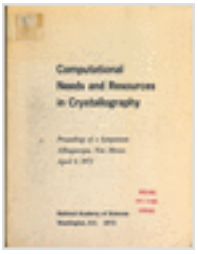

Computational Needs and Resources in Crystallography: Proceedings of a Symposium, Albuquerque, New Mexico, April 8, 1972. (1973)

Pages
144

Size
8.5 x 11

ISBN
0309296897

Division of Chemistry and Chemical Technology; National Research Council

Computational Needs and Resources in Crystallography

Proceedings of a Symposium

Albuquerque, New Mexico

April 8, 1972

Division of Chemistry and Chemical Technology
National Research Council

with support from the National Science Foundation
Contract No. NSF-C310, Task Order No. 233

NATIONAL ACADEMY OF SCIENCES
Washington, D.C. 1973

NOTICE

IN MEMORIAM

Walter Clark Hamilton

February 16, 1931 - January 23, 1973

CONTENTS

PREFACE

On April 8, 1972 a symposium was held on the topic "Computational Needs and Resources in Crystallography". This symposium, sponsored by the Division of Chemistry and Chemical Technology of the National Academy of Sciences - National Research Council, was arranged to follow immediately the Winter Meeting of the American Crystallographic Association and was attended by 111 participants, most of them from that meeting (see Appendix 2). Financial support was provided by the Chemistry Section of the National Science Foundation under Contract No. NSF-C310 Task Order No. 233. The aims of the symposium were similar to those of the conference on "Computational Support for Theoretical Chemistry" sponsored by the Division's Committee on Computers in Chemistry in Bethesda, Maryland, May 8-9, 1970, also with support from the Chemistry Section of the National Science Foundation.

The motivation was spelled out in the call to the symposium as follows:

> Crystallographers are among the major users of computers in chemistry and physics; they have also had a long history of innovative uses of computers, both in computation and in on-line control of experiments. In view of the continually expanding number of scientists engaged in structural crystallography, the increasing amount of computer usage by these scientists, and the realization of shrinking research budgets, the time seems ripe to bring together a group of crystallographers and representatives of the federal funding agencies to explore together questions such as the following:
>
> 1. What are the computing needs of crystallographers, now and during the next decade? Are these needs being adequately met?
> 2. Are there new hardware, software, and theoretical crystallographic developments that may induce marked changes in computational methods?
> 3. Are the present methods of funding and operation of computer centers working to everyone's satisfaction?
> 4. Are there substantial benefits to be derived from the establishment of regional computer centers or computer networks for large-scale or special-purpose computations? Should a task force be commissioned to explore such possibilities in depth?
>
> The answers to all these questions will surely not be completed in a one-day session. It is hoped however that the presentations of the invited speakers and the open discussion will result in clear statements of the present position and the spectrum of existing opinion, and will sharply define those questions that may form the

basis for action by the relevant part of the scientific community.

The symposium did not have as its outcome any formal recommendations, per se, though some consensus can be inferred on various points. It did reflect accurately the viewpoints of a fair cross-section of the crystallographic community, both in the prepared lectures and in the spontaneous discussion. A transcript of this discussion is included here, along with manuscripts of the prepared talks. The discussion has received a minimum amount of editing from the participants and the editors, and gives a flavor of the current state of thinking of crystallographers concerning their computing activities.

The sponsors of the symposium hope this document will be useful to all concerned with the support of scientific computing, as an example of where computation stands today in one segment of physical science.

(This preface was prepared by Walter C. Hamilton, who organized and gave direction to the Symposium. The following summary was drafted at Dr. Hamilton's request by R. A. Young.)

SUMMARY

Present Activity in Crystallographic Computing in the U.S.A.

Present crystallographic uses of computers include instrument control, data handling and reduction, application of direct methods of structure solution, refinement of both general and detailed features of structural models, and display of the results in graphic form including three-dimensional dynamic displays utilized in an interactive mode.

The total annual cost of crystallographic computing being done in the U.S.A. is estimated at 10 million dollars of which, probably, only about half is charged directly to crystallography budgets. Those crystallographers heavily engaged in structure determination spend about 80% of their computer time on refinement of structural models.

The computing cost per structure varies too greatly to be a meaningful figure for planning purposes; costs in Hamilton's survey reported at this symposium range from less than $1,000 to more than $100,000, the differences being attributable to the number of structural parameters being refined and this number, in turn, being dependent both on the number of atoms in the unit cell and on the detail in which the refinement was to be carried out. It is a truism that the cost per calculation goes down as the size of the computer goes up. A rule-of-thumb suggested is that a factor of two increase in machine cost is accompanied by a factor of eight increase in computational speed in solving crystallographic problems. However, the actual gain depends on the particular calculation and machine combination. Further, it refers only to computation time and not to input-output time.

Hamilton's survey shows that crystallographers, as a group, have experience with a great variety of computer facilities. The size of the computers used ranges from stripped-down minicomputers to the largest computers now generally available. Of the responding crystallographers, 62% reported experience with remote terminals, 37% had used computers at locations other than their home installations, and 42% had made use of program lists.

Needs

Although no poll was taken, the combination of the papers presented and the vigorous discussion of them suggest the existence of a fair consensus on the following needs, perhaps among others.

1. There is a need for computing capacity one to two orders of magnitude larger than is presently available. Refinements of protein structures are now becoming possible, and such refinements will make significant demands on large-core, very fast computers. Such large-capacity computers are also needed for application of direct methods of structure solution to large structures, i.e., containing more than 40 to 50 atoms in the asymmetric unit. A third area in which the extra capacity is needed is in refinement of

structural models accurate in greater detail; such detailed information may be important to molecular biological function, on the one hand, or to fundamental understanding of the solid state, on the other.

In Hamilton's survey, 22% of the respondents reported that their work is seriously limited by lack of adequate computer capacity and another 21% reported that they are moderately unhappy with their computer resources. The capacity for full-matrix least-squares refinement of up to 240 parameters is typical of one of the major types of large computers now in use (e.g., CDC 6400). For gross structural refinement this number of parameters corresponds to only about 60 atoms; for detailed refinements it corresponds to only about 25 independent atoms. The need for larger capacity is particularly evident when it is noted that the magnitude of the calculation goes up as N^2, where N is the number of parameters. The parallel vector processing machines now coming into operation may have a significant impact on this problem.

2. There is a need for improved approaches, algorithms, and computer programs that will do more efficiently, or in better ways, what is done most, i.e., refinement of structural models now done by least-squares methods. Concomitant with this need are expressed needs for improved information and program exchanges, and for overcoming compatibility problems that presently limit the general utilization of desirable programs, including systems approaches, developed by certain leading groups.

3. There is a need for local computing facilities that would give faster response and would be more convenient to use. Two routes to this goal are indicated, (i) through local computer centers that are more responsive, with fewer bureaucratic problems and communication failures, and (ii) enhancement of computational capabilities through minicomputers on line with data-collection instruments in the crystallographic laboratories.

4. There is a need for attention to the computing problems of the crystallographer by the funding agencies, especially with regard to the systematic development of computing resources (as differentiated from continued use of present resources in the present manner); there is now little programmatic support in the agencies for such development.

Trends

From a combination of the invited presentations and the discussion, certain trends can be identified that will affect substantial portions of the crystallographic community:

1. Increasing computer usage by crystallographers, both because of new computational techniques and because the number of crystallographers working on larger structures is steadily increasing.

2. More work on protein structure, as it becomes possible for protein structures to be refined.

3. Increasing use of direct methods for large structures.

4. Further program development to take up significant parts of the structure-refinement load with faster (less costly) methods than the full-matrix least-squares refinement method, which still is needed in the final stages.

5. Increasing demands for accuracy in details of structures; examples are the determination of small differences associated with biological function, and the precise determination of thermal motions and of electron densities for their contributions to fundamental understanding of the solid state.

6. Increasing up-grading of minicomputers used for on-line diffraction-instrument control (i) so that they can be programmed conveniently to operate the instrument with better optimized and more flexible data-collection strategies, and (ii) so that the relatively large portion of the computer's time presently unused can be applied to data processing and interpretation including, in some cases, least-squares refinement of structural models.

7. Further emphasis by some groups on facilitating program exchanges.

8. Growing interest in computerized data banks, including banks remotely addressable.

9. Increasing use of graphic displays generated and manipulated by computer, sometimes in an interactive mode.

10. Increasing emphasis at state, city, and multiply-located-company levels on state, regional, or company-wide computer-communication networks; fueled by visions of economy in the totality of all computing, this emphasis may not always work to the advantage of the crystallographic or other scientific user.

11. Growing budgetary pressures from sponsoring agencies toward development of either (i) regional or national crystallographic computing centers addressable from remote terminals or (ii) regional or national computer-communication networks accommodating heirarchical computing attuned to crystallographic needs.

Pros & Cons of Remote Computing

Most of the discussion of remote computing dealt with networks, the ARPA network being used as the prime example. A "network" in this context is actually a computer-communication network. It consists, basically, of single or multiply redundant communication links between various traffic-controlling computers which, in turn, communicate with other larger computers at their respective sites. A computational problem entered on the net can, in principle, be assigned to the available computer most suited to that particular kind of problem.

A second type of remote service is the regional computing center which is addressable from remote terminals located at the user sites. The contemplated national computing center for theoretical chemistry was cited as an example. Such a discipline-oriented facility could be expected to optimize its program holdings and service capabilities for that particular discipline.

Although these two types of remotely usable computing services differ considerably in their desirability for various crystallographic computing purposes, some advantages and disadvantages applicable to both were brought forth in the symposium. In favor of the remotely usable service, the following five points can be summarized:

1. Protein crystallographers, and some others, need access to larger computers than are now ordinarily available in any one laboratory.

2. A quantum jump in crystallographic productivity might result from what would amount to a new dimension of computing capacity made available through a network, or a fully implemented regional center.

3. Sufficiently reliable high-speed data transmission, available either now or in the near future, may obviate most needs for physical transport of data or results, possibly including graphic displays.

4. The best computing system for carrying out the items of highest cost in crystallographic computing, primarily least-squares refinement of structural models, would be available to all in an operating and fully checked-out form.

5. Information retrieval would be possible from a central library of crystallographic data, in the forms of both hard copy and graphic display, such as a stereographic view of the structure.

Several points of disadvantage of remote-service systems were also brought out:

1. There is a cost threshold that could keep out the small user. Figures cited indicate a basic cost of about $12 000 per year, exclusive of computing costs, for a remote job-entry terminal (card reader and line printer) such as might be desired for access to a regional computing center.

For the ARPA network, the basic cost varies with the type of entry. For connection of a local campus computer (through an Interface Message Processor, or IMP), the cost is presently $21 500 per year just for the prorated share of communication and maintenance costs. [On the other hand, information provided several months after the symposium suggests that some users—in this case, of course, doing ARPA-approved work—could be accommodated from their own terminals through a 32-port Terminal IMP, or TIP, with a lower data-transfer rate (4800 baud) at much lower cost. In this case the costs, apparently, would be comparable to those for terminals remotely accessing regional computer centers via voice-grade telephone lines. For the most rudimentary type of terminal, the fixed costs could be an order of magnitude less than those quoted, but, of course, with a considerable compromise in utility.]

2. The full real cost of the computing service would be charged to the user. This could be a considerable disadvantage to many present users in institutional environments where real costs of computing are partially subsidized.

3. The growth of local resources for large computational problems could be impaired both through the diversion of dollar support to the remote-user system and through a sapping of local interest in program development.

4. The effects of less-than-optimum program operation or computing services would be magnified on a regional or a national scale. One weakness could be lack of dynamic program development. Another could be the widespread effects of any programmatic error that remained undetected for a time in a program used by a large number of remote users. Still another could be the disaster wreaked if, for political or economic reasons, the network or regional center failed or stopped rendering service.

Recommendations

Although no formal recommendations *per se*, were made by the symposium group, some that probably would have met with approval can be inferred.

First, it would be recommended that sponsoring agencies, and others in position to influence these matters, give particular attention to the needs mentioned.

Second, it would be recommended that all parties, whether sponsors, scientists, or administrators, be sensitive to the trends in crystallographic computing with a view toward enhancing and exploiting those that are desirable.

Third, the symposium group's demonstrated interest in, and apparently positive attitude toward, the potentialities of networks and regional centers for crystallographic computing lead to the inferred recommendation that the possibilities be seriously investigated by some competent group. This competent group should be appropriately related to organized crystallography and should be funded as necessary to carry out substantive investigation of all factors involved, including attitudes and preferences of the crystallographic community, of which the members of this symposium were not necessarily a representative sample.

INTRODUCTORY REMARKS

Walter C. Hamilton

Crystallographers have long been among the major users of computers in American scientific laboratories. The massive calculations involved in crystal structure determination--least-squares adjustments of hundreds of parameters derived from thousands of observations, and Fourier series calculated at hundreds of thousands of points--led the crystallographer to early exploitation of the computer. In fact, many of the advances in computer technology were stimulated by crystallographic needs.

The purpose of this symposium is to obtain a spectrum of opinion of crystallographers on such questions relating to their computing as whether their needs are presently being met, whether developments in hardware, software, and remote terminals are going to change radically the pattern of crystallographic computing, and whether centralized crystallographic computing facilities may play a role.

To provide background for the symposium, a questionnaire (see Appendix 1) was circulated to all the approximately 1800 members of the American Crystallographic Association. This questionnaire received some criticism; it is clear that one never knows exactly which questions to ask until the answers are in.

There was particular difficulty with Question 3: "What is your real computing cost per year?" Many crystallographers simply do not know, mainly because the amount of the university or departmental subsidy is never reported to them. One response reads as follows:

> "A user on our campus does not know the real cost. Why? On his output he gets a compute cost which is subsidized (or may be if his chairman wishes) up to 87%. The terminal, connect, paper, maintenance, etc. costs (inhouse) are extra and are covered by the department. The user has no knowledge of the cost of this subsidy."

This is of course a problem, but I think that judicious spade work should allow anyone to find out what his institution is spending on computing (including all these extra items) and to determine his prorata share. It behooves anyone doing scientific research to know exactly how much money is being spent in the support of his research. I might note that in an AEC laboratory, the machines have been purchased and the user is charged only operating costs--again resulting in artificially low values. The total costs, including machine amortization, can however be obtained, and AEC users should be aware of this kind of subsidy.

Now to a few results of the questionnaire:

First of all, 142 questionnaires were returned. Since I asked for only one member of each group to return the questionnaire, and since not all ACA members use computers, I would guess that these returns represent about 50% of the laboratories. This estimate is borne out by my own judgment that, among the major laboratories I know, about 50% of the large users were not represented.

What Computers Are Crystallographers Using?

The results are summarized in Table 1. It is clear that IBM still dominates, with the 360 series the major contributor. A few 370 series machines are in operation, and there are still three or four 7000 series machines in use. Most of the CDC usage is in the 6000 series, with three or four users of 7600's and a number still using 3000 series. PDP-10, UNIVAC 1108, and XDS Sigma series also have fair representation.

