

PARTICLE ACCELERATORS

Volume 19, Numbers 1-4 (1986)

Proceedings of the
**WORKSHOP ON ORBITAL DYNAMICS AND
APPLICATIONS TO ACCELERATORS**

March 7-12, 1985

Lawrence Berkeley Laboratory
Berkeley, California

Organizing Committee

Jay Marx, Chairman
Lawrence Berkeley Laboratory

Anastassios Bountis
Clarkson University
Alex Chao
Stanford Linear Accelerator Center
Max Cornacchia
Lawrence Berkeley Laboratory
George Dell
Brookhaven National Laboratory
Alex Dragt
University of Maryland
Michael Harrison
Fermi National Accelerator Laboratory

Robert Kee
Sandia National Laboratory
Michael Lieberman
University of California, Berkeley
Thomas Nash
Fermi National Accelerator Laboratory
Ronald Peierls
Brookhaven National Laboratory
David F. Sutter
U.S. Department of Energy

This workshop was supported by the
High Energy Physics Division,
Office of High Energy and Nuclear Physics,
Office of Energy Research,
United States Department of Energy

NUMERICAL TECHNIQUES FOR THE STUDY OF LONG-TIME CORRELATIONS†

CHARLES F. F. KARNEY

Plasma Physics Laboratory, Princeton University, Princeton, NJ 08544

(Received March 7, 1985)

In the study of long-time correlations, extremely long orbits must be calculated. This may be accomplished much more reliably using fixed-point arithmetic. Use of this arithmetic on the Cray-1 computer is illustrated.

I. INTRODUCTION

There has been considerable interest recently in simple dynamical systems which exhibit very complicated behavior. One of the simplest such systems is a two-dimensional area-preserving map. This shares many of the properties of Hamiltonian systems with two degrees of freedom. The phase space for these systems is typically divided into integrable and nonintegrable (or stochastic) portions. An interesting problem is the behavior of an orbit in the nonintegrable portion of phase space when it approaches an integrable portion. This can introduce very long-time correlations into the stochastic orbits. The first systematic study of this problem was given by Channon and Lebowitz,¹ who studied orbits in the Hénon map.² This study was based on 7750 orbits of length 10^4 . However, subsequent studies have recognized that longer orbits must be considered in order to determine the long-time behavior accurately. For example, in their work on the whisker mapping Chirikov and Shepelyansky³ studied a single orbit of length 10^8 . More ambitious calculations were performed by the author⁴ on the periodic quadratic mapping

$$Q: \quad y' - y = g(x), \quad x' - x = y', \quad (1)$$

where

$$g(x) = \begin{cases} 2(x^2 - K), & \text{for } x_{\min} \leq x < x_{\max}, \\ g(x \pm L), & \text{otherwise,} \end{cases}$$

and $L = x_{\max} - x_{\min} > 0$. In this study 1600 orbits of length 2×10^9 were used (i.e., a total of 3.2×10^{12} iterations).

II. THE NUMERICAL PROBLEM

In a calculation of this magnitude, we must ask whether the results obtained on a computer have any relevance to the study of the exact system. The problem arises

† This work was supported by the U.S. Department of Energy under Contract DE-AC02-76-CHO-3073.

because only a finite set of numbers can be represented on a computer. Consequently, *every* orbit in a numerical mapping is periodic. The numerical mapping does not give the generic behavior for an exact mapping in which periodic orbits are a set of measure zero. Nevertheless, useful information can be obtained if the average period T of orbits in the numerical mapping exceeds the time in which we are interested. On the other hand, the numerical calculations are useless if T is less than the time in which we are interested.

For a two-dimensional map, such as Eq. (1) implemented in floating-point arithmetic, we estimate that $T \sim 2^{p/2}$, where p is the number of bits of precision. On the Cray-1, we have $p = 48$ and $T \sim 10^7$. Therefore, single-precision floating-point arithmetic on the Cray-1 cannot be used to study orbits of length 10^9 . One possible solution (the brute-force approach) is to go to double precision. This extends T to about 10^{14} but at a cost of a factor of 2–4 in speed. It is preferable, however, to understand what defects in the floating-point number system cause T to be as short as it is, and then to employ a system of arithmetic which doesn't have such defects.

