p \leq \ell k < \frac{\ell}{\ell-1} p \leq 2p$. Q.E.D.

Theorem 16. If the base moduli are ordered $m_1, m_2, \ldots, m_n$, only the mixed base number system gives a partitioning of the elements into those elements which represent integers in the range less than $\frac{(c+1)M}{m_1}$ but greater than or equal to $c \frac{M}{m_1}$, where c is the first coordinate of the element.

Proof: We will consider all number systems which give the desired partitioning and conclude that they are all identical to the mixed base system.

Consider the number system based on the vectors $< \beta_1, \beta_2, \ldots, \beta_n >$. It is assumed that this number system achieves the desired partitioning. Here β_1 corresponds to B_1; β_2 to B_2; etc. An element of this number system $\mathfrak{X} = (y_1, y_2, \ldots, y_n)$ represents the integer

$$y_1 B_1 + y_2 B_2 + \ldots + y_n B_n \bmod M.$$

For the set $< \beta_1, \ldots, \beta_n >$ to be a basis it must be triangular; therefore, one may state

$$B_n = k_n$$

$$B_{n-1} = k_{n-1} m_n$$

$$\ldots$$

$$B_j = k_j \prod_{i=j+1}^{n} m_i$$

$$\ldots$$

$$B_1 = k_1 \prod_{i=2}^{n} m_i$$

where

$$0 < k_j < \prod_{i=1}^{j} m_i \ .$$

Consider $y_1 = c$ and $y_i = 0$ for $i \neq 1$. The expression $y_1B_1 + y_2B_2 + \ldots + y_nB_n \bmod M$ becomes $c\, k_1 \frac{M}{m_1} \bmod M$.

Clearly, $\frac{cM}{m_1} \leq ck_1 \frac{M}{m_1} < (c+1) \frac{M}{m_1}$ for $c < m_1$ requires $k_1 = 1$. One condition which must be satisfied is $y_1 = 0 \equiv Y < \frac{M}{m_1}$. This requirement becomes

$$y_2B_2 + y_3B_3 + \ldots + y_nB_n = z \bmod M \tag{5-1}$$

where $z < \frac{M}{m_1}$ for all $y_2, y_3, \ldots, y_n$.

Consider first the case

$$y_2 \neq 0,\ y_3 = \ldots = y_n = 0.$$

We have

$$y_2B_2 = y_2k_2 \frac{M}{m_1m_2} \ .$$

If there is an integer ℓ in the range $0 < \ell < m_2$ such that

$$\frac{M}{m_1} \leq \ell k \ \frac{M}{m_1m_2} < M \tag{5-2}$$

condition (5-1) is violated. Expression (5-2) may be written

$$m_2 \leq \ell k_2 < m_1m_2. \tag{5-3}$$

By Lemma 2, it is seen that for $k_2 > 1$, we can find an integer in the range $0 < \ell < m$ such that $m_2 \leq \ell k_2 < 2m_2$. Since $m_1 \geq 2$, inequality (5-3) is satisfied. Therefore, $k_2 = 1$.

Assume that we have shown $k_i = 1$ for $i = 1, 2, \ldots, m$.

Let $y_i = m_i - 1$ for $i = 1, 2, \ldots, m$, $y_{m+1} \neq 0$, and $y_{m+2} = y_{m+3} = \ldots = y_n = 0$.

Consider

$$(m_2 - 1)\frac{M}{m_1 m_2} + (m_3 - 1)\frac{M}{m_1 m_2 m_3} + \ldots + \frac{(m_m - 1)M}{m_1 m_2 \ldots m_m} + y_{m+1}k_{m+1}\frac{M}{m_1 m_2 \ldots m_m}$$

or

$$(m_2 - 1)m_3 \ldots m_n + (m_3 - 1)m_4 \ldots m_n + \ldots + (m_m - 1)m_{m+1} \ldots m_n + y_{m+1}k_{m+1}m_{m+2} \ldots m_n$$

which yields

$$(m_2 m_3 \ldots m_n) - (m_{m+1} \ldots m_n) + y_{m+1}k_{m+1}m_{m+2} \ldots m_n .$$

But $\frac{M}{m_1} = m_2 m_3 \ldots m_n$.