Table 1 Computers Used By Crystallographers

Manufacturer	Number in use by respondents to questionnaire	Percentage of total number reported
IBM	92*	52
CDC	44**	25
DEC	15	8
UNIVAC	12	7
RCA	6	3
XDS	5	3
Honeywell	2	1
Burroughs	1	0.5
Electrologica	1	0.5

*Including 68 in the 360 series.
**Including 30 in the 6000 series.

What Kind of Work Do the Respondents Do?

Structure determination	111
Small Structures	31
Medium Structures	88
Large Structures	21
Other	38

Several respondents defined small, medium, and large structures and were remarkably consistent. Small seems to indicate a molecular weight below 300; large, a molecular weight greater than 800, or 50 to 100 heavy atoms. A few of the respondents are doing very large structures (proteins), although the response from protein crystallographers was much smaller than the 50% quoted above. The other uses of computers by crystallographers (Table 2) are extensive but in toto probably represent a small percentage of the total computer time. Many respondents make use of the computer for both structural and other work. Those that do structure determination spend most of their time on refinement (Figure 1). It would seem that a great effort should be spent on making refinements more efficient, a topic discussed later in this symposium.

TABLE 2 Computer Usage by Crystallographers Other Than in Single-Crystal Structure Determination

Methods research and program development
Liquid diffraction: Amorphous radial distribution functions
Powder diffraction: Phase and mineral identification
Electron microscopy
Optical and geometrical crystallography
Electron spectra for chemical analysis (ESCA) of solids
Thermodynamics of solids
Small angle diffraction
Membrane structure analysis
Diffuse scattering
Magnetic studies
Anomalous transmission
Interferometry
On-line control of experiments and displays
Supervision and management

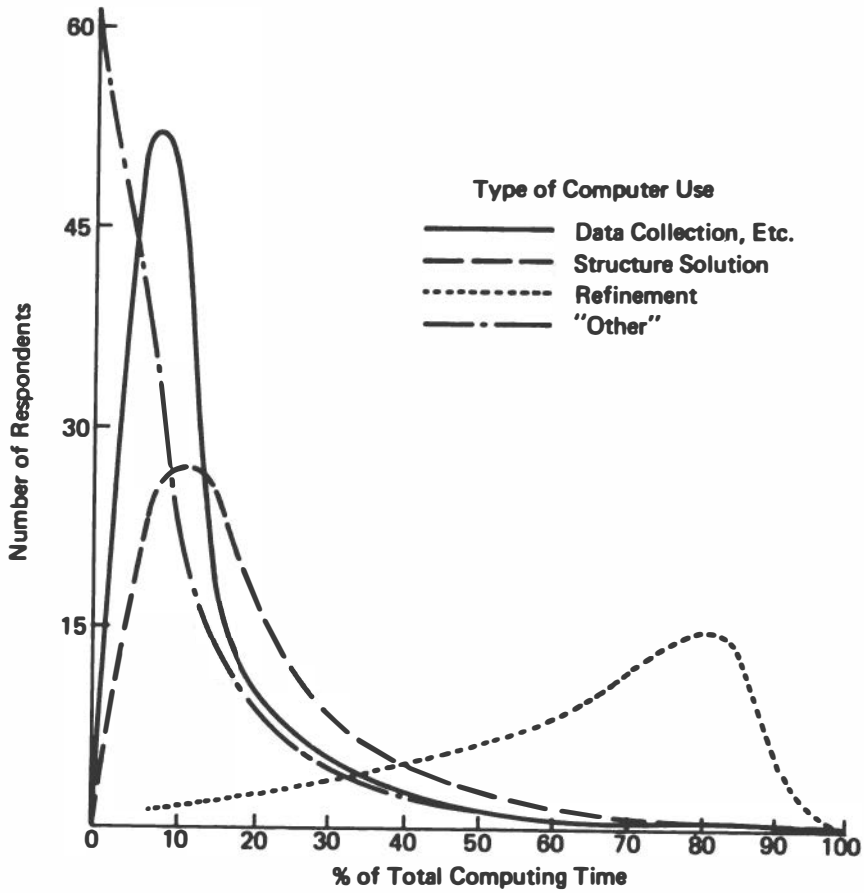

Figure 1 Estimated Time Devoted to Various Types of Computer Use by Crystallographers

What Are the Costs of Crystallographic Computing?

A summary of the responses is given in Table 3. As noted previously, the figures are difficult to come by. The maximum amount reported was $150 000 per annum. The total of $2.4 M is probably low by a factor of 4. I surmise that the response was about 50% and that hidden subsidies account for another 50% error. Thus, my estimate of the total amount of money spent in the USA on crystallographic computing is about $10 M per year. The $400 000 reported in direct government grants to individual scientists is illuminating. It means, if accurate, that a major share (more than 80%) of the computing dollar is supplied by grants to the central facilities, probably mostly from federal and state governments.

TABLE 3 Annual Computing Costs in Crystallographic Laboratories

Estimated Cost Range in 1000 $	Number of Respondents	Percentage of Respondents
0-5	41	29
5-10	20	14
10-15	13	9
15-20	10	7
20-30	13	9
30-40	11	8
40-50	9	7
>50	8	6
Don't know	16	11

Mean cost: $18 000

Total for responding laboratories: $2.4 M

Direct support through research grants: $0.4 M
Other direct computing charges: $0.1 M

Although crystallographers are becoming more knowledgeable concerning their costs for carrying out crystal structure determination, estimates (see Table 4) are still hard to come by, as witnessed by the following responses:

"Estimate is probably ±100%."

"It depends on the size and method."

"We don't have enough statistics to know."

Nevertheless, the mean value of $5400 seems reasonable to me and provides confirmation of my estimated total by the following argument.

TABLE 4 Estimated Cost per Structure

Estimated Cost	Number of Respondents
$1000-5000	78
5001-10 000	13
10 001-20 000	7
20 001-30 000	2
30 001-50 000	1
>100 000	1
Mean Estimate: $5400	

Figure 2 shows the increase in the number of crystal structure determinations over the past few years. There were probably 1500 in 1971. If half of these were in the USA, at $5400 per structure, we derive a total of $4 M. Double this to allow for hidden subsidies, add $2 M for other than structure determinations, and we arrive at the estimate of $10 M stated previously.

How Much Will Computer Usage Increase?

A summary of the responses is given in Table 5. These are of course subject to considerable variation, but the mean response does not seem surprising. There is an approximate doubling every five years as we are able to do more things and have access to more sophisticated hardware and software. This increase in computer usage does not necessarily imply a 100% increase in computing dollars, for the cost per unit calculation has been steadily dropping.

What Experiences Have People Had With Sharing of Information and Facilities?

62% have used remote terminals

37% have used computers other than in their home installations

42% have used program lists

There does seem to be substantial amount of such experience, and most active crystallographers are attuned to the feasibility of remote computing. Program lists are not so widely used, possibly because they rapidly become out of date.

How is Your Work Affected by Limitations of Available Computational Facilities?

Unfortunately, I did not include a category, "Not at all", in the questionnaire, but 9% of the respondents wrote this in. If we combine "not at all" with "slightly", we find that 55% are quite happy, 22% are very unhappy, and 21% only moderately unhappy. The nature of dissatisfaction in a few cases is illustrated by the following two comments:

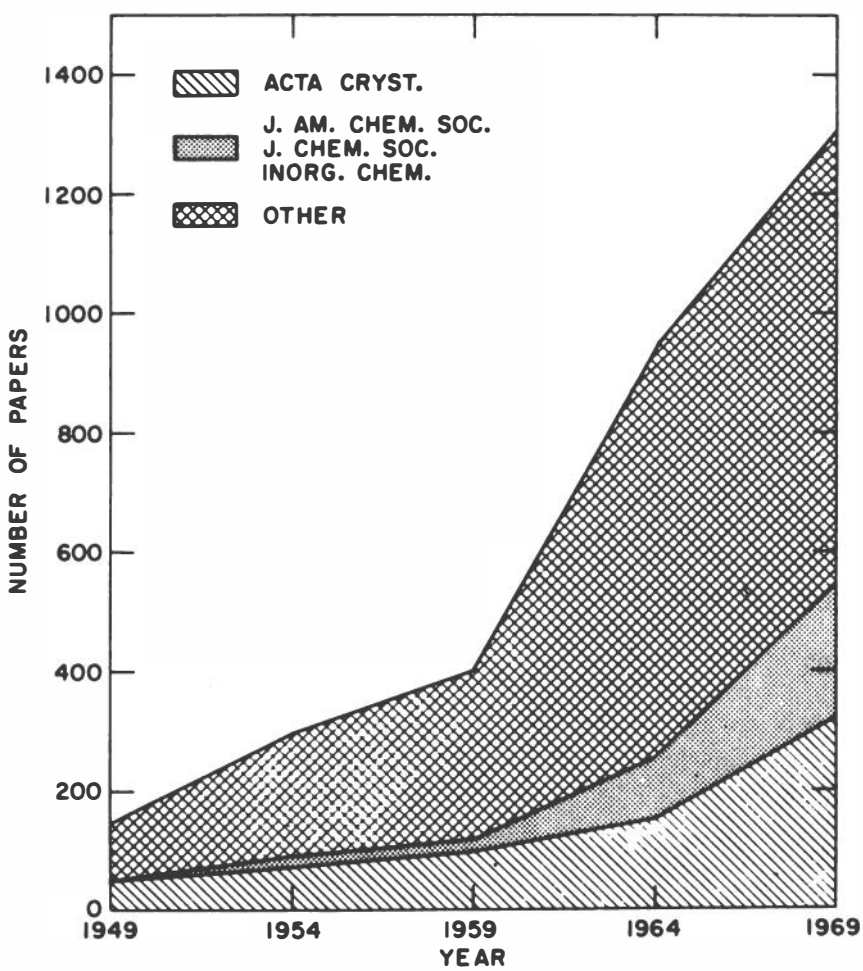

Figure 2 Growth in Annual Number of Papers Published Reporting Crystal Structure Determinations

Under Explanation of Other Computing: "Debugging newly obtained programs and adaptation to our computers; coping with frequent changes in computer hardware; poor service relationships with the university computer center, especially sheer red tape."

Under Limitations on Computing: "But these limitations are primarily imposed by security, budgetary and supervisory restraints or constraints; secondarily by the Byzantine complexity of non-scientific non-mathematical, clerical problems of getting on and off the blankety-blank computer."

This background information sets the stage for the prepared papers and discussions that follow.

TABLE 5 Estimated Increase in Computer Use

Numbers of responses in various categories	by 1977	by 1982
0%	23	22
0-25%	28	9
25-50%	22	19
50-100%	34	15
100-200%	0	6
200-300%	6	0
300-400%	1	3
400-500%	1	4
500-1000%	2	3
DECREASE	2	2
DON'T KNOW	15	34

Session I

What Are the Needs of Crystallographers?

Session Chairman: Philip Coppens

Computing Needs of the Structural Chemist

James A. Ibers

Dr. Hamilton has asked me to outline our research group's computational needs and our thoughts on how such needs may best be met. Although I am a structural chemist with considerable computational needs, I must emphasize that structural chemists are a heterogeneous group and that our group's needs and attitudes may not be typical. In outlining computational needs, I will comment also on the operation of the Vogelback Computing Center at Northwestern University. In my position as Chairman of the University Computing Committee over the past four or so years, I have gained some insight into the support problems of a large computer in a private educational institution and I hope information on this subject will be of interest to others.

Mainly for the benefit of noncrystallographers, I present in Figure 1 some features of our data-collection operation. I must emphasize that we are indeed data processors, and this fact has implications with respect to remote computing, a subject I will take up presently. Figure 1 indicates that we accumulate roughly two crystallographic data sets per month.

Picker FACS-1 with monochromator and magnetic tape output

Down time: Less than 3%

Idle time (mostly set-up): 40%

Actual data collection: about 60%

Typical data rate: 350 reflections/24 hour day

Data sets October, 1971 through March, 1972: 11

Figure 1 Data Collection

Figure 2 provides information on our university computer, on the average cost per structure, and on the breakdown of our crystallographic computations by type. Most of the money is spent on refinement, generally computations involving non-linear least-squares analysis. Again with

relevance to remote computing, I should add that the UPDATE system of CDC is very convenient for program management, and that we have ceased some years ago to maintain card images of our programs. The fact that our program library is on a permanent file is also important, and assures that all allowed users of the file have access to the same versions of the programs.

1. CDC 6400 with expanded core storage, 64 000-word core, 4 tape units, plotter

2. Program library on permanent file

3. Source library in UPDATE form

4. Average cost per structure: $2500 at $500/hr.

5. Breakdown of crystallographic uses:
   Program development: 2%
   Processing of data: 5%
   Structure solution: 5%
   Refinement: 80%
   Error analysis, drawings, etc: 8%

Figure 2 Computations

Figure 3 indicates the types of problems we have handled recently. Those familiar with our work know that without triphenylphosphine as a ligand to stabilize a transition metal in a low-valence state, we would feel very lost indeed. In this respect our computational needs may not be typical of structural chemists; even with the theoretical dodges we use, such as treating the phenyl rings as rigid groups, our problems tend to be large by ordinary standards. The size of the problem is really defined by the number of variables to be determined, and this number also defines the cost, because of the predominance of the refinement procedure over all other computations in terms of time.

Two of the eleven problems summarized in Figure 3 exceeded the capacity of our CDC 6400. By this I mean the following. In carrying out the non-linear least-squares refinement, one sets up the matrix of the normal equations, which will be of order N x N if there are N variables to be refined. Because the matrix is symmetric, only N x (N+1)/2 cells are needed. On our CDC 6400 we can obtain a maximum of about 140 000 cells for program and data. This proves to be a limitation at about N = 240.

When we have more than 240 variables we must resort to dubious mathematical tricks to carry out the refinement. These tricks cost much more than the hypothetical solution of carrying out the normal calculations on a machine of similar speed but greater storage capacity.