We are approximating a mapping Q (the "exact" mapping) with another mapping Q^* (the realization of Q on a computer, the "numerical" mapping). We can estimate the error in a single iteration of the mapping as 2^{-p} . As we iterate the mapping the error grows. However, we can seek to control the error by making sure that Q^* has some of the same properties as Q . Indeed if Q^* has enough of the "interesting" properties of Q , we might tolerate quite a large error in a single iteration.

(A parallel situation exists in the approximation of the collision operator in a plasma by the Landau collision operator. This is known to be in error by about 5%. However, because the Landau operator conserves all the quantities conserved by the exact collision operator—number, momentum, and energy—and because it has an H theorem, we are sure that the errors won't affect anything "important." Indeed, for these reasons, most plasma physicists are happy to regard the Landau collision operator as exact!)

Perhaps the most important property of Q is that it is area-preserving. It is known that a small amount of dissipation greatly alters a mapping. Now Q^* is a mapping defined on a discrete set of numbers so that the area-preserving property has to be translated into the analogous property for discrete mappings. However, because the floating-point number system is nonuniform, this property is very complicated and consequently difficult to implement. On the other hand, if a uniform number system is used, then area-preservation in Q corresponds to the mapping Q^* being *one-to-one*. Such a number system is implemented on most computers and is called the "fixed-point" number system.

III. FIXED-POINT NUMBERS

Floating-point numbers are conventionally represented as $2^m \times f$, where $\frac{1}{2} \leq f < 1$ and m can vary (the binary point can *float*). The fraction f and the exponent m are then stored in different parts of the computer word. Floating-point numbers

are ideal for representing numbers which may be very small or very large. In typical mapping calculations, this flexibility is not needed. The numbers representing the coordinates are usually bounded, so we do not need to be able to represent very large numbers. Furthermore, the added precision available for small floating-point numbers is wasted since these are often added to much larger numbers.

Fixed-point numbers are also represented as $2^m \times f$, but now $0 \leq f < 1$ and m is fixed. (Note a possible confusion in the terminology: *fixed-point* numbers have nothing to do with the *fixed points* of a mapping.) Now only f need be stored in the computer word, and it is the programmer's responsibility to remember m . (Because a whole word is used to store f , it is often possible to increase the precision. Thus on a PDP-10, 36 bits are available for f , which is considerably more than the 27 bits used to represent the fraction in floating-point numbers on that machine.)

Addition and subtraction of fixed-point numbers are exact. Thus, if Eq. (1) is implemented in fixed-point arithmetic to give a numerical mapping Q^* , then this can be represented by Eq. (1) with the exact $g(x)$ replaced by an approximation, $g^*(x)$. Furthermore, Q^{*-1} exists because the operations in Q^* can be reversed to give x and y in terms of x' and y' . In fact, there is a way to compute Q^{*-1} numerically without having to program the reverse operations. If we define an involution ($J^2 = \text{identity}$)

$$J: \quad y' = -y, \quad x' = x - y,$$

then we can show that $Q^{*-n} = J^{-1}Q^{*n}J$, which may be exactly computed because J can be exactly carried out in fixed-point arithmetic. Thus each point on the plane has a unique successor (given by Q^*) and a unique predecessor (given by Q^{*-1}), and the mapping is one-to-one. Such a fixed-point mapping is a *permutation* on phase space. In contrast, a floating-point mapping is a *many-to-one* mapping, or a *function* on phase space.

One-to-one mappings have been used in the study of dynamical systems by Miller and Prendergast⁵ and by Rannou.⁶ These authors implemented the mappings with integer arithmetic and used rather coarse representations of phase space. Thus Rannou divided phase space into a maximum of 800×800 cells. (A fixed-point mapping implemented on the Cray-1 has $2^{48} \times 2^{48}$ cells.) However, these studies are useful for providing theoretical results about permutations. In particular, Rannou⁶ considers random permutations of N points defined as the ensemble of all possible such permutations. She shows that the average period of orbits is $\frac{1}{2}(N+1)$. The average period of orbits in a random function is given by Knuth⁷ as $\sqrt{\pi N/8} + \frac{1}{3}$. The reason that permutations have longer orbits is that they practice collision avoidance. The only way such an orbit can become periodic is by landing on the initial point. In a random function, an orbit can become periodic by landing on any of its previous points. Now the mappings describing dynamical systems are definitely not random. In particular they often possess symmetries (for example, the symmetry defined by the involution J connecting forwards and backwards trajectories). Rannou finds⁶ that the average length of the orbits of a random symmetric permutation is reduced to $O(N^{1/2})$. We