Thus the sum s is

$$s = \frac{M}{m_1} - m_{m+1} \cdots m_n + y_{m+1}k_{m+1}m_{m+2} \cdots m_n .$$

If

$$\frac{M}{m_1} \leq s < M \qquad (5\text{-}4)$$

condition (5-1) will be compromised. Expression (5-4) becomes

$$m_{m+1} \cdots m_n \leq y_{m+1}k_{m+1}m_{m+2} \cdots m_n < (m_1 - 1)\frac{M}{m_1} + m_{m+1} \cdots m_n$$

which upon substitution for M yields

$$m_{m+1} \cdots m_n \leq y_{m+1}k_{m+1}m_{m+2} \cdots m_n < [m_2 \cdots m_m(m_1 - 1) + 1]\, m_{m+1} \cdots m_n .$$

Removing the common factor in the above expression one obtains

$$m_{m+1} \leq y_{m+1}k_{m+1} < [m_2 \cdots m_m(m_1 - 1) + 1]\, m_{m+1} . \qquad (5\text{-}5)$$

If $m_{m+1} < k_{m+1} < (m_1-1)m_2 \cdots m_{m+1}$, (5-5) is satisfied by $y_{m+1} = 1$. If $k_{m+1} < m_{m+1}$ and $k_{m+1} > 1$, Lemma 2 guarantees that there exists a y_{m+1} such that

$$m_{m+1} < y_{m+1}k_{m+1} < 2m_{m+1}$$

but since

$$2 < [m_2 \cdots m_m(m_1 - 1) + 1] .$$

It is necessary that $k_{m+1} = 1$ or that

$$(m_1 - 1)\, m_2 \cdots m_{m+1} < k_{m+1} < m_1m_2 \cdots m_{m+1} .$$

To see that k_{m+1} cannot be in the range

$$(m_1 - 1)\, m_2 \cdots m_{m+1} < k_{m+1} < m_1m_2 \cdots m_{m+1} ,$$

consider $y_{m+1} = 1$ and $y_i = 0$ for $i \neq m + 1$. Then

$$s = B_{m+1}\, k_{m+1}m_{m+2} \cdots m_n$$

and

$$\frac{M}{m_1} \leq (m_1 - 1)m_2 \cdots m_{m+2} \cdots m_n < s < m_1m_2 \cdots m_{m+2} \cdots m_n = M$$

Again we have satisfied condition (5-4); it is necessary that $k_{m+1} = 1$.

Therefore, the only number system which effects the partition described is that system having $k_i = 1$ for $i = 1, \ldots, n$. This system is the mixed base number system.

Theorem 17. If the base moduli are ordered $m_1, m_2, \ldots, m_n$, there is only one number system with the property that the first coordinate partitions elements which represent consecutive integers into m_1 classes. That number system is the mixed base number system.

Proof: Theorem 16 shows that there exists but one number system which partitions the elements into those elements which represent integers in the range less than $(c + 1)\frac{M}{m_1}$ but greater than or equal to $c\,\frac{M}{m_1}$, where c is the first coordinate of the element. By definition this number system is the mixed base number system.

It remains to show that no number system exists which effects the same partitioning for $y_1 \neq c$. Assume such a number system exists with basis $\beta'_1, \beta'_2, \ldots, \beta'_n$. An element of the number system represents the integer

$$y_1B'_1 + y_2B'_2 + \ldots + y_nB'_n \mod M . \tag{5-6}$$

Again we may state

$$B'_n = k_n$$

$$B'_{n-1} = k_{n-1}m_n$$

$$\ldots$$

$$B'_j = k_j \prod_{i=j+1}^{n} m_i$$

$$\ldots$$

$$B'_i = k_1 \prod_{i=2}^{n} m_i$$

where $0 < k_j < \prod_{i=1}^{j} m_i$. The range imposed on k_1 is $0 < k_1 < m_1$.

Thus we have

$$B_1' = k_1 \quad M/m_1$$

and Expression (5-6) becomes

$$y_1 k_1 \frac{M}{m_1} \mod M \quad \text{when} \quad y_1 \neq 0,\ y_2 = \ldots = y_n = 0.$$

It will now be shown that there exists an integer c in the range $0 \leq c < m_1$ such that

$$c \frac{M}{m_1} \leq y_1 k_1 \frac{M}{m_1} \mod M < (c+1) \frac{M}{m_1} \tag{5-7}$$

cannot be satisfied for $0 < k_1 < m_1$, $0 < y_1 < m_1$, and $y_1 \neq c$.

Take $c = 0$.

Expression (5-7) becomes

$$0 \leq y_1 k_1 \frac{M}{m_1} \mod M < \frac{M}{m_1} \quad . \tag{5-8}$$

Condition (5-8) will be satisfied only if

$$y_1 k_1 \equiv 0 \mod m_1 \ . \tag{5-9}$$

Since $y_1 \neq 0$, (5-9) requires that $k_1 = 0$. This is the desired contradiction which completes the proof.