Average number of independent atoms: 15

Average number of phenyl groups: 5

Typical number of variables: 200

Number of variables exceeded 6400 capacity in 2 cases out of 11

Figure 3 Problem Type

Figure 4 gives an estimate of our sources of computing money for crystallographic calculations. I have divided the money into "hard" and "soft", and it is necessary to define these terms. Hard money comes from the outside. It is the kind of money that computer center directors and university controllers are very fond of. Such money comes from the various funding agencies, from private sources, etc. Soft money is a short name for internal regulation of computing. The process may work something like this: In one manner or another a Dean obtains a certain amount of soft money to be handed out to his department heads. He puts this in his bottom drawer and doles it out to the department heads who in turn may dole it out to various potential users. This process provides both regulation and power. In a typical university, control of money means power and hence the Dean maintains power. But more important, a regulatory system is built in. If Professor X goes through his soft computing money like a shot, he must ask his department head for additional funds. The department head may take funds away from others in the department or he may go back to his Dean for more money. In either case, someone at the appropriate level in the university is going to question Professor X's proclivity to spend computing money. This is as it should be, for the University Computing Committee and similar committees are not in the regulatory business. They simply have no way of assessing the value of Professor X's calculations. Over the years we, at Northwestern University, have devised various schemes for the distribution of soft money, one in particular is this: If Professor X brings in large amounts of hard money, he is rewarded with similar amounts of soft money, and the reward is on a graduated scale. This procedure has worked very well. It has enabled us to make realistic requests to the granting agencies and has assured Northwestern that money allocated to computing in such grants is indeed spent on computing, for the incentive is there to

spend it to obtain soft money. If one transfers hard computing money at the end of the year to cover debts in the supplies account, then one clearly does not gain additional computing capacity through soft money. Such book transfers were far more common before the incentive system was devised. Figure 4 indicates also roughly the amount of hard dollars that Northwestern might lose if I switched the bulk of my computations to a CDC 7600 outside. I return to this point later.

Hard money:	$15 000
Soft money:	$30 000
Loss to university computing center if CDC 7600 is used off campus:	$10 000

Figure 4 Payment for Crystallographic Computations

Figure 5 presents a rough breakdown of the sources of support for our computing center. The University puts in the bulk of the money to run the center. Sponsored research, covered by grants, provides about 25%, and about 10% comes from other outside sources, for example, other academic institutions using our computer. As a non-profit institution we cannot, of course, accept computing from profit-making organizations. The budget for the computing center is based on staff, maintenance, and amortization of the computer. The hourly cost is derived basically by dividing the budget figure by the number of hours available for computing.

University	65%
Sponsored research	25%
Outside sources	10%
Annual budget: $1M (including $200 000 amortization)	

Figure 5 Sources of Support for Northwestern University's CDC 6400

Figure 6 indicates the major types of operations that take place on our computer. These various operations represent diverse needs of diverse individuals, and it is probably impossible to provide for all of them efficiently on a given computer. Moreover, the needs are ever changing: computer-aided instruction was not even with us five years ago. It is clearly necessary to consider these operations when projecting computer needs for the next decade on a given campus, and such a projection is difficult. For example, the medical, dental, and law applications are in their infancy,

but with the big push toward providing better medical and dental care, not only will there be vastly increased uses of computers in these fields, but funding will be available to obtain giant computers.

1. Administrative data processing
2. Library processing
3. Batch processing
4. Number crunching (crystallography)
5. Interactive processing
6. Interactive programming and time-sharing
7. Information retrieval and large-scale data bases
8. Laboratory processing
9. Medical, dental, and law applications
10. Computer-aided instruction
11. Computer graphics

Figure 6 Major Needs of Northwestern University Users

Some time ago I sat on a Long-Range Computer Needs Committee which fearlessly tackled the question of the projection of computing needs at Northwestern University. The conclusions reached are shown in Figure 7. It is not appropriate here to go through our reasoning. The salient points are that we could not justify a larger machine, and that the number-crunchers and some other users will be expected to go elsewhere for their computing in the future. Our CDC 6400 is currently used about 140 hours per week, and we anticipate that the load it can handle can be extended considerably by the hardware additions indicated in Figure 7. The decision that some users will go elsewhere for their computing is a painful one to a university, because, generally speaking, those with big computer demands also bring in hard dollars. Nevertheless, there seems to be no choice.

1. Projected usage does not justify bigger machine
2. Administrative processing should remain separate
3. Additions to hardware are necessary to handle many of the tasks (expanded core storage, expanded discs, additional printer and plotter, at a total cost of about $500 000)
4. Number crunchers will have to go elsewhere with consequent loss in income to Northwestern University.

Figure 7 Conclusions of Long-Range Computer Needs Committee

What crystallographers need are faster machines with increased high-speed memory. Figure 8 indicates two of the great advantages of the CDC 7600 over the CDC 6400 for our purposes. On the CDC 7600 one can carry out in 5 to 10 minutes of central processor time what one might have been able to do on the CDC 6400 in 3 hours. Not only is the resultant calculation much cheaper but it has more chance for success, because the system is not given as long a time for a possible failure. Since furthermore the CDC 7600 has effectively infinite core in its large core memory, an entire new set of crystallographic problems become open to successful attack. In this discussion I have limited myself to the CDC 7600 as an example of an extant, giant computer. I presume similar machines are available from other manufacturers, but I am personally unfamiliar with the problems of using such machines.

1. "Infinite" core and hence computations can be made that are not possible on 6400.
2. The 7600 is 30 times faster than the 6400 with consequent savings in dollars.

Figure 8 Advantages Of Remote Computation (CDC 7600)

The following remarks, appended since the Albuquerque meeting, relate to some of our experiences with remote computing that have occurred since then. We have recently been performing remote calculations on the CDC 7600 at Lawrence Berkeley Laboratory. Transmission to and from the 7600 is done over telephone lines through the 200 User's Terminal at our Computer Center. We have not carried out large data-processing calculations, but rather have restricted ourselves during this trial period to potential-function calculations that require only modest amounts of data transmission, but do require the speed and capacity of the 7600. Figure 9 is one I showed at Albuquerque. It was my guess concerning disadvantages of remote computing. I am now able to comment on some of these points. We have had essentially no compatibility problems. A 6400 UPDATE tape compiled directly on the 7600, and the resultant program, produced answers that were identical with those obtained on the 6400. We have had some difficulties with 7600 control-card instructions, mainly because of the difficulty of finding the right source of information at LBL. We have also had some difficulty obtaining up-to-date information on system changes, although the most important ones are put on a common file which we can obtain daily. We have not tried to transmit large data files, nor do I believe that this is feasible. I think it is inevitable in the types of operations we do that there be sympathetic cooperation at the other end. We are going to have to mail data tapes and answer tapes back and forth; we are going to have to create permanent files that don't get wiped out on

Monday mornings, etc. Thus far we have not experienced any difficulty with tape storage at LBL.

1. Machine incompatibilities
2. Remote and unexpected software changes
3. Difficulty of transmitting large data files
4. Difficulty of tape and disk storage
5. Government bureaucracy

Figure 9 Problems in Remote Computation

On the basis of this limited test I am pleased with the results and am optimistic that remote computing on problems too large or too expensive to handle locally will become commonplace among crystallographers. Moreover, as I have tried to indicate, we really have no choice. It is hoped that a symposium such as the present one will not only make our needs known but will facilitate the establishment and efficient use of national or regional computing facilities.

DISCUSSION

Jeffrey: How big is Northwestern University? That is a parameter relating to usage. How many faculty and how many students?

Ibers: Northwestern has 6500 undergraduates and approximately 2000 graduates on the Evanston campus. The expectation for growth in either of these is zero. We are a private institution and are not planning to expand. We have a large graduate school on the Chicago campus, in Medicine, Dentistry and Law, and these programs will expand under government pressure. But on the Evanston campus, we can estimate computing need on the basis of a fixed population.

Jeffrey: How many faculty?

Ibers: There are about 400 faculty members in the College of Arts and Sciences, representing some 75 per cent of the total faculty.

Caughlan: What fraction of your usage is administrative and what fraction is educational, including instruction and research of graduate students?

Ibers: The administrative computing is done on a separate computer, and so does not affect us. The keeping track of where books are from the library is also done on that computer. The 6400 is used only for research and instructional purposes, and the usage is approximately 35 per cent undergraduate, 35 per cent graduate, and 30 per cent faculty. The distinction in the last two areas clearly is not easy because some professors assign computer numbers to individual graduate students, while others use a blanket number for themselves and all their graduate students. But the instructional uses are about one third.

Coppens: What is done with faculty members who have no funds for research, do they get time?

Ibers: The faculty members who have no funds get soft money up to a point, and as far as I know this point, with rare exceptions, provides them with enough to do what they need to do. A man who has no money from the outside and wants to become a major user of the computer is clearly destined for some talks with his dean.

Dewar: I would like to pursue the point you made that the 6400 in its present state will meet needs other than in the one category of number crunching. You mentioned ten years as a possibility. I wonder whether that is likely, because I think the potential for explosion in usage is even greater in some of those other areas, computer-assisted instruction and data-base retrieval, for example. You may be talking about demands for a thousandfold increase in on-line storage, which is conceivable with projected hardware. What are some of the parameters that go into that decision?

Ibers: It is difficult to feel confident about projections of computer usage. Nevertheless the Long-Range Computer Needs Committee at Northwestern did talk extensively with diverse groups of people, many of whom had grandiose schemes for computing in the next decade. Obviously one of the major parameters that went into our thinking was hard money, or rather the lack of it, for extensive computer changes. Perhaps I left the wrong impression in my talk. We did not conclude that only the number crunchers would go off campus to compute; they were simply the most obvious group at the present time. But should we get heavily involved in computer-aided instruction, then money will not be spent at the computer center but will possibly buy remote terminals for a tie-in with the Plato system at Illinois. In any event, I too have been around the computer game for many years, and know your worries when you imply that estimates on computer usage always fall short.

Nevertheless we think our estimates are based on reasonably hard facts.

Dewar: There are other areas where going off campus probably looks very attractive, for example, large scale manipulations with the census files.

Ibers: Another example at this time is that we run our Chemical Abstracts searches at Argonne National Laboratory, not because we want to but, as I understand it, because their format is difficult to change away from IBM. So we pay for the Chemical Abstracts searching at Argonne.

Williams: I have a question related to funding but in an indirect way. As you know, the government is about to buy 10-12 large machines and I am wondering about the problems of conversion. I've heard, being an IBM 360 user, that the conversion between the CDC 6000 and 7000 series involves quite a change. I wonder if you could comment on the conversion problem, particularly with respect to interconversion between IBM and CDC 7000 series machines.

Ibers: The incompatibilities between us and the LBL at Berkeley are no more serious than incompatibilities between one 6000 center and another, and these incompatibilities will lessen as the 7600 at Berkeley institutes more of the Scope system.

Jeffrey: At the University of Pittsburgh we have disposed of the soft-money problem by putting the computer on overhead. It raised the overhead by about three per cent, but saved real money by reducing the bookkeeping and bureaucracy of the soft-money operation.

Coppens: Does that mean free access for everybody?

Jeffrey: Yes. The resources are allocated on a departmental basis, based on what is considered to be the department's justifiable use. The priorities and departmental use are programmed into the computer. Responsibility for proper use of the computing resource is then placed at the departmental level.

Sayre: You have indicated that crystallographers may often want to turn to special off-campus machines for their large number-crunching computations. You have pointed out two associated problems, that of data transmission and that of administrative complexity. What about cost? Do you have a cost figure for the 7600 you are trying to get on to, that can be compared with the $500 per hour you pay on your own campus?

Ibers: Yes, and this is part of the problem. The rate of the 7600 at LBL for contract users, people with AEC money, is $600 per hour for a machine that is 30 times faster than one for which we pay $500 per hour. The reasons for this of course have already been pointed out. The AEC

considers only operating costs in arriving at an hourly charge. They do not consider amortization or the initial expense of buying the machine. So it is a very attractive computer.

Crystallographic Computing in a Small Institution Without Large Inhouse Computers

Helen M. Berman

Introduction

In an institution the size of ours, with about four hundred personnel of whom perhaps twenty have any use for computer facilities at all and only four or five need extensive computing capabilities, it is clearly not practical or desirable either to own or to rent and maintain a large computer installation. Among the options that were open to us were to (1) rent or buy a small computer or (2) rent or buy a remote terminal that can access the large computers at computer centers. We chose to rent two types of remote terminals, a key punch, and a card sorter. I will describe our facilities and try to evaluate them in terms of the needs of a small crystallography laboratory.

A Remote Batch/Intelligent Terminal: Univac DCT 2000

The DCT 2000 operates as a remote arm of a central computer and also has some processing capability. Table 1 lists the specifications of our present system. It has a line printer, card reader, and card punch which can function both on and off line.

We have a Bell Telephone Data Set which modulates and demodulates digital data to go over the analog common carrier. It operates over a public dial-switched network and allows a maximum transmission rate of 2000 bits per second. The connections between the data set, the terminal, and the telephone lines are electrical. The data are transmitted in eighty-character blocks. When errors are detected in transmission there is automatic retransmission. We also have a modem that transmits the data over a private four-wire line to one particular computer. Its maximum transmission rate is 4800 bits per second. The principal reason for the increased rate is that one pair of wires allows continuous transmission of data while the other pair is used for checking. With public lines we can use only two wires for both processes. The private line is also considerably less noisy and there are fewer automatic retransmissions.

Requirements for the Central Computer

The computer must have the hardware to accept the fast transmission rate and the software to decode ASC-II and the specially blocked character strings. In practice this limits us to UNIVAC 1108 facilities and some CDC 6600 facilities. IBM uses the EBCDIC code.

TABLE 1

PRINTER CHARACTERISTICS

Printing Speed: Maximum rate of 250 lines of alphanumeric characters per minute; 60 lines per minute with voice-grade telephone line
Printing Positions: 128
Paper Speed: 25 lines per second
Special Features: Transmit/Receive Monitor, Offline Listing Form Control

READER/PUNCH CHARACTERISTICS

Cards: Standard 80-column cards
Reading Speed: Maximum rate of 210 cards per minute, 75 cards per minute on voice-grade telephone
Reading Method: Photoelectric read station
Punching Speed: Maximum rate of 75 cards per minute for 80-column punching
Punching Method: Two columns at a time
Input Hopper Capacity: 1200 cards
Output Stacker Capacity: 850 cards

CONTROL UNIT CHARACTERISTICS

Transmission Method: Block by block
Transmission Mode: Half duplex; 2 or 4 wire (nonsimultaneous; two-way transmission)
Transmission Facilities: Voice-grade telephone toll exchange or private line
Transmission Rate: 4800 bits per second (private line); 2000 bits per second (switched telephone network)
Transmission Code: ASC-II, XS-3 (DLT compatible)
Buffer Storage: 256-Character capacity - Two 128-character core memory buffers
Translation Capabilities: Card Code/Transmission Code, Hollerith/ASC-II, Hollerith/XS-3 (DLT compatible)
Special Features - Error detection and retransmission, telephone alert, select character capability, short block capability, peripheral Input/Output channel, unattended operation

Description of the Actual Operation

At present we have contacts with two commercial data processing centers in the Philadelphia area and one university center in New York City. We transmit our data over a voice-grade public telephone line via the card reader. One company gave us a private line to encourage us to dedicate all our computing to them.