conjecture that a similar phenomenon occurs with symmetric functions reducing the average length of the orbits to $O(N^{1/4})$. Substituting $N = (2^{48})^2$, which is appropriate for a two-dimensional mapping on the Cray-1, we find that $T \sim 10^{14}$ with fixed-point arithmetic and $T \sim 10^7$ with floating-point arithmetic. Clearly, fixed-point arithmetic allows us to examine much longer orbits.

IV. OPERATIONS ON FLOATING-POINT NUMBERS

Let us briefly describe how to perform some useful operations on fixed-point numbers. We begin with the elementary operations: Addition and subtraction are performed with the same instructions as for integer addition and subtraction. Multiplication by a power of two $2^{\pm m}x$ may be accomplished by a left/right shift of x by m bits (possibly with sign extension). The computation of the integer part, $\text{int}(x) \equiv [x]$, and the fraction part, $\text{fract}(x) \equiv x - [x]$, of a number may be performed by ANDING the computer word with appropriate masks. (The computation of $\text{int}(x)$ for floating-point numbers needs a sequence of instructions in FORTRAN: `xi = aint(x); if (xi.gt.x) then xi = xi - 1.0.`)

The method for multiplying by an arbitrary number depends on the computer. For instance, on a PDP-10 the MUL instruction multiplies two 36-bit numbers to give the 72-bit product. This result can be shifted to align the binary point with the assumed position within the word. Similarly, on the mc68000 microprocessor the MULS instruction gives the 32-bit product of two 16-bit quantities. (Higher-precision multiplication can be performed with a sequence of these instructions.) The situation is slightly different on the Cray-1. When two numbers with zero exponent fields are multiplied using the floating-multiply instructions, the product of the fraction fields is returned with no normalization performed. Thus, if we represent fixed-point numbers as a Cray-1 word with the binary point 48 places from the right, then we can multiply two numbers in $[0, 1)$ with the rounded floating-multiply instruction *r.

Special functions may be calculated by converting the fixed-point number to floating point, calling a library routine for the special function, and converting the result back to fixed point. Alternatively, a subroutine calculating the special function directly with fixed-point arithmetic may be written. The program METAFONT by Knuth⁸ contains a complete collection of routines for the elementary functions assuming a fraction size of 16 bits. An interesting shortcut is available for the calculation of random numbers. Usually, there is a library routine which returns a random floating-point number uniformly distributed in $[0, 1)$. The fraction field of this floating-point number is uniformly distributed in $[\frac{1}{2}, 1)$; thus a uniformly distributed fixed-point number may be obtained by taking all but the first bit of the fraction.

V. EXAMPLES OF MAPPINGS

Care must be taken when implementing an area-preserving mapping in fixed-point arithmetic to ensure that the numerical mapping is one-to-one. We have

seen that the mapping Q may be implemented in a straightforward way. What about other mappings? The problem is well illustrated by the mapping $x' = 2x$, $y' = \frac{1}{2}y$. If this is coded as it stands, then on each iteration the least significant bit of y is lost and the mapping is clearly not invertible. Some way is therefore needed for remembering that lost bit.