Collary 1: If the base moduli are ordered $m_1 = 2, m_2, m_3, \ldots, m_n$ there is one and only one number system with the property that the first coordinate partitions the elements into two classes, the representations of the integers less than $\frac{M}{2}$, and the representations of integers greater than or equal to $\frac{M}{2}$.

Proof: This corollary follows from Theorem 17 with $m_1 = 2$.

One now asks whether a number system with base moduli $m_1, m_2, \ldots, m_n$ exists which partitions elements which represent consecutive integers into m_j classes with the first coordinate where $j \neq 1$?

The number of elements having the same j-th coordinate in any system having m_j as a base prime is

$$\frac{M}{m_j} = \prod_{\substack{i=1 \\ i \neq j}}^{n} m_i$$

Likewise, the number of elements associated with a particular value of the first coordinate is

$$M/m_1 = \prod_{i=2}^{n} m_i$$

The number of elements in the two cases differ and the greatest common multiple is one. Therefore, the answer to the question posed in the preceding paragraph is negative.

A similar argument shows that no number system with base moduli $m_1, m_2, \ldots, m_n$ where $(m_1, m) = (m_2, m) = \ldots = (m_n, m) = 1$ exists which partitions with the first coordinate elements which represent consecutive integers into two classes, one class the elements of which represent integers less than M/m and the other class representing integers greater than or equal to M/m. The argument follows:
Since M is relatively prime to m, $m \nmid M$, but $m \mid M - \ell$ when $0 < \ell < m$. Let $M - \ell = d_m$. If $m < m_1, m_1 \nmid M - \ell$. If $m > m_1$

$$m_1 \nmid M - \ell \quad \text{for} \quad cm_1 < \ell < (c+1)m_1 .$$

Suppose $km_1 = \ell$, then $m - \ell = m_1 m_2 \ldots m_n - m_1 k$ but $(m, m_1) = 1$; therefore, $m \nmid M - \ell$ which is a contradiction. The only related number system which produces a partitioning is the mixed base number system and; therefore, the proof is completed.

From the theorems and arguments advanced thus far in this chapter, one concludes that the mixed base number system is the only number system which partitions the elements representing consecutive integers. In particular only the mixed base system with m_1 even partitions the elements representing consecutive integers into two groups—(1) elements corresponding to positive integers and (2) elements representing complements. Thus if one wishes to use a number system to determine the sign of the residue element, he will find it necessary to use the mixed base system.

5.2 Number Systems Allowing Sign Detection with Fewer Than n-Carries

It has been suggested that the use of number systems which are neither strictly residue nor strictly mixed base might ease the carry situation in sign determination. Such systems are those in which a certain number of carries are eliminated from the operation of expressing a vector in the mixed base system.

As developed in the chapter concerning the Carry Algorithm, a vector X with coordinates $(x_1, x_2, \ldots, x_n)$ relative to the basis $< \beta_1, \beta_2, \ldots, \beta_n >$ is expressed with coordinates $(y_1, y_2, \ldots, y_n)$ relative to the mixed base basis $< \alpha_1, \alpha_2, \ldots, \alpha_n >$ by determining the $Q = ||q_{ij}||$ matrix and following with the matrix multiplication $X \cdot Q = Y$. The elements of the Q matrix are governed by the equation

$$\beta_i = q_{i1}\alpha_1 + \ldots + q_{in}\alpha_n \quad \text{for} \quad i = 1, 2, \ldots, n . \qquad (5\text{-}10)$$

The Y so obtained will not in general be a linear form with restricted coefficients. Carrys must then be propagated from each position to the next more significant position. The general conversion will require up to n-1 carries. To reduce the maximum possible number of carries which can occur by say m-1 carries, it will be necessary and sufficient to guarantee

$$y'_k < m_k \quad \text{for} \quad k = n-m, \ldots, n \; . \tag{5-11}$$

This is in turn equivalent to the condition

$$q_{ik} = \delta_{ik} \quad \text{for} \quad i,k = n-m, \ldots, n \; . \tag{5-12}$$

Condition (5-12) may be expressed as

$$\beta_k = \alpha_k + \sum_{i=1}^{n-m-1} c_{ik}\alpha_i \quad \text{for} \quad k = n-m, \ldots, n \; . \tag{5-13}$$

If $||a_{ij}||$ is the array of the mixed base basis vectors, consider the m x m sub-array in the lower right corner. Equation (5-13) states that this sub-array must be preserved in $||b_{ij}||$, the array of the β vectors. The only other requirement is that $||b_{ij}||$ be triangular. Other considerations which affect the selection of the remaining elements in the β array stem from a desire to simplify the carry structure in the β system.