Our intensity data are collected on magnetic tape and sent via mail or messenger service to the computer center. As much as possible, we try to store all large data sets and programs on disk files or magnetic tapes so that we do not have large input decks for everyday work. We use the X-ray 70 system for most of our standard crystallographic calculations and use stand-alone programs for special applications. In practice, we do about ten calculations a day via the DCT and divide our work about equally between the two commercial centers. If we wish, we can submit several jobs at once. Almost all output comes over the line printer and card punch. Occasionally, for very large jobs, we have a messenger deliver the output. All our CALCOMP plots are delivered. The computers have software such that we can submit a job, terminate, and retrieve the output later in the day. The turn around time at the commercial center is usually about one hour. If we wish, we can leave the terminal on line all day and allow the output to come back when it is ready. While it is theoretically possible to leave the terminal completely unattended we almost never do because of printer jams.

We run our full-matrix least squares at the university center because of the very low rates for central processor time. However, the transmission rate is extraordinarily slow due to the bad phone connections between Philadelphia and New York City. Furthermore, the long-distance telephone rates add to our costs.

In Table 2 are listed our costs. It is clear from these numbers that the charges for central processor time are very variable and at the commercial centers very high. However, the other aspects of the service are good and the telephone connections are satisfactory so that at present we are not inclined to use less expensive but more inconvenient services.

TABLE 2

DCT 2000 terminal (rental and maintenance)	$917/month
Telephone service	$200/month
UNIVAC 1108 CPU time, commercial rate	$800/hour
CDC 6600 CPU time, AEC rate	$200/hour
Magnetic tape storage	$10/month

The Simple Input/Output Terminal

We have recently experimented with the use of an ordinary input/output terminal with nothing more than a keyboard and acoustic coupler. Our transmission speed is 300 bits per second in ASC-II code via an ordinary telephone line. The purchase cost of such a terminal is about $3000, or $150 per month for rental. The asynchronous eleven-bit character transmission is compatible with a large number of computers. At present we access the CDC 6600 at Brookhaven National Laboratory with this terminal. To test this computing arrangement we refined a structure by full-matrix least squares. The data were sent via mail since we do not have a card or tape reader. The least-squares program was modified so as to trim the output considerably. We use the FOCUS system which is a multiple access file handling system. All the appropriate input parameters were stored on files which were edited at various stages of the refinement. The editing features were also used to output selectively the results of the refinement. The turn around time varied from 10 minutes to about three hours. With the present system at Brookhaven we can only submit one job at a time from a particular terminal; we cannot terminate and retrieve the output at some later time. When a job is complete we can make the output files into permanent files and print the full contents later. An Fo, Fc listing for about 1200 reflections takes about half an hour. In practice we can run about two big calculations a day with one terminal. The unreliability of the telephone lines causes us to be disconnected about twice in an eight hour period.

Appraisal of Remote Terminals

In general we are satisfied with our computing system, with some reservations which I shall outline. Certainly in our situation it would be impossible to maintain a large computer, and the nearest large computer center is fifteen miles away. Our choice then is between utilizing a mini-computer to its utmost or using a remote terminal system. At the time the decision was made to change from the IBM 1620 to a more modern system, the only small computer available to the laboratory that was within the budget was the IBM 1130. Programming the 1130 or a mini-computer to do crystallographic calculations was and still is a formidable task, and the laboratory did not have the personnel or inclination to approach the task. It seemed more prudent to use computers that were already programmed for crystallographic calculations, which at the time meant having access to an IBM, Univac, or CDC computer.

I outline below the advantages and disadvantages of our system.

Advantages

1. The turn around time is quite short and one bypasses the usual delay incurred in waiting for an operator to remove output from the line printer.

2. We have tremendous versatility and can access a variety of large computers. We are not committed to any one computer, so that if service deteriorates at a particular center we can easily switch to another.

3. On the I/O terminal using the FOCUS system at BNL, one can take advantage of the editing features to easily change data cards and shorten the output. For example, if all you want to see is how one atom has refined, you can search for that atom alone and by-pass the rest of the output. Debugging of programs also is simpler. One learns very quickly how not to be "card" bound.

Disadvantages

1. The public telephone lines are not reliable and as a result the transmission rates are slowed. The telephone rates are high for long distance calls. Leased lines are more reliable but they force us to dedicate all computing to one center.

2. The commercial computer centers are not stable and we must be on our guard against business failures.

3. The cost for operating time is high and, since we must pay for our computing in real cash, this considerably limits any experimental computing we might contemplate.

4. We do not have a cathode-ray-tube or magnetic-tape reader.

Our Concept of an Ideal Computer System

For us the ideal system would provide a remote terminal with display capability and enough memory so that we could do small calculations on site. The ideal computer center with which we would communicate should have the hardware and software to handle a variety of remote terminals. It should be fast and have a large core. The center itself should have some programming experts to aid us in some of our computing problems, and a well documented library of essential crystallographic programs that can be easily accessed. Finally, the rates should be low. If all these criteria were met by the center we should probably be able to invest in a leased line and thereby by-pass our problems with the public telephone lines.

DISCUSSION

Baur: It would be interesting to get from you the cost per structure, because you really know the amount.

Berman: When I saw the graph I was surprised, because our cost comes out much higher. Our computing budget is about $35 000 a year in real money. It's hard to say how many structures we've solved per year because we got our diffractometer only last year. Our total cost of computing is high and I think it could be lower if we didn't have to deal with a commercial company.

Baur: There are several universities in the Philadelphia area. Can't you deal with them through your remote terminal?

Berman: No. The University of Pennsylvania has a 360 which people on site seem to have trouble using. We can't communicate with the 360 at all. Really our way is the only way we can do it.

Coppens: I think we should be careful about talking too much about the cost per structure because your structures may be larger.

Berman: Yes. One structure may take much longer to solve and therefore we may have fewer structures. Also we are dealing with very rigid systems. The commercial centers have two priorities. We do overnight computing but it really doesn't save us much.

Dewar: In my experience these terrible troubles about telephone lines are confined to the east coast in general, or non-Bell Telephone areas. Certainly we have had no problems whatsoever in the mid-west, which I think is worth mentioning for those who don't have the experience. Secondly, the DCT 2000 is rather a curious choice given your aims. What were the parameters of that decision? There are at least a dozen devices on the market that are fully programable and can communicate with any computer in sight.

Berman: Not at the time we made the decision, which was in 1967. I agree that there are many good terminals now - better than the DCT.

Dewar: It's much easier to make your end flexible than it is to march around the country trying to make an arbitrary number of computers flexible.

Berman: Yes, but to get it taken out and install something now would be quite difficult. I agree with you.

Hamilton: Your computing costs are about $10 000 per man year and my averages seem to be about $5000.

Berman: Yes, as I said, I think our computing costs are too high and I would like to see something done to lower them, mainly lower CPU rate.

Ibers: Personally I want to run my programs at someone else's institution. If I must run their programs, then easy access to modification of those programs is essential. Let me hasten to add that even our own programs must be modified on occasions for a particular problem.

Coppens: I understand you can do modifications of programs.

Berman: Yes, It's done all the time but it turns out that for routine structure analysis we tend to use X-ray 70 and are happy with it. Yes, we can do any kind of computing we want.

Young: There may be another side to the point that Ibers raises, which is, in most of the computing we do we would be happy to use his program wherever it is.

Okaya: Since you have an automatic diffractometer, that means you have a small computer in your laboratory.

Berman: Yes.

Okaya: Is it possible for you to use this in a kind of time-shared mode?

Berman: No.

Okaya: It must be much cheaper in the long run than spending so much money. Could you perhaps do all the refinement on that small computer?

Berman: We have so few people in the laboratory that if one person devoted all his time doing the programming for that, there would be so much less that we could do of other types of work.

Okaya: Perhaps Syntex could make that possible for you.

Berman: Yes, that's what I'm hoping.

Caughlan: About ten years ago, we used a remote terminal to connect with the UCLA 7094. We had a lot of trouble with telephone lines there too, and this was card input-output. It was very difficult, and this indicates something about distance transmission.

Zalkin: I have a question regarding cost. The NYU computer costs $200 per hour and the other commercial ones are much higher. I don't quite

understand why you just don't use that one all the time.

Berman: Because it's very frustrating trying to get through the telephone lines.

Zalkin: To New York?

Berman: To New York. It would be great to do all our computing at Brookhaven or at NYU.

Zalkin: Are most of the commercial outfits close by?

Berman: They're in the Philadelphia area.

Bernstein: I'd like to comment on the telephone lines. We've also had quite a few problems. We investigated the situation. I spoke to the people who design some of these lines. It seems the ones you are using are voice-transmission lines and were not designed for data transmission. Also, there has been a strike against New York Telephone, and we are assured that things are going to improve. Even though they are voice-grade lines they should transmit data. The company has been working on ways of improving it.

Ibers: I have heard that the FCC is considering the problem of broadband microwave transmission. If such a means of transmission is allowed, then it should materially improve the possibilities of handling large data sets. Does anyone have any knowledge of the situation with respect to microwave transmission?

Suddath: Our 370 is tied with Harvard by microwave and my understanding is that it's working out quite well, very high transmission rates.

Berman: Just using the leased telephone line makes all the difference in the world.

Murphy: We're on the ARPA net. It's a fifty-kilobit line and extremely reliable. If more networks like this could be developed so that smaller institutions might get on, it might really relieve the problem.

Wolten: A little over 12 years ago I worked for North American Aviation. They had computers at several of the various installations spread out over the Los Angeles area. They were all linked by microwave. Any program could go to any computer that was available at the time and the answer would come back the same way.

Dewar: In case people do have trouble with transmission, there exist modems that will solve these problems completely if you can afford

to install two of them, one at each end, error-correcting modems. There are several of these on the market and I'm sure they would eliminate most of your problems.

Berman: If we had done that, it would have dedicated us to one computer center, and we haven't yet found the computer center we are willing to make that dedication to.

Meyer: There are several commercial firms that plan within the next few years to blanket the country with data transmission networks. DATRAN for one.

Computing Needs of Protein Crystallographers

Keith D. Watenpaugh

The growth of crystallography has been closely linked to the growth of computer science and technology in general. As computers became faster and more sophisticated, the rate of growth of crystal-structure determinations and their degree of refinement and accuracy increased. In no field of crystallography is this more evident than in protein crystallography. Protein molecules are at least an order of magnitude larger than those studied in normal crystallography, and along with this large size comes very real problems and experimental limitations associated with the collection and treatment of data. In fact, many analogies may be drawn between the state of the art of protein crystallography now and that of ordinary crystallography 15 to 20 years ago when smaller and slower computers were just coming into use.

Computers are important not only in the processing of crystallographic data but also in the collecting of the data. The mass of data necessary to solve protein structures even in moderate detail, as well as protein's almost universal instability (especially under X-ray bombardment), requires automated high-speed data collecting and processing. Presently, either computer-controlled diffractometers or computer-controlled film-scanning densitometers are used for this purpose. This aspect of crystallography is discussed later in this symposium, and I mention it here only because it is an indispensable part of protein crystallography. Also, new, extremely high-speed data-acquisition systems are now being developed (Xuong and Vernon, 1972).

Computer application to solving protein structures through phasing by multiple isomorphous-replacement techniques (Blow and Crick, 1959) has become quite routine. Also, improving heavy-atom parameters by alternating cycles of least-squares refinement with cycles of phasing by multiple isomorphous replacement has become standard (Dickerson et al., 1968). However, this method of determining phases cannot usually be extended beyond 2.0 Å resolution (resolution is usually defined as the minimum interplanar spacing to which data were collected). A Fourier synthesis (electron-density map) calculated using these phases and fitted by some model giving approximate atomic positions usually is the final step in the structure determination. This is so for a number of reasons, including the difficulty of obtaining good higher-resolution data and the limitations of present-day computers.

Further computer applications at this point may include fitting a model to the electron-density map while maintaining some constraints

on the model (Diamond, 1971). This allows approximate atomic positions to be calculated, but their accuracy is quite uncertain. Important use of computers has also been made in studying the protein conformations with computer-controlled display systems (Barry and North, 1972). Tollin and Rossman (1966) have described various rotation-function programs. Programs of this type may be used to fit known protein models to the crystallographic data of similar unknown proteins in order to solve related protein structures without using isomorphous-replacement techniques. However, the most demanding use of computers in the near future is going to be in the refinement of protein structures to produce much more accurate models.

Since the phased data in even a "high resolution" protein structure do not approach the precision required to resolve individual atomic positions, the current protein models are poor by regular crystallographic standards. The need to improve these models is indisputable. As more and more structures are being determined to 3.0 Å, 2.5 Å or 2.0 Å resolution, it is becoming painfully obvious that the models simply are not accurate enough to explain the unusual and unique physical and chemical properties of many proteins. Practically nothing can be said about bond lengths or angles in proteins, and even atomic positions have uncertainties of the order of a half an ångström or more.

Following are just a few of the questions that may be answered if more accurate protein structures are obtained, with computer refinement of protein structures playing a primary role:

1. Vallee and Williams (1968) have proposed an entatic state or region of abnormal conditions in the proteins as giving rise to internal activation by geometric and/or electronic strain. Stretching a bond by 0.2 Å or twisting it through 20° can produce very large changes in energies, yet be entirely missed by present protein X-ray crystallographic analysis.

2. High orientation dependence has also been proposed as contributing to the unusual catalytic properties of enzymatic proteins. Strom and Koshland (1970) have suggested that large rate increases may be realized by proper orientation of reacting molecules, and that enzymatic reativity may be due to this "orbital steering" ability. They propose that changes in angles of as little as 10° may produce rate changes of 10^4, again outside the range of present protein crystallographic accuracy.

3. Nonplanarity of peptide groups as well as the close proximity of atoms appear to be implicated in the activity of lysozyme (Barry and North, 1972). An accurate structure is required to assess the degree of nonplanarity of its peptides.

4. Chromatium high-potential protein (HIPIP) and bacterial ferre-

doxins have similar iron-sulfur clusters in 2.2 Å resolution electron-density maps, yet their physical properties are very different (Carter et al., 1972; Adman et al., 1972). Ferredoxin has an oxidation-reduction potential of approximately -400 mV whereas that of HIPIP is +350 mV. An accurate description of the cluster and its environment is needed to explain the difference.

Refinement of Protein Structures

Procedures for refining protein structures fall into two classes. In the first are those that attempt to produce the best fit of the model to the electron-density map generated by determining phases through multiple isomorphous-replacement methods. The second includes procedures that improve the phases and/or extend the phases to higher-resolution structure factors to produce a more accurate model than can be derived from the experimentally determined phases.

R. Diamond (1971) has written a sophisticated computer program that optimizes the fitting of a model to an electron-density map while maintaining some constraints on the model. Bond lengths are kept constant while overlapping densities of neighboring atoms are accounted for. Some interbond angles may either be constrained or allowed to vary. This procedure appears to lead to an improvement of the model with respect to experimental data to 3.0 Å or 2.0 Å resolution if the electron-density map is reasonably good. However, it must be noted that refinement of a model by electron-density maps has several disadvantages when errors exist in the data or when atoms are not resolved. Computer requirements for this type of refinement are not particularly large, but the model produced would be considered only a reasonable starting model for refinement according to regular crystallographic standards.