We begin by observing that a succession of nonlinear shifts

$$x' = x + f(y), \quad y' = y + g(x')$$

is in a form that is invertible. The trick is to combine mappings of this form to produce the desired mapping. Thus the scaling mapping

$$x' = sx, \quad y' = y/s$$

may be implemented as

$$x^* = x - y/s, \quad y^* = y + (s-1)x^*, \quad x' = x^* + y^*, \quad y' = y^* - (s-1)x'/s.$$

Similarly, the rotation

$$x' = \cos \theta x - \sin \theta y, \quad y' = \sin \theta x + \cos \theta y$$

can be coded as

$$x^* = x - \alpha y, \quad y' = y + \beta x^*, \quad x' = x^* - \alpha y',$$

where $\alpha = (1 - \cos \theta)/\sin \theta$ and $\beta = \sin \theta$. Note that α diverges for $\theta \rightarrow \pi$. But this is no serious limitation because rotations by multiples of $\frac{1}{2}\pi$ can be realized exactly merely with sign changes; thus we can restrict $|\theta| \leq \frac{1}{4}\pi$. Interestingly, this form for the rotation involves only three multiplications, and so may be faster than the "standard" form which involves four. If θ is small, then the rotation can be approximated by⁹

$$x' = x - \epsilon y, \quad y' = y + \epsilon x',$$

where $\epsilon = 2 \sin \frac{1}{2}\theta$. In fact, with exact arithmetic this produces a slightly eccentric ellipse $x^2 - \epsilon xy + y^2 = \text{const}$.

VI. REALIZATION ON A CRAY-1

The biggest disincentive to using fixed-point arithmetic is the absence of any support for this arithmetic in languages like FORTRAN. Usually, one has to resort to assembly language to utilize this arithmetic. However, it is possible to make a substantial gain in speed by coding in assembly language, so the effort is often worth it. (Mapping calculations tend to benefit from hand coding in assembly language because most of the running time is spent in a small mapping subroutine.) In order to illustrate the benefits, we show how the mapping Q may be written in assembly language CAL on the Cray-1. The reader is referred to the CAL manual¹⁰ for further details on the language, and the CFT manual¹¹ (Appendix F) for details on how to interface assembly-language routines to a FORTRAN program.

Fixed-point numbers are represented as a 64-bit word in twos-complement notation with the binary point 48 places from the right. The floating-point multiply instruction *r can multiply two such fixed-point numbers provided they lie in the range [0, 1). It is convenient to scale the variables in Eq. (1) so that the mapping is periodic in x and y with period 1. The running time is slightly shortened if the inverse of the mapping is used. The mapping then becomes⁴

$$Q^{*-1}: \quad x' = x - y + \frac{1}{2} \pmod{1}, \quad y' = y - 2^m(ax'^2 - bx' + c) \pmod{1},$$

where a, b, c are all in $[0, 1)$ and m is non-negative. In addition, we wish to keep track of when x or y leave the unit square. The speed can be increased by following 64 particles at once (to make use of the vectorization capabilities of the Cray-1) and by iterating the mapping many times in one call. Thus a FORTRAN realization of this procedure (using floating-point arithmetic) reads

```

subroutine quade (n, x, y, lv, a, b, c, m)
parameter (l = 64)
logical lv
dimension x(l), y(l), xi(l), yi(l), lv(l)
do 1 i = 1, l
    lv(i) = .false.
1 continue
do 2 j = 1, n
    do 2 i = 1, l
        x(i) = x(i) - y(i) + 0.5
        xi(i) = aint (x(i))
        xi(i) = cvmgm (xi(i) - 1.0, xi(i), x(i))
        x(i) = x(i) - xi(i)
        y(i) = y(i) - 2.0 ** m * (a * x(i) ** 2 - b * x(i) + c)
        yi(i) = aint (y(i))
        yi(i) = cvmgm (yi(i) - 1.0, yi(i), y(i))
        y(i) = y(i) - yi(i)
        lv(i) = lv(i) .or. (xi(i) .ne. 0.0) .or. (yi(i) .ne. 0.0)
2 continue
return
end

```

This iterates Q^{*-1} n times, and returns x and y . (The CFT function `cvmgm(a, b, c)` returns a if $c < 0$ and b otherwise.) The variable lv returns TRUE if the particle left the unit square during any of these n iterations.