Since $a_{ij} = b_{ij}$ for $i, j = n-m, \ldots, n$ the carry structure in the last m positions is fixed. Carries from the k-th components to the j-th components where

$$k = n-m, \ldots, n$$

$$j = 1, 2, \ldots, n-m-1$$

can be prevented by the following constraint

$$b_{ij} = 0 \quad \text{for} \quad i = n-m, \ldots, n$$
$$j = 1, 2, \ldots, n-m-1 .$$

Further carries can be eliminated by making the upper right partition of the β array the identity matrix.

Since $a_{ij} = b_{ij}$ for $i, j = n-m, \ldots, n$ carries from the ℓ-th component to the $(\ell-1)$-st component will occur for $\ell = n-m+1, \ldots, n$. Therefore, at least $m-1$ carries will occur in the β system. The total number of carries in the β system for a subtraction followed by sign detection is at least as great as for subtraction in the residue number system and conversion to the mixed base system.

Such number systems do not appear practical, for nothing is gained in addition and sign detection. In fact, from the Multiplication Algorithm of Chapter III, it is clear that much speed is sacrificed in multiplication.

CHAPTER VI

SUMMARY AND CONCLUSIONS

The general question of the algebraic properties of the residue number system was treated. By considering the set of elements of the residue number system as a pseudo-vector space (the R-space) it was possible to define and describe the properties of a whole class of number systems related to the residue number system. Meaningful definitions of linear independence, linear transformations, and matrix multiplication were formulated. However, such vector space concepts as rank and row echelon form have no analogous interpretations in the R-space.

It was proven that any triangular array of vectors will define a number system related to the residue number system if the elements on the principal diagonal are relatively prime to the associated base modulus. The Carry Algorithm and a variation, the Borrow Algorithm were given which allow arithmetic operations in number systems related to the residue number system.

The mixed base number system has been known to afford solutions to the problems of sign determination, magnitude comparison, and additive overflow. Solutions to the problems of multiplicative overflow digit fill-in, and division were shown. The R-space analysis led quite naturally to the known solutions as well as the new solutions of the problems of the residue number system. Solutions are possible because the mixed base number system partitions elements representing consecutive integers with the first coordinate. Thus when employing the mixed base number system the magnitude of the first coordinate gives the sign. The partitioning places

weights on the coordinates and allows magnitude comparison. Since overflow may be regarded as a variation of magnitude comparison, overflow problems may be solved. With these problems solved division by the Euclidean Algorithm is possible.

The mixed base number system was proven to be the only number system related to the residue number system which incorporates the desired partitioning. Further, number systems were investigated which would require fewer carries for sign detection than the residue number system and possess fewer arithmetic carries than the mixed base system. It was shown that no net savings would be possible.

The number systems related to the residue number system may be classified according to the maximum length of carry chains involved in addition. If this is done it is noted that the residue number system with carry chains of zero length possesses the simplest carry structure while the mixed base system with a possible $n-1$ carries is the most complex. The desirable properties and the problems of the residue number system arise from its carryless structure. Similarly the solution to the residue number system problems exist because the mixed base system involves carry chains of maximum length. Thus one concludes that of all the number systems related to the residue number system only the residue and mixed base systems are of primary interest.

The solutions to the problems of the residue number system depended upon the partitioning properties of the mixed base number system. In a certain sense these solutions are unique; therefore, it is not anticipated that the residue number system will achieve wide application in general purpose digital computers.

BIBLIOGRAPHY

Birkhoff, G. and MacLane, S. A Survey of Modern Algebra. New York: The Macmillan Company, 1941.

Cheney, P. W. A Digital Correlator Based on the Residue Number System. Technical Document LMSD-702670, Lockheed Aircraft Corporation, Palo Alto, Calif., 1960.

Garner, H. L. "The Residue Number System." IRE Trans. PGEC, Vol. EC-8, June 1959, p. 140-7.

Hardy, G. H. and Wright, E. M. An Introduction to the Theory of Numbers. London, England: Oxford University Press, 1956.

McCoy, N. H. Rings and Ideals. Buffalo, New York: The Mathematical Association of America, 1948.

Svoboda, A. Rational Number Systems of Residual Classes. Stroje Na Zpracovani Informaci, Sbornik, V, 1957.

van der Waerden, B. L. Modern Algebra. New York: Frederick Ugar Publishing Company, Vol. I and II, 1950.