A second procedure, which involves refining phases and extending them to higher resolution, is the so-called "direct methods." Use of direct methods is discussed by D. Sayre in this symposium. Application of direct methods to protein crystallography has not proved very successful yet, but some promising results have been obtained. However, computing times can be enormous.

In the course of the refinement of small structures, it was found that ΔF syntheses (difference maps) provided advantages over Fourier syntheses in the refinement of structures when atomic positions were not resolved, as is the case when working with two-dimensional data or when there are series-termination errors due to lack of higher-order data. Shifts in atomic positions are proportional to the slope at the assumed atomic positions and can be determined more easily and reliably. Apparent thermal motion is also more easily determined. Moreover, gross errors in the model, such as misplaced or missing atoms, can be detected. A

reasonable analogy may be drawn between refinement of small structures with two-dimensional data and refinement of proteins with three-dimensional data when atomic positions are not quite resolved. It is questionable, however, whether unrestricted use of ΔF refinement is justifiable without data near atomic resolution. Computer times required for ΔF refinements in general are not large.

With the advent of large high-speed computers, the most powerful method of refining small structures was by means of full-matrix least-squares. Even on small structures, least-squares can tax the largest and fastest computers. True full-matrix least-squares refinement on a protein structure cannot be reasonably accomplished on today's computers. Reducing the magnitude of the task by neglecting off-diagonal terms, as is sometimes done on small structures, proves disastrous with proteins, since at low resolution the correlation between neighboring atoms is high and cannot be ignored. Further problems may arise because the number of structure factors does not greatly exceed the number of parameters to be refined and because an appreciable fraction of the crystal may be composed of solvent.

The numerous difficulties and limitations associated with the refinement methods have prevented people from refining protein structures in spite of the great need to do so. However, it is important to know whether it is possible to refine proteins and to what extent the model may be improved.

Refinement of a Protein Structure

A brief summary of the refinement of rubredoxin provides a convenient way of describing the magnitude, merits, and limitations of various refinement procedures (Watenpaugh *et al.*, 1971).

An accurate description of the iron-sulfur complex as well as the chain conformation is essential in understanding the unusual physical properties of this protein. Also, it is a good structure on which to test protein refinement because of its relatively small size and because significantly observable data exist to a resolution of 1.5 Å (approximately atomic resolution).

Atomic positions with which to begin the refinement were picked off a 2.0 Å electron-density map phased by multiple isomorphous replacement. Structure factors based on these parameters had an R-index of 0.37. The R-index, defined by

$$R = \frac{\sum ||F_o| - |F_c||}{\sum |F_o|}$$

is used to measure the discrepancy between the observed and calculated structure factors. Models of small structures with R-index in the vicinity of 0.4 can usually be refined.

It is convenient to quote computer time and costs of the CDC 6400 computer at the University of Washington to compare the magnitudes of the various steps in the refinement of rubredoxin. The present charges on this computer are approximately $5.00 per minute for the central processing and a variable amount for input-output and peripheral processing. The X-ray 70 crystallographic computing system supplied by Dr. Stewart was used for all major computing (Stewart, 1967). The structure factor calculation (F_c calculation) over the approximately 500 atoms in the asymmetric unit or 1500 atoms in the cell on more than 5000 observed reflections requires approximately 65 minutes of central processor time and costs $350. A ΔF synthesis following the F_c calculation, with 2 grid points per ångström, takes 40 minutes and costs $250.

Initial refinement was with ΔF syntheses for several reasons. The maps provide a constant check on whether sensible corrections are being applied, allow modifications in assignment of peptide residues since the sequence was not known, and keep computing time and costs within reasonable limits.

Calculations of the shifts in atomic positions from the ΔF syntheses were done by hand. Approximately 30 man hours were required at this task of calculating shifts on about 500 atoms per cycle. The ΔF refinements were sufficiently well behaved to improve the structure significantly and allow identification of additional residues the assignments of which had been questionable at the outset. Four cycles of ΔF refinement decreased the R-index from an initial value of 0.37 to 0.22.

Since the ΔF refinements were fairly well-behaved it seemed feasible to attempt least-squares refinement. It is impossible to take full advantage of the superior characteristics of full-matrix least-squares refinement because of the magnitude of the problem. Even this small protein, with approximately 600 atoms including water molecules in more or less discrete locations, involves 2400 parameters to vary (x,y,z and isotropic thermal parameter). Even to store the unique part of the symmetric matrix would require 3 million words of storage. The computer time to build such a matrix in the course of a refinement is not feasible on currently available computers. The CDC 6400 computer at the University of Washington with a core size of 64000 words is capable of refining 240 parameters. Therefore the matrix was partitioned into blocks of about 240 parameters associated with neighboring atoms and requiring 10 passes through the computer to complete one full cycle of refinement. One cycle of refinement requires approximately 17 hours of central processor time and costs about $5400 to $6000 as compared with about $600 and a week of hard work for ΔF refinement. Bond lengths and angles were calculated after each least-

squares cycle, as were ΔF syntheses at various stages to check the behavior of the refinement. The least-squares refinement behaved surprisingly well, with no general tendency for atom positions to oscillate or diverge, in spite of the lack of complete atomic resolution and the not highly overdetermined constraints on the parameters. After four cycles of least-squares refinement, the R-index is 0.126, but some regions of the molecule are still poorly defined because of either high thermal motion or disorder.

It is now evident from the refinement that there is very significant distortion in the tetrahedral configuration of the iron-sulfur complex. Three iron-sulfur bond lengths may be not significantly different from each other (2.34 Å, 2.32 Å, 2.24 Å with standard deviations of 0.03 Å) and agree well with those observed in small crystal structures, while the fourth is short (2.05 Å) suggesting an entatic nature for this protein. The more accurate iron-sulfur cluster as well as the more accurate description of the surrounding polypeptide should allow for more quantitative theoretical calculations to be performed to explain both electron-transport mechanisms and the energetics of protein folding.

Perhaps the most important outcome of this successful refinement of rubredoxin has been to stimulate refinement of other proteins in which better accuracy is required to explain their mechanisms. ΔF refinements are currently under way on both bacterial ferredoxin and high-potential iron protein at 2.0 Å resolution in hopes of explaining their very different physical properties. Subtilisin and pancreatic trypsin inhibitor are being refined to better understand behavior of proteases. Refinement is beginning on the triclinic form of egg-white lysozyme, which holds promise of being capable of refinement beyond any other protein currently under study.

Suddenly, refinement of protein structures is no longer in the future but in the present. The limiting factor is not whether proteins can be refined but the computational aspects of refinement. New computer programs designed for the refinement of protein structures, not small crystallographic structures, must be forthcoming. For example, ΔF refinement techniques, which disappeared from use on small structures with the advent of high-speed computers, must be reexamined keeping in mind current computer technology and speed. New methods of least-squares refinement are required that take into account the overlap of electron densities of neighboring atoms and allow more nearly diagonalized matrices, so as to increase speed and efficiency of refinement. However, in final analysis, the dynamic growth of protein crystallography is dependent on the increasing availability of larger and higher-speed computers.

Acknowledgments

I am indebted to Dr. L. H. Jensen for many helpful discussions and to the USPHS for support under Grant GM-13366 from the National Institutes of Health.

References

Adman, E., L. C. Sieker and L. H. Jensen. 1972. The structure of a bacterial ferredoxin. Amer. Cryst. Assn. Abstr. p. 66. Albuquerque.

Barry, C. D. and A. C. T. North. 1972. The use of a computer-controlled display system in the study of molecular conformations. Cold Spring Harbor Symp. on Quant. Biol. 36: 577.

Blow, D. M. and F. H. C. Crick. 1959. The treatment of errors in the isomorphous replacement method. Acta Cryst. 12: 794.

Carter, C. W., Jr., S. T. Freer, N. H. Xuong, R. A. Alden and J. Kraut. 1971. Structure of the iron-sulfur cluster in Chromatium iron protein at 2.25 Å resolution. Cold Spring Harbor Symp. on Quant. Biol. 36: 381.

Diamond, R. 1971. Real-space refinement procedure for proteins. Acta Cryst. A27: 436.

Dickerson, R. E., J. E. Weinzierl and R. A. Palmer. 1968. A least-squares refinement method for isomorphous replacement. Acta Cryst. B24: 997.

Stewart, J. M. 1967. X-ray 67: program system for x-ray crystallography. TR-67-58 (NSG-398), Computer Science Center, University of Maryland.

Storm, D. R., and D. E. Koshland, Jr. 1970. A source for the special catalytic power of enzymes: orbital steering. Proc. Nat. Acad. Sci. 66: 445

Tollin, P. and M. G. Rossman. 1966. A description of various rotation function programs. Acta Cryst. 21: 872.

Vallee, B. L. and R. J. P. Williams. 1968. Metalloenzymes: the entatic nature of their active sites. Proc. Nat. Acad. Sci. 59: 498.

Watenpaugh, K. D., L. C. Sieker, J. R. Herriott and L. H. Jensen. 1972. The structure of a non-heme iron protein: rubredoxin at 1.5 Å resolution. Cold Spring Harbor Symp. on Quant. Biol. 36: 359.

Xuong, N. H. and W. Vernon. 1972. A rapid data acquisition system for protein crystallography. Amer. Cryst. Assn. Abstr., p. 59, Albuquerque.

DISCUSSION

Freer: I wish to comment on the refinement of HIPIP, the high potential iron protein from Chromatium D. We were so encouraged by the progress of the Seattle group that last December we wrote a numerical differential-synthesis program which we hoped would mate protein refinement. This procedure has worked amazingly well. A complete refinement cycle consists of an F_c, a ΔF map and then automatic calculation of slopes and parameter shifts. For HIPIP, where we're talking about 800 atoms, the F_c runs 7 minutes, the Fourier 4 minutes, and the differential synthesis about 30 seconds. For a total cost of approximately $600 (for about 3 hours of CDC 3600 time), we reduced the R factor for HIPIP from 34 to 16%.

Coppens: That's even lower than doing it by hand, but much faster.

Freer: Yes. All I want to emphasize is that since this program has come into being such refinement is becoming practical.

Watenpaugh: In starting out, for example, we've used the X-ray system designed for people who are solving lots of different structures and lots of different space groups, but when we come to protein structures we're going to spend a significant amount of time on refinement so that it will be very important that we optimize the system for a particular space group and a particular protein.

Stewart: People tend to refer to this X-ray system as being mine. The authorship extends over a great many people. In fact, Steve Freer who just spoke is himself one of the original authors. And it was not our intent that the least-squares nor this Fourier program be used for protein-structure analysis. It was written with the idea of space-group flexibility and convenience, and therefore you're paying for this in overhead in a real way. If it's used for these large structures it does cost more, and I think Steve's remarks are especially important. In the old days we really pushed for efficiency rather than convenience. There could be many short cuts made. I'm sorry to say also that I completely obliterated about two years ago our differential synthesis, destroyed all vestiges of its existence, and threw it away, believing it had been an exercise in futility. So that has saddened me a little bit.

Dewar: Bearing in mind the whole purpose of this symposium, particularly in terms of looking forward to what could be done with large computing facilities, one conclusion I seem to hear is that one really isn't that far away. The costs you quoted are high, but entirely reasonable for the size of structure that's being tackled. Does it seem fair to conclude that even in connection with protein structures, we're talking not about some miraculous new hardware two orders

of magnitude faster, but about very large existing computers, perhaps 7600's, at the top of the scale? This would be a significant conclusion from the point of view of building a gigantic computer for crystallographers.

Coppens: We have heard discussion of the unit cost per man year in computing time; would it be higher for this kind of work?

Watenpaugh: Well, just in the refinement I've spent over $30 000 in one year. The amount required for the solution of protein structures has been small, and this is about where protein people have stopped. However, as more proteins are going to be refined, you're going to see an astronomical increase in the amount of computing time on the part of protein crystallographers.

Johnson: How much money goes into computer graphics on any protein structure?

Watenpaugh: We don't do any, but there are some groups that do, and I imagine quite a bit of time is spent in some laboratories.

Freer: We actually have been spending as much on computer graphics as on refinement.

Schomaker: Keith, is it true though that you are substantially inhibited in your progress by lack of money?

Watenpaugh: Yes.

Schomaker: You could have spent $100 000 or $200 000, or at least at that rate?

Watenpaugh: Well, we're not even at convergence yet. It's just that we actually have a good start, and we still have fairly low-resolution data. We want to collect data to high resolution.

Sayre: I think it should also be noted that rubredoxin is a small protein and that the cost of a least-squares refinement rises approximately at a rate between n^2 and n^3 where n is the number of atoms.

Coppens: So perhaps there is a need for larger computers.

Sayre: Yes. That's the point.

Koetzle: There must be some resolution limit beyond which this sort of refinement that you're talking about is not possible. To what resolution would you say the data on the protein ought to go before you can initiate this process?

Watenpaugh: On the Chromatium HIPIP, they're working with 2 Å data but I think this is the bare minimum of the resolution at which we can work. It's becoming increasingly obvious that the positions of the atoms are fairly well behaved in proteins, much as they are in small structures, and therefore, with high-speed data-collection techniques to collect the data before our crystals go to pot, we should be able to take lots of proteins to atomic resolution in the future.

Jeffrey: I have an idea that differential syntheses are like block-diagonal refinement in their convergence. So you may not gain as much by differential synthesis because the refinement would be less rapid than full-matrix.

Freer: The fact is that I can do the differential synthesis on a CDC 3600 with 32 000-word core, and that's what I have now. We're going to try to get on the 7600 at LBL. Being members of the University of California, perhaps we might have a better chance.

Jeffrey: I seem to remember a paper where someone related differential synthesis to a diagonal matrix refinement.

Freer: Yes. I think it's equivalent to least-squares weighted by the reciprocal of the atomic scattering factors.

Coppens: This was discussed in a paper by Cruickshank (Acta Cryst. 5, 511, 1952).

Seeman: Can you give the accuracy of your starting model? What's been your actual shift from the model to your current coordinates?

Watenpaugh: We've had shifts of over half an ångström, but the average shifts were probably about two-tenths of an ångström or so per cycle in the initial stages of refinement.

J. D. H. Donnay: I should like to address my comments to the funding agencies. About twenty years ago, at the Paris International Congress of Crystallography, the late Professor Mauguin referred to "those crystallographers who had the courage and the audacity to tackle the protein structures." That was in 1954. What was almost foolhardy at that time is still a job that requires considerable boldness and courage today. It seems to me that funding agencies should know that we as a profession (1) have high respect for the people who work on protein problems and (2) feel that, if a nation is lucky enough to have research people willing and able to solve such problems, it should make sure that they receive full support from their government.