Let us see how this can be coded using fixed-point arithmetic in CAL. For brevity, the loading and storing of the arguments is omitted. At this point we have stored the vector length (64) in the register vl , the vectors x and y in the vector registers $v2$ and $v1$, the constants a, b , and c in $s2, s3$, and $s4$, the count n in $a1$, and the shift m in $a4$. We begin with some initialization:

```

v0    0          v0 ← 0 (used for ORing with  $x$  and  $y$ )
s1    1
s1    s1 < 47   0.5

```

$s6$	$-s1$	$s6 \leftarrow -0.5$
$s5$	<48	$s5 \leftarrow$ fraction mask
$a1$	$-a1$	$a1 \leftarrow -n$ (flip sign of count)

We next turn to the main loop. Only the j loop in the FORTRAN code needs to be explicitly written. The i loop is implicitly performed by the vector instructions. The calculation is carried out entirely in registers. However, at the end of one iteration, x and y are in different registers ($v7$ and $v4$). It is therefore necessary to repeat the coding with interchanged registers to get x and y back to $v2$ and $v1$. The test for the particle having left the unit square lv is given by ORING the successive x 's and y 's together into register $v0$. At the end we check whether the integer part of $v0$ is nonzero. We are able to postpone the operation $y \leftarrow \text{fract}(y)$ until the end. (With floating-point arithmetic, this would result in a loss of precision, so it is necessary to extract the fractional part with each iteration.)

<i>loop</i>	$v3$	$s6 + v1$	$y - 0.5$
	$v5$	$v2 - v3$	$x \leftarrow x - y + 0.5$
	$v7$	$s5 \& v5$	$x \leftarrow \text{fract}(x)$
	$v6$	$s2 * r v7$	ax
	$v2$	$v0 ! v5$	$lv \leftarrow lv \text{ OR } \text{int}(x) \neq 0$
	$v3$	$v6 * r v7$	ax^2
	$v4$	$s4 + v3$	$ax^2 + c$
	$s0$	$s3 * r s4$	use multiply unit for 1 cycle
	$v0$	$s3 * r v7$	bx
	$v6$	$v4 - v0$	$(ax^2 - bx + c)$
	$v5$	$v6 < a4$	$2^m(ax^2 - bx + c)$
	$v4$	$v1 - v5$	$y \leftarrow y - 2^m(ax^2 - bx + c)$
	$v0$	$v2 ! v4$	$lv \leftarrow lv \text{ OR } \text{int}(y) \neq 0$
	$a1$	$a1 + 1$	increment loop counter
	$v3$	$s6 + v4$	now repeat everything with
	$v5$	$v7 - v3$	$v1 \rightleftharpoons v4$ and $v2 \rightleftharpoons v7$
	...		
	$a1$	$a1 + 1$	
	$a0$	$a1$	
	<i>jam</i>	<i>loop</i>	jump back if more to do

We end by taking the fractional part of y and by testing $v0$ for a nonzero integer part.

$v3$	$s5 \& v1$	$y \leftarrow \text{fract}(y)$
$a3$	48	fraction shift
$v6$	$v0 > a3$	shift out fraction part of $v0$
vm	$v6, n$	$lv \leftarrow \text{int}(y) \neq 0 \text{ OR } \text{int}(x) \neq 0$
$s1$	vm	transfer to $s1$

At this point x and y are in $v2$ and $v3$, and lv is in $s1$.

There are two sources of speed on Cray-1. The first is the ability to process vectors. This enables a given functional unit to produce one result every cycle

(12.5 ns). Utilizing this feature in CAL is relatively easy. The second source of speed is the ability of different functional units to be operating at the same time. Depending on the degree of overlap, this can speed up the program by a factor 1.5–3. However, taking advantage of this feature is made difficult by a complicated set of rules for when a particular instruction can issue. An extremely useful tool is the timing code CYCLES,¹² which produces a detailed timing analysis of the code. The application of this code to the main loop of the mapping routine gives:

			W	D	I	C	O	F	R
<i>loop</i>	<i>v3</i>	<i>s6 + v1</i>			0	5	64	68	69
	<i>v5</i>	<i>v2 - v3</i>	68	25	69	74	133	137	138
	<i>v7</i>	<i>s5 & v5</i>	4	10	74	78	138	142	142
	<i>v6</i>	<i>s2 * r v7</i>	3	10	78	87	142	146	151
	<i>v2</i>	<i>v0 ! v5</i>	63	07	142	146	206	210	210
	<i>v3</i>	<i>v6 * r v7</i>	8	25	151	160	215	219	224
	<i>v4</i>	<i>s4 + v3</i>	8	10	160	165	224	228	229
	<i>s0</i>	<i>s3 * rs4</i>	58	01	219	226			
	<i>v0</i>	<i>s3 * r v7</i>			220	229	284	288	293
	<i>v6</i>	<i>v4 - v0</i>	8	10	229	234	293	297	298
	<i>v5</i>	<i>v6 < a4</i>	4	10	234	240	298	302	304
	<i>v4</i>	<i>v1 - v5</i>	69	27	304	309	368	372	373
	<i>v0</i>	<i>v2 ! v4</i>	4	10	309	313	373	377	377
	<i>a1</i>	<i>a1 + 1</i>			310	312			
	<i>v3</i>	<i>s6 + v4</i>	62	05	373	378	437	441	442

For each instruction is given the wait time, the delay code, the issue time, the chain slot time, the operand ready time, the functional unit ready time, and the result ready time. The times are all in units of the clock of the Cray-1, namely 12.5 ns. The delay code is an octal number describing why the instruction could not issue. The meanings of the bits in this code are

- 1 functional unit not ready
- 2 result register not ready
- 4 operand register not ready
- 10 waiting for chain slot
- 20 missed chain slot

From the last line in the timing analysis, we see that one complete iteration (for 64 particles) takes 373 cycles or 73 ns/particle. Since there are 12 vector instructions in one iteration, the machine is computing results at the rate of about 2/cycle. Very little improvement is possible beyond this because the vector-add unit is busy 91% of the time. For comparison, if the same mapping is implemented with floating-point in FORTRAN, it takes 410 ns/particle/iteration, a factor of 5.6 slower.

The complexity of timing on the Cray-1 can be understood by considering the “dummy” scalar multiply which issues at 219. This produces no useful result but

causes the next vector-multiply instruction to issue one cycle later. Because of this, the vector-add unit is ready at the chain-slot time for this instruction (229), and the next instruction chains to this one. Without this dummy instruction, the vector multiply would issue at 219, the vector add would then miss the chain slot and have to wait until time 292 to issue; i.e., there would have been additional delay of 63 cycles.

VII. CONCLUSIONS

In this paper, we have shown that the area-preserving nature of mappings implemented on a computer can be preserved by using fixed-point arithmetic. Because of this, much longer orbits can be studied. The precision is comparable to that of floating-point numbers; however, roundoff error is much easier to control with fixed-point arithmetic. At present, access to fixed-point numbers is through assembly language. However, it is often faster than floating-point arithmetic. Indeed, since floating-point instructions are not available on several modern microprocessors, fixed-point arithmetic would be a natural way of studying mappings on such devices.

REFERENCES

1. S. R. Channon and J. L. Lebowitz, in *Nonlinear Dynamics*, Annals of the New York Academy of Sciences **357**, 108 (New York, 1980); S. R. Channon, Ph.D. thesis, Rutgers University, 1981.
2. M. Hénon, *Quarterly App. Math.*, **27**, 291 (1969).
3. B. V. Chirikov and D. L. Shepelyansky, *Physica*, **13D**, 395 (1984).
4. C. F. F. Karney, *Physica*, **8D**, 360 (1983).
5. R. H. Miller and K. H. Prendergast, *Astrophys. J.*, **151**, 699 (1968).
6. F. Rannou, *Astron. and Astrophys.* **31**, 289 (1974); F. Rannou, Ph.D. thesis, University of Nice, 1972.
7. D. E. Knuth, *The Art of Computer Programming* (Addison Wesley, 1981) 2nd ed., vol. 2.
8. D. E. Knuth, *The METAFONTbook* (Addison Wesley, 1985).
9. M. Minsky, in MIT Artificial Intelligence Laboratory Report AIM-239, (1972).
10. *Cray Assembly Language Reference Manual* (Cray Research, Inc., 1980).
11. *CFT, the Cray-1 FORTRAN Compiler*, Publication SR-0009 (Cray Research, Inc., 1984).
12. H. L. Nelson, LLL Report UCID-30179, Revision 2 (1981).