Session II

What New Developments Are in the Wind?

Session Chairman: William R. Busing

New Computational Techniques, Particularly For Refinement (1)

Carroll K. Johnson

The two principal numerical techniques used to refine crystal structures (2) are the Fourier transform method and the method of linearized least-squares. The following remarks will be restricted to the least-squares approach; however, significant developments are also occurring in the Fourier field, the Fast Fourier Transform algorithm being used to decrease computing time substantially.

An important preliminary for any crystal structure refinement is the selection of an appropriate mathematical model for the structure under study. The selection is usually influenced by the following three considerations.

1. What is the relative importance to the investigator of the different types of information that can be obtained from a structure refinement?
2. Are there any unusual problems involved, such as major disorder in the structure or poor quality diffraction data?
3. Are the available computer hardware, program software, and computing budget adequate to handle the proposed refinement?

Ideally, consideration number 1 is of greatest importance, and the refinement model should be based on the particular type of chemical or physical information that the investigator wants to gain from the structure refinement. There seem to be two different areas of interest to crystallographers doing crystal structure analysis. The first area concerns the geometrical properties of the idealized configuration of point atoms (i.e., metrical properties such as distances and angles), and the second area concerns the elucidation of atomic density function properties such as electron density.

There are two different schools of thought concerning what is the best method to use in refining a crystal structure. These two schools may be termed the free-model school and the constrained-model school.

The free-model school reasons that we should refine a structure in the least restrictive way possible, with independent parameters for each atom so that the final results are unbiased by preconceived chemical concepts incorporated into the model. The most commonly used model, with 3 positional parameters and 6 anisotropic temperature factor parameters for each atom, is an example of an unconstrained model.

The constrained-model school argues that we should put as much chemical information as possible into the model so that the variables to be

determined are reduced to the basic parameters of direct interest to the investigator. Examples of constrained models are the rigid-body model, the segmented-body model, and the models which force chemically symmetrical groups to be geometrically symmetrical even though they are not crystallographically equivalent. Such constraints can be applied to both positional and thermal-motion parameters.

The majority of the crystallographers seem to follow the free-model school of reasoning. The advantage of the unconstrained model is its simplicity and easy direct application to a wide variety of problems. A disadvantage is often the large number of variable parameters that must be handled when crystal structures of even modest complexity are refined. For example, a full-matrix refinement with anisotropic thermal parameters for a 45-atom structure will involve at least 406 variables and will require 82 621 words of core storage for the least-squares matrix alone.

The economic importance of the least-squares calculation is emphasized by the survey taken by Dr. Hamilton for this symposium. The survey shows that 80 to 90% of the computing time used by U.S. crystallographers is spent in the structure-refinement step. Furthermore, the greater part of this computer time is used in forming the matrix of the least-squares normal equations; consequently, it is often worthwhile and sometimes essential to approximate the matrix by an alternate matrix requiring less computer time and less computer memory.

Table 1 lists some old and some new methods for approximating the crystallographic least-squares matrix. The principal approach used to minimize computer core requirements is to omit as many off-diagonal terms as possible, thus transforming the full matrix to a sparse matrix. The block-diagonal matrix with one atom per block is the most commonly used sparse-matrix approximation although further reduction is possible. Diagonal matrix approximations are of little value for general crystallographic refinement because of the oblique coordinate systems used for trigonal, monoclinic, and triclinic crystals.

TABLE 1 Approximations For The Crystallographic Least-Squares Matrix

1. Sparse matrix approximations
(a) Block diagonal with one atom per block
(b) Cross-word puzzle (block diagonal + first neighbor interaction terms)

TABLE 1 continued

TABLE 1 continued

2. Recycle and update approximations
(a) Use the same full matrix unchanged for several cycles
(b) Recalculate only the block-diagonal submatrices and simply rescale the rest of the old full matrix
(c) Recalculate only the matrix elements influenced by parameters which undergo appreciable shifts
3. Analytical matrix approximations

An untried but seemingly logical extension from the one-atom block-diagonal matrix is the "cross-word puzzle" matrix where all interaction terms between close-neighbor atom are added to the block-diagonal matrix. It is well known that close-neighbor atoms have a greater least-squares interaction than distant-neighbor atoms. The cross-word matrix would have to be stored by blocks and inverted with a partitioned-matrix inversion scheme.

The Recycle and Update Procedures listed in Table 1 assume that the complete matrix has been calculated and stored once and that the changes in it from cycle to cycle are small. The option of using the same matrix unchanged for several cycles was available in the original Busing and Levy least-squares program for the IBM-704. Unfortunately, there is no recorded evaluation of the usefulness of this approximation; however, most of the verbal reports received indicate erratic behavior. There are several rather obvious modifications of the Recycle Procedure which might prove to be useful. For example, the atom-block-diagonal submatrices might be recalculated each cycle and the rest of the matrix simply rescaled by the new over-all scale factor. Alternatively, an algorithm might be devised whereby the only matrix elements to be updated would be those involving parameters that shifted appreciably in the preceding cycle.

The final method listed in Table 1 utilizes a completely different approach to reduce computer time. It replaces the time-consuming numerical summations over the thousands of reciprocal lattice points by analytical integrations. The results on analytical matrix approximations presented here are from my own work; however, I recently learned that Professor Verner Schomaker at the University of Washington has derived independently a number of related results.

There are five factors which are functions of the scaled reciprocal-

lattice vector $\underset{\sim}{t}$ (i.e. $\underset{\sim}{t} = 2\pi\underset{\sim}{h}$) in each term of the sum for any particular matrix element. The five factors for a centrosymmetric structure with refinement based on F^2 are listed in Table 2.

TABLE 2 Factors In The $P\bar{1}$ Crystallographic Least-Squares Matrix Sums Which Are Functions Of The Reciptocal Lattice Vector $\underset{\sim}{t} = 2\pi\underset{\sim}{h}$*

(1) $F_c^2(\underset{\sim}{t})/\sigma^2[F_o^2(\underset{\sim}{t})]$

(2) $f_m(\underset{\sim}{t})f_n(\underset{\sim}{t})$

(3) $\exp\{-\frac{1}{2}\,\underset{\sim}{t}'[(\underset{\approx}{b}_m + \underset{\approx}{b}_n)/2\pi^2]\underset{\sim}{t}\}$.

(4) $t_i t_j$ or $t_i t_j t_k$ or $t_i t_j t_k t_\ell$; $(i,j,k,\ell = 1,2,3)$.

(5) $\left\{\begin{matrix}\cos\\ \sin\end{matrix}[\underset{\sim}{t}'(\underset{\sim}{x}_m - \underset{\sim}{x}_n)] \pm \begin{matrix}\cos\\ \sin\end{matrix}[\underset{\sim}{t}'(\underset{\sim}{x}_m + \underset{\sim}{x}_n)]\right\}$

* $\underset{\sim}{x}_m$, $\underset{\sim}{x}_n$ are positional vectors and $\underset{\approx}{b}_m$, $\underset{\approx}{b}_n$ are anisotropic thermal-motion matrices for atoms m and n.

The first factor in Table 2 contains the calculated squared structure factor divided by the variance of the observed squared structure factor. This factor can be eliminated from the list by making the following approximation:

> Approximation 1 - The magnitude of the calculated squared structure is assumed to be proportional to the variance of the observed squared factor. Consequently, the ratio $F_c^2/\sigma^2(F_o^2)$ becomes a constant for all reciprocal lattice points.

Approximation 1 is completely valid for the special case where variances are based on counting statistics alone with no correction for background counts.

The second factor in Table 2, a product of atomic scattering factors

for atoms m and n, may be replaced by an analytical expression.

Approximation 2 - The product of two atomic scattering factors is assumed to be approximated adequately by a short sum of spherical Gaussian functions.

Sums of 3 to 5 Gaussian functions currently are used successfully to replace scattering-factor table-look-up procedures in crystallographic programs. The same tabulated Gaussian coefficients could be used in a double summation; however, a more efficient procedure is to fit new Gaussian coefficients directly to the scattering factor product, taking care to make the fit acceptable for the entire range of $\sin(\theta)/\lambda$.

The third factor in Table 2, the product of anisotropic Gaussian temperature factors for atoms m and n, presents no difficulty. The fourth factor $t_i t_j \ldots t_n$ is the n^{th} degree product of the components for the three-dimensional reciprocal lattice vector $\underset{\sim}{t}$. The 2^{nd}, 3^{rd}, and 4^{th} degree products occur in position-position, position-thermal, and thermal-thermal matrix elements respectively. The fifth factor is a product of trigonometric terms which can be rewritten as a sum of trigonometric terms with arguments containing inner products of $\underset{\sim}{t}$ with an interatomic vector between atoms m and n. When the periodic properties of the trigonometric functions and the crystal lattice are considered, this factor is seen to contain the Patterson vectors between all atoms of types m and n in the crystal.

By incorporating approximations 1 and 2, we can write a simplified equation for any element in the crystallographic least-squares matrix $\underset{\simeq}{L}$. For example, for space group $P\bar{1}$, the equation for the interaction term relating the i^{th} component of the position vector $\underset{\sim}{x}_m$ for atom m and the j^{th} component of the position vector for atom n (i,j = 1,2,3) is

$$L(x_m^i, x_n^j) = K \sum_{\underset{\sim}{y}} \sum_{p} \alpha_p \sum_{\underset{\sim}{t}} t_i t_j \exp(-\underset{\sim}{t}' \underset{\simeq}{M} \underset{\sim}{t}/2) \cos(\underset{\sim}{t}' \underset{\sim}{y}) \qquad (1)$$

with the matrix $\underset{\simeq}{M}$ defined as

$$\underset{\simeq}{M} = \beta_p \underset{\simeq}{G}^{-1} + (\underset{\simeq}{b}_m + \underset{\simeq}{b}_n)/2\pi^2 . \qquad (2)$$

In these equations, K is a constant, $\underset{\sim}{y}$ is an interatomic vector between atoms on crystallographic sites m and n (i.e., $\underset{\sim}{x}_m - \underset{\sim}{x}_n$, $\underset{\sim}{x}_n - \underset{\sim}{x}_m$, $\underset{\sim}{x}_m + \underset{\sim}{x}_n$, and $-\underset{\sim}{x}_m - \underset{\sim}{x}_n$), $\underset{\simeq}{b}_m$ and $\underset{\simeq}{b}_n$ are anisotropic temperature factor matrices for atoms m and n, $\underset{\simeq}{G}^{-1}$ is the contravariant metric matrix, and α_p, β_p are

coefficients in the Gaussian expansion for the scattering-factor product for atom pair m,n.

The main step in the approximation procedure involves replacing the summation over $\underset{\sim}{t}$ in Eq. (1) by an integration over $\underset{\sim}{t}$.

Approximation 3 - We assume that enough reciprocal-lattice points are included in the reciprocal-lattice summation to justify the replacement of the summation by an integration without including higher-order correction terms.

Approximation 3 is based on the classical Euler-MacLaurin summation formula suitably generalized to the three-dimensional case. The one-dimensional Euler-MacLaurin summation formula is

$$\sum_{k=0}^{m} f(a + kh) = \frac{1}{h}\int_{a}^{b} f(t)dt + \frac{1}{2}[f(b) + f(a)] + \cdots , \quad (3)$$

where h = (b - a)/m. The higher-order terms (not shown) involve powers of h and odd-order derivatives of f at the limits a and b.

A number of special cases now occur, depending upon the integration limits. In the simplest case, the integration extends over all of the reciprocal space and we obtain the result,

$$L(x_m^i, x_n^j) = K' \sum_{\underset{\sim}{y}} \sum_{p} \alpha_p [-|\underset{\approx}{M}|^{-\frac{1}{2}} H_{ij}(\underset{\sim}{y}, \underset{\approx}{M}^{-1}) \exp(-\underset{\sim}{y}' \underset{\approx}{M}^{-1} \underset{\sim}{y}/2) \quad (4)$$

with $$H_{ij}(\underset{\sim}{y}, \underset{\approx}{M}^{-1}) \equiv z_i z_j - M_{ij}^{-1}, \quad \text{and } \underset{\sim}{z} = \underset{\approx}{M}^{-1} \underset{\sim}{y}.$$

The tensor component $H_{ij}(\underset{\sim}{y}, \underset{\approx}{M}^{-1})$ is a second-order three-dimensional Hermite polynomial. Corresponding formulas for the position-thermal and the thermal-thermal interaction have the same form as the position-position interaction equation shown in Eq. 4 except that the second order H_{ij} is replaced by the third and fourth order polynomials H_{ijk} and H_{ijkl} respectively.

Equation 4 represents the asymptotically limiting situation which is approached only when the entire reciprocal-lattice data set is included. A more common experimental practice is spherical truncation of the data set at some fixed value of $|\underset{\sim}{t}|$. An exact solution for this general case with anisotropic temperature factors and spherical truncation is quite difficult, but some success has been achieved with empirical correction factors applied to Eq. 4. If the temperature factor for each atom is isotropic and the truncation is spherical, the finite summation over $\underset{\sim}{t}$ in

Eq. 1 can be replaced by an integration which has an anlytical solution involving Legendre polynomials and error functions.

Equations 1 and 4, which are specialized for space group $P\bar{1}$, can be generalized to include any centrosymmetric space group by incorporating a double summation over the symmetry operation for the space group. The task of obtaining an analytical formulation for the noncentrosymmetric space groups is less straightforward because the first factor given in Table 1 (i.e., $F_c^2(\underset{\sim}{t})/\sigma^2[F_o^2(\underset{\sim}{t})]$) is replaced by three terms. For example in the noncentrosymmetric space group $P\bar{1}$, the matrix element for a position-position interaction may be written

$$L(x_m^i, x_n^j) = K \sum_p \alpha_p \sum_{\underset{\sim}{t}} \frac{A_c^2 + B_c^2}{\sigma^2(A_o^2+B_o^2)} t_i t_j \exp(-\underset{\sim}{t}' \underset{\approx}{M} \underset{\sim}{t}/2)\cos[(\underset{\sim}{x}_m - \underset{\sim}{x}_n)'\underset{\sim}{t}]$$

$$- K \sum_p \alpha_\rho \sum_{\underset{\sim}{t}} \frac{A_c^2 - B_c^2}{\sigma^2(A_o^2+B_o^2)} t_i t_j \exp(-\underset{\sim}{t}' \underset{\approx}{M} \underset{\sim}{t}/2)\cos[(\underset{\sim}{x}_m + \underset{\sim}{x}_n)'\underset{\sim}{t}] \qquad (5)$$

$$- K \sum_p \alpha_p \sum_{\underset{\sim}{t}} \frac{2A_cB_c}{\sigma^2(A_o^2+B_o^2)} t_i t_j \exp(-\underset{\sim}{t}\ \underset{\approx}{M}\ \underset{\sim}{t}/2)\sin[(\underset{\sim}{x}_m + \underset{\sim}{x}_n)'\underset{\sim}{t}],$$

where A and B are the real and imaginary parts of the structure factor and $F^2 = A^2 + B^2$. The problem is to predict the behavior of the factors $(A_c^2 - B_c^2)/[\sigma^2(A_o^2 + B_o^2)]$ and $A_cB_c/[\sigma^2(A_o^2 + B_o^2)]$. Intuitively, it seems that the terms containing these factors may tend to integrate to zero if the entire reciprocal lattice is included and if the structure is a "random structure"; however, the conjecture has not been proven. The integral behavior of these terms for a real structure with a truncated data set seems rather unpredictable.

Evaluation of the analytical matrix technique is underway. With favorable conditions (i.e., low crystallographic symmetry and extensive, but finite, diffraction data) the computing time required to form the matrix has been reduced by an order of magnitude. For cases where the symmetry is very high and the data collected are not extensive, there may be no saving of time.

The principal testing of the procedure to date has been for an application rather different from least-squares refinement. We use the inverted analytical matrix to calculate the complete variance-covariance matrix for a published structure without computing structure factors or their derivatives. The only data needed to generate the analytical matrix are the structural parameters, a matrix scale factor, and a truncation parameter. The latter two parameters can usually be obtained quite

accurately from the standard deviations, which usually are published with the crystal-structure paper.

In addition to the matrix approximations described above, there are also possibilities for saving computer time by utilizing some special redundancy properties of the full crystallographic least-squares matrix in space group P1. The basic approach is simply to examine the equations for the elements in the matrix. An example is for a hypothetical structure displaying space group symmetry $P\bar{1}$, with two atoms (m and n) in the asymmetric unit. If we write out the equation for the 171 supposedly unique elements in the symmetric 18 by 18 matrix, $\underset{\sim}{L}$, for positional and anisotropic thermal parameters, we quickly discover that considerable redundancy is present, only 103 elements actually being unique. The remaining 68 elements are simple multiples of other elements. For example, we find that $L(b_m^{1\,2}, b_m^{1\,2}) = 4L(b_m^{1\,1}, b_m^{2\,2})$ and $L(b_m^{1\,3}, b_n^{2\,3}) = L(b_m^{2\,3}, b_n^{1\,3}) = 2L(b_m^{3\,3}, b_n^{1\,2}) = 2L(b_m^{1\,2}, b_n^{3\,3})$. For other centrosymmetric space groups, the redundant linear combinations are fewer and more complicated.

DISCUSSION

Sparks: Concerning the calculation of correlation coefficients from position parameters and thermal parameters, I always thought that wouldn't work if you had a situation where you had refined a set of data having not only the termination problem you mention but also a lot of weak reflections that are just left out of the data set. It would seem to me this would tend, in some peculiar way, to bias the results.

Johnson: The numerical agreement between the variances calculated from the regular inverted least-squares matix and those calculated from the inverted analytical matrix is usually quite good for any data set. The agreement for the covariances become much better if the data set is fairly extensive. Missing reflections may present a serious problem if the data set is quite sparse, but we have not examined this aspect numerically or theoretically.

Templeton: This sounds like magic until you think about it, but there's a way of restating it which I think makes evident what you're doing. If you have a published structure, and use this to calculate structure factors, this is a data set more or less like the experimental data set, more to the structure's right and less to the structure's wrong. From that data set you select, depending on your knowledge, which reflections have been left out. For example commonly people say, "We observed 1600 reflections of which 400 were zero." If you simply chop off the 400 smallest ones then you would have a very good replica. One thing I noticed in your suggestion that one leave out matrix elements not affected by temperature, it is not evident how you know which ones these are because they're not necessarily just the ones labeled by the subscripts of the parameter that is shifted.

Johnson: I have to admit that I have not thought this through in detail and cannot at present describe an algorithm that would keep track of the major changes from cycle to cycle.

Templeton: Then part of my next question is, how do you know which ones they are, because all of the derivatives include in them the structure factor?

Johnson: But we have in this formulation eliminated the structure factor, but you're right, very good point. However, the structure factor was eliminated in the analytical formulation and perhaps, as an approximation, the same reasoning could be extended to this approach.

Hamilton: Would you predict that this may be the answer for people who are refining protein structures and that they should really be thinking seriously about this method?

Johnson: I must admit I harbor the fond hope that the analytical approach might someday be applicable to protein refinement. Unfortunately, I cannot at present see how to handle the non-centrosymmetric problem properly. Luckily, you scheduled me in a session where I can discuss what should be done and not necessarily what can be done. I think it is certainly worthwhile to put some additional effort into this approach to see if it might be a feasible solution for proteins.

Busing: You say this speeds up the computation of the matrix by perhaps a factor of 10. If the number of observations and parameters is very large, does this method become even more favorable?

Johnson: The approximation improves as the number of observations increases and we are in best shape if the data set includes everything that can possibly be measured. In this case we also obtain our maximum time advantage. Additional parameters may also improve the time advantage because the sum over the Patterson vectors converges rapidly as a function of interatomic separation; consequently the long vectors can safely be omitted from the summation.

New Computer Needs for Direct Methods

D. Sayre

Our purpose in this part of the symposium is to try to foresee whether new developments in crystallographic techniques are likely to generate changes in the computational resources that crystallographers should have. In this paper we consider whether such changes are likely to arise out of developments in direct methods.

Structures of Moderate Size

In the direct methods, the problem of structure determination is converted into the mathematical problem of solving a system — usually a large system — of equations or relations involving the structure factors as principal variables. The equations or relations express the non-negativity, atomicity, or other property of the structure, and the phases of the structure factors are the quantities solved for. Up to a certain size of structure it is feasible to attempt the solution without any initial information about its location, i.e. the solution can be attempted ab initio. In this case, which is called the pure direct method case, if the solution is successful a very convenient structure-determining method results.

In the usual formulation of this approach, the basic step is that of passing from a situation in which the phases of i of the structure factors exist (these phases being such as to satisfy the relations being solved) to the situation in which i+1 values (still satisfying the relations) exist. The information available for this i^{th} step consists of all relations involving the as yet unassigned structure factor phases, which can be written in terms of the structure factor magnitudes plus the i phases already assigned. Unfortunately, the phase-limiting relations known to us today have the property of being generally weak for small values of i, though they become reasonably strong by the time i reaches, say, 10N, where N is the number of independent atoms in the structure; moreover the larger N is the weaker the relations become for small i. The result is that the solution process must pass through an initial stage in which the partial solutions that must be retained in some form if the possibility of missing the solution altogether is to be avoided, may branch to impractically large numbers, and the severity of this stage rises steeply with N. (It rises roughly as m^n, where n is the number of branch-points and m is their average multiplicity, and where both are increasing functions of N.) This rapid rise with N in the average time needed to produce a solution, although its exact behavior as a function of N is not known, is what puts the limit on the size of structure that can be handled by purely direct methods.

For a procedure of this kind, if its method of exploring partial

solutions allows it quickly to go deeply into some of those solutions, it may produce a solution to a problem of size N in a time short compared to the time needed to guarantee a solution at that size. This effect complicates the interpretation of empirical data on the relationship between the difficulty of solution and N. Thus we may state on the basis of the recent solution by pure direct methods of the structure of adenosine triphosphate, with 72 atoms in the asymmetric unit of $P2_12_12_1$ (Kennard et al., 1971), and of that of valinomycin, with 78 atoms in $P2_1$ (Duax and Hauptman, 1972), that the difficulty of solution is probably not in general beyond the range of our computers up to N=75, but this conclusion must be somewhat tentative at present.

Much of the work going on in direct methods today can be traced to the desire to strengthen the phase-limiting relations at low values of i. Thus Karle (1971) has recently re-examined the determinantal inequalities introduced into crystallography by himself and Hauptman and has suggested that E_h may be more closely located by higher-order forms of the determinants than by the m=3 forms commonly used until now; at the i^{th} step in the solution process orders up to the $i+1^{st}$ would be available. Similarly Weeks and Hauptman (1971) have re-examined the tangent formula which is often used to produce an estimate of an $i+1^{st}$ phase from the values of i phases and have suggested a modification intended to reduce somewhat the errors which occur when i is small. Again the interest in the method of Tsoucaris (1970) stems from the possibility that his principle of estimating the phase of E_h by maximizing a certain Karle-Hauptman determinant may provide a still more accurate localization of that phase. Finally both Karle (1970) and Hauptman (1970) have recently been concerned with improving the accuracy of the type of formula that permits approximate values, if not of the phases at least of the cosines of the structure invariants, to be computed from the structure factor magnitudes alone; these values are therefore available at every step i. Indeed the most interesting thing about this latter type of relation may be the fact that the possession from the outset of the full set of cosine invariants, even in approximation, allows one to think in terms of pure direct methods that are not step-by-step in nature; these may possibly have no greater overall chances of success than the step-by-step technique, but would at least be free to some extent of the extreme branching in which the latter can so easily get caught. For example, in the centrosymmetrical space groups the cosine invariants are exactly what is needed to compute the double Patterson function.

Large Structures

Above the size of structures we have been considering thus far lies the entire range of large structures. Although these have not been approachable by direct methods up to now, it may well be that the most significant developments in direct methods will relate to structures of this class. The circumstance making this possible is that certain equation-solving

techniques can accept some initial information about the location of the solution, while physical phasing techniques based upon the inclusion of heavy atoms can provide such initial information, though not sufficient to determine all desired phases. The most dramatic instance of this mixed type of solution to date is that of Sobell et al. (1971) in solving the structure of an actinomycin-deoxyguanosine complex containing 140 atoms in the asymmetric unit of $P2_12_12_1$. In this case the single heavy-atom isomorphous replacement technique was followed up both by tangent-formula refinement and Patterson search methods. Going upward in size, tangent-formula refinement has been attempted without success in the extension of protein phases beyond 3 Å resolution, by Hoppe and Gassmann (1964), and by Reeke and Lipscomb (1969). An attempt has been made also by Dickerson, but with results that have not been published. Barrett and Zwick (1971) attempted phase extension beyond the limit of multiple isomorphous replacement on myoglobin by a different method, which is not really a direct method, but without success.

Some recent work of my own on a method similar in aim to tangent-formula refinement may be of interest. The method may be characterized as a stronger solution technique than tangent-formula refinement, but also computationally more expensive. The first stage of the work (Sayre, 1972) consisted of a demonstration that structure factor phases can be refined in much the same way as we refine atomic parameters, i.e. by least-squares minimization of an appropriate function of the quantities being refined. The structures considered obeyed the relations $a_{\underset{\sim}{h}}F_{\underset{\sim}{h}} = \sum F_{\underset{\sim}{k}}F_{\underset{\sim}{h}-\underset{\sim}{k}}$ and the refinement process consisted in minimizing the function $\sum|a_{\underset{\sim}{h}}F_{\underset{\sim}{h}} - \sum F_{\underset{\sim}{k}}F_{\underset{\sim}{h}-\underset{\sim}{k}}|^2$. The favorable thing about these refinements was that they could be successfully initiated from a quite incomplete set of starting phases. For example, it was sufficient to have the phases out to 3 Å resolution, with no phase information beyond that point initially available. The principle requirement in using the above function for minimization is that there be available a reasonably complete set of magnitudes out to 1.5 Å or better; thus, given an initial set of phases over the low-resolution terms, the effect of the refinement process is to produce a set of phases extending over the higher-resolution terms as well. Significant errors in the phases initially given could be tolerated and corrected by the refinement process. Final mean errors for the phases ran from 8° to 10° in these experiments, in which all magnitudes were assumed to be accurately known.

This direct phase-refinement process for phase extension is now being tried on the protein rubredoxin. L.H. Jensen at the University of Washington has kindly supplied observed magnitudes for rubredoxin to 1.5 Å together with phases to 3 Å as determined several years ago by multiple isomorphous replacement and anomalous dispersion techniques. A refined set of phases for the full set of magnitudes (5034 terms in all) has been produced by the above process, and the resulting Fouriers at 1.5 Å resolution have been sent to Jensen for evaluation. The Fouriers are definitely those of an atomic structure, but it is not yet known whether the method placed the atoms in the correct positions.

The use of this technique in protein crystallography carries with it certain difficulties. Firstly, the computation is a massive one, owing to the large set of quantities to be refined. On rubredoxin each cycle of the refinement takes just under an hour on a large machine (IBM 360/91), with a total of approximately 35 hours required for the refinement as a whole. The cost of this amount of computation at a typical large computing installation might be about $25 000. This figure could be reduced through reprogramming and redesign of the algorithm, but no real estimate of the reduction has been made. On the negative side, the computation increases approximately as N^2, N being the number of independent atoms. Of course if one drops below the protein size-range, the N^2 dependence helps considerably; thus at 200 atoms a refinement might cost about $5000.

Secondly, although the method will seek out a minimum of the function in question, there is little experience as yet with real structures concerning the likelihood that the minimum is the one corresponding to the physically correct structure. In any event, the phases obtained by this type of refinement should not be regarded as final phases but as being subject to final refinement by the usual methods. The phase-refinement technique, in other words, should be thought of as a possible bridge from the preliminary phasing techniques to final refinement of atom parameters.

In another development Luzzati, Tardieu, and Taupin (1972) have described the phasing of reflections from biological membranes and related systems by a direct method. The method is based upon the recognition that structures of this type consist of separate constituents (water, hydrocarbon, and protein, for example) with a somewhat different electron density for each. At low resolution, therefore, the electron-density function for the structure should be divisible into regions such that the density is reasonably flat within each region and changes fairly rapidly only at the region boundaries. The authors have found a means of expressing this requirement in terms of conditions among the structure factors, and have devised a discrete search method for selecting sets of phases by computer that approximately satisfy the requirement. The method is apparently in regular use at their laboratory, where it facilitates considerably the structure determination for this type of system. It shows the possibility of direct methods in structure work that does not even approach atomic resolution. The authors indicate a belief that the technique may be applicable to other biological macromolecules, at low resolution, including proteins.

Conclusions

1. For the crystallography of structures below 30-40 atoms in size, developments in direct methods will probably not have much future effect on the scale of computational needs. This does not mean that future

technical improvements in direct methods for such structures are in any way ruled out, or that each individual crystallographer working in this area today necessarily has adequate computing resources, but simply that the main impact of direct methods on computation for such structures has already largely taken place.

2. The steep general rise of difficulty of the pure direct methods with N makes a gradual extension of their range beyond the present approximate limit at N=75 much more likely than a rapid one. From the point of view of computational resources, maintaining this gradual extension is likely to involve a rather steeply increasing need for resources. In addition, as N increases, not only the mean time but the uncertainty of the time to solution can be expected to rise, i.e. the actual time to solution in any given case may prove to be much smaller, or much larger , than the predicted mean time. As larger structures are attempted by purely direct methods, this unpredictability of machine time will increasingly have to be allowed for in the budgeting of resources. This latter conclusion may not necessarily hold for the non-step-by-step type of solution, if that type of solution should develop.

3. The next few years may see for the first time the application of direct methods, sometimes in combination with other methods, to proteins and other large structures. If this occurs, very large machine resources will be required.

References

Barrett, A.N. and Zwick, M. (1971). Acta Cryst. A27, 6.

Duax, W.L. and Hauptman, H. (1972). Program of ACA Winter Meeting, Albuquerque, paper 03.

Hauptman, H. (1970). Proceedings of the International Summer School on Crystallographic Computing, Ottawa, August 1969, pp. 41-51. Copenhagen: Munksgaard.

Hoppe, W. and Gassmann, J. (1964). Ber. Bunsenges. 68, 808.

Karle, J. (1970). Acta Cryst. B26, 1614.

Karle, J. (1971). Acta Cryst. B27, 2063.

Kennard, O., Isaacs, N.W., Motherwell, W.D.S., Coppola, J.C., Wampler, D.L., Larson, A.C., and Watson, D. G. (1971). Proc. Roy. Soc. Lond. A325, 401.

Luzzati, V., Tardieu, A., and Taupin, D. (1972). J. Mol. Biol. 64, 269.

Reeke, G.N. Jr. and Lipscomb, W.N. (1969). Acta Cryst. A25, 2614.

Sayre, D. (1972). Acta Cryst. A28, 210.

Sobell, H.M., Jain, S.C., Sakore, T.D., and Nordman, C.E. (1971). Nature New Biology 231, 200.

Tsoucaris, G. (1970) Acta Cryst. A26, 492.

Weeks, C. and Hauptman, H. (1971). Program of ACA Winter Meeting, Columbia, South Carolina, paper H4.

DISCUSSION

Kartha: While it is quite true that the additional reflections did sharpen up the map, the crucial question is, have you got the phases between 3Å and 2Å correctly? Did you make any comparison of these phases with those obtained by the use of isomorphous series and anomalous scattering data in this region?

Sayre: Until now I have not known any of Jensen's phases except his experimental phases to 3Å, and as a consequence I have not been able to make the comparisons you mention. At this meeting, however, Jensen has given me a tape containing his latest set of phases, and I expect to start making comparisons between the two sets within a few days.

Kartha: A map using only the 3Å-2Å would be worth calculating to see how the map with your phases on the one hand and the isomorphous series phases on the other compares.

Sayre: Perhaps so. I personally have thought that the most interesting comparison would be with Jensen's final 1.5 Å set.

Dewar: If direct methods start to involve large amounts of computer time, we should be wary of using criteria like computer costs per

structure and computer costs per man year, because it may actually represent a good use of resources to crank those figures up to a much higher level than they are now. We tend to write off other costs to some extent, partly because the cost of estimating other costs is difficult in university environments. We may worry about the computer costs, but personnel costs are extremely difficult to assess. I think you should not pay too much attention to statistics of this type, as this kind of work will tend to change the picture considerably.

Hall: I've been assisting Dr. Sobell at Rochester in trying to solve another form of the actinomycin complex you mentioned. So far this has been unsuccessful but it may be of interest to note the computer times involved in the calculation of cosine averages for this particular complex. There are approximately one thousand E's above 1.4, ten thousand reflections altogether. A cosine calculation for one invariant takes about two seconds on the UNIVAC 1108 and when you consider that there are about 80 000 invariants, this represents a sizable amount of computer time. I don't think we can justify that sort of expenditure now. You might, of course, only consider looking at a thousand of those invariants. Even so, two thousand seconds on UNIVAC 1108 time is still a significant cost in such an analysis.

Sayre: I interpret your remark as pointing out the cost of a structure-determining method that uses most or all of the cosine invariants, even if it permits the avoidance of a step-by-step approach. Doubtless your point is well taken. I think we might agree that the central question concerns the expansiveness of the number of cosine invariants relative to that of the solution tree in a step-by-step solution procedure.

Hall: I'd like to acknowledge that Dr. Hauptman and the group at the Medical Foundation in Buffalo have been working in collaboration with us on this structure and that it was in fact the MDKS average that I referred to.

Sayre: Which is a more complicated one.

Hall: Right. And it was just a preliminary calculation to try and get some estimate of the reliability of the invariant relationships.

Donnay: With that beautiful map you have shown us, are you not at a point where you can place atoms and get some idea of an R index?

Sayre: I suppose that would be possible, but in fact the map was produced only about 5 days ago.

Donnay: But this could eventually be an indication of whether or not you are heading in the right direction?

Sayre: Not as yet. In this work I have until now avoided thinking about atom positions, and have been careful to treat the problem entirely as a problem in equation-solving. I plan to continue in this vein for the present. Of course, should the phase-refinement method prove to be a generally valid one, it would as you say be at the present stage that one would try to place the atoms.

Kartha: You mention the difficulty of using direct methods as a function of the number of atoms, but in protein crystallography usually there is another problem—that not only the number increases but the ratio of the number of parameters to number of reflections goes up very much. What is the correlation?

Sayre: Regarding the minimum-finding process, suppose there are p phases to be adjusted to produce a minimum. The number of available squaring-method equations, s, will be at least p, and the number of observational equations will therefore be at least 2p, since each squaring-method equation is an equation between complex quantities. Thus the minimization process can always be carried out under conditions of at least 2-fold overdetermination. When data are complete within the CuK sphere, the minimum corresponds to the true phases with an average error that may be only 1 - 2°. Under conditions of more severe data-incompleteness, the position of the minimum may move further away from the position corresponding to the true phases. Based on experiments with artificial structures, an average error of 8 - 10° may be expected from this cause as a consequence of the data incompleteness typical of protein crystallography. In these experiments an average temperature factor of 20 was assumed, and an instrumental sensitivity such that 8 independent reflections per atom (excluding H atoms) would be observed.

Xuong: What is the influence of the data accuracy on the results of this kind of refinement?

Sayre: I have only preliminary results bearing on that question. It appears, again using artificial structures, that the structure factor magnitudes can be perturbed with random errors up to a standard deviation of 30% without as a rule preventing the refinement from locating the correct minimum. Similarly the procedure appears normally to tolerate the introduction of random errors up to a standard deviation of 30° in the starting set of phases (to 3 Å).

The Role of the Minicomputer in the Crystallography Laboratory

Robert A. Sparks

It is universally recognized that the minicomputer plays an important role in the crystallographic laboratory. The semi-automatic and automatically controlled diffractometers have offered welcome relief for crystallographers who previously had to measure large amounts of data manually.

As with most computer-controlled instruments, early programs tended to be written to collect data in much the same way that the crystallographer would have used to operate the instrument manually. Later programs have taken advantage of the flexibility of the computer to perform tasks that would have been virtually impossible to perform with the manual or semi-automatic diffractometer.

Thus, it is now possible to:

1. Automatically center reflections.
2. Sample the profile for each reflection during data collection.
3. Choose different scan speeds dependent on the intensity of the reflections.
4. Search for peak maxima for all but the weakest reflections.
5. Measure reflections at many azimuthal angles about the diffraction vectors.
6. Measure regions of reciprocal space in a three-dimensional fashion to obtain diffuse-scattering information.
7. Automatically redetermine the crystal orientation if the crystal should move during data collection.
8. Obtain information about crystal quality and crystal symmetry.

All of this can be done with a slow computer having a minimal amount of core (4000 words). For subsequent processing of the collected data a magnetic tape drive is desirable. Magnetic tape is chosen because it is an inexpensive means of storing large amounts of data in formats that are universally recognized by large and small computers.

As collection methods become more complex one soon realizes that the limiting factor is the amount of core available. Thus, it is becoming fairly common for computer-controlled diffractometers to have more than 4000 words of core or to have a disk for program overlays.

For reasons of economy, manufacturers of computer-controlled diffractometers have chosen the least expensive computers that can easily be interfaced to the many control and acquisition functions of the diffractometer. Almost all commercial instruments are of the one instrument-one minicomputer type. Other configurations, such as several instruments-

one medium-size computer or one instrument-one minicomputer-communication link-large computer, have the possible advantage that more computer capability becomes available for the diffractometer for at least part of the time. However, such approaches are expensive because they are almost always one of a kind. The advantage of the one instrument-one minicomputer is that the development cost, of which a large part is for software, can be spread over many identical instruments.

Although the diffractometer experiments are slow, the inexpensive minicomputers used to control the diffractometers are not slow. The state of computer technology is such that computers with memory-cycle times of about one microsecond and execution times of one to three microseconds for most commands are no more expensive to build than computers that are one-half or one-tenth as fast. Therefore, the minicomputer is used to perform some calculations that do not use the diffractometer-control features of the computer. Thus, the minicomputer is used to determine indices for reflections, best least-squares unit-cell parameters, and Lorentz-polarization factors. All of these tasks can be performed by the large computer but are more conveniently done by the minicomputer.

Some crystallographers have used the small computer for tasks more traditionally performed on the large computer. Thus, Eric Gabe uses the PDP-8 which controls his Picker diffractometer to do structure-factor calculations. Shiono for many years has used the IBM 1130 to do almost all types of crystallographic calculations. For the most part, however, crystallographic computations are done on the most powerful computer available. Why this is so is illustrated in the first two columns of Table 1, which compare the characteristics of the Nova 1200, used to control the Syntex $P\bar{1}$ Autodiffractometer, with the characteristics of the CDC 6600, one of the most powerful computers used for crystallographic calculations. Although core speeds are not very different, the CDC 6600 can achieve effective speeds of up to 100 nanoseconds because the memory has been divided into independent blocks of 4096 words each. In every other respect the CDC 6600 is a much more powerful computer. Because of the large core memory, all structure factor data and the normal equations of a least-squares program can be resident. Because of the fast instruction registers, tight loops can be executed with no need to continually reference slow core. Because of the many arithmetic units and addressing and indexing registers, many operations can take place simultaneously. Not of least importance is the fact that CDC has an excellent FORTRAN compiler which makes efficient use of all this sophisticated hardware.

On the other hand, the minimal Nova 1200 configuration does not have enough core and is so slow that many of the important crystallographic programs would be virtually impossible to run for all but the smallest structures. I believe, however, that the most serious limitation is the unavailability of compilers of higher-level languages producing progams to minimize amount of core needed at run time. This deficiency has

meant that the crystallographer has not easily been able to tailor his data-collection programs to meet his requirements.

The minicomputer industry however, is, advancing rapidly. The industry is extremely competitive and prices for all parts of the hardware are decreasing at a phenomenal rate. New innovations - for example, semiconductor memory - are introduced into minicomputers almost simultaneously with introduction in large computers. Good compilers with full FORTRAN IV capability are available for computers with larger core memories (12 000 words or more for the NOVA computers). Disk operating systems that are as flexible and easy to use as those found on large computers are now arailable. Finally, fast floating-point hardware is now optional from some of the minicomputer manufacturers and also from several independent firms.

The third column of Table 1 lists the characteristics of a system that would satisfy all or almost all of the crystallographer's computing needs. In addition to the basic 4000-word NOVA with a magnetic tape drive required for the $P\bar{1}$ Autodiffractometer, this system has an additional 12 000 words of core, 131 000-word fixed-head disc, and floating-point hardware. Software consists of FORTRAN IV and a Disc Operating System. Crystallographic data-processing programs would be written in FORTRAN IV. Programs would reside on magnetic tape reels and be loaded on to disc when needed. Large programs would consist of several overlays. Large arrays would also reside on disc and be brought in to core one sector at a time. Diffractometer programs would also be written in FORTRAN IV but using machine-language subroutines for driving the goniometer axes, reading the encoders and scaler, opening and closing the shutter, etc.

Table 1 Comparison of Nova 1200 and CDC 6600

	CDC 6600	Minimal Nova 1200	Nova 1200 with Structure Determination Package
Magnetic Tape Drive	many	1	1
Core Speed	1.0 μs	1.2 μs	1.2 μs
Word Size	60 bits	16 bits	16 bits
Core Size	131 000 words	4000 words	16 000 words 131 000-word disk

Table 1 continued

Table 1 Comparison of Nova 1200 and CDC 6600

	CDC 6600	Minimal Nova 1200	Nova 1200 with Structure Determination Package
Operand, Addressing, and Indexing Registers	24	4	6
Fast Instruction Registers	8 (60 bits each)	None	None
Floating Multiply	1μs (60 bits)	2 ms (32 bits)	15.6 μs (32 bits) 24.2 μs (64 bits)
Arithmetic Units	10	1	2
FORTRAN IV & Operating System	Very Good	No	Good

Almost all crystallographic programs could be run on such a system. It is hard to justify the cost of a plotter for the infrequent use crystallographers would make of it. Therefore, it would probably be most economical to generate plotting information on magnetic tape on this system and then have the actual plotting done at central facilities. Fourier maps would be generated on magnetic tape and either printed at central facilities or printed on the slow printer by the NOVA 1200. At ten characters per second a large Fourier map could take several hours to print. In many cases, good peak-picking programs exist that eliminate the need to print the maps.

There is no question that such a system is feasible for almost all crystallographic calculations. The structure of vitamin B12 was solved and refined on a computer with a configuration closer to the basic NOVA 1200 than to the system proposed here. Indeed much of the philosophy of disc (or drum) usage and external plotting and printing of large files is identical to that used on the large computers of 5 and 10 years ago.

Time-sharing of data collection and data processing presents problems not associated with the amount of core or the arithmetic processing speed, but rather with the allocation of peripherals to the two tasks. Data collection must have the magnetic tape drive available for output of the intensity data. Therefore, production of a Fourier map could not be done simultaneously with data collection. However, Fourier calculations are quite fast (except for printing) and interrupting data collection for the few minutes necessary to generate the map and output it to magnetic

tape is not a serious limitation. Happily, the least-squares calculation which takes the bulk of time for structure determination requires the magnetic-tape drive only for the brief time necessary to dump all the data onto disc. After this, several operations could be performed without using the magnetic-tape drive and could be effectively overlapped with data collection.

The proposed system in the crystallographer's laboratory is clearly more convenient than a centralized computing facility. It is, in most cases, also more economical.

The inner loop of a least-squares program (namely the generation of the normal equations) was written in FORTRAN and executed on a number of different computers. The program is shown in Figure 1 and the results of the test in Table 2. The FORTRAN compilers that produce the most efficient code were used on the CDC 6600 and IBM 370/155. Because the floating-point hardware is fairly new for the NOVA machines, the FORTRAN compiler has not yet been modified to produce code for this feature. Reasonable substitutions were made in the assembly listing generated for the software floating-point version in order to produce the "FORTRAN-like" code. The hand-optimized version was an assembly language program written to be executed as efficiently as possible. If the matrix is large enough to require that it be stored on disc, the data-channel transfers would increase the NOVA 1200 times in this example by about 0.8. If 64-bit floating-point numbers are required, an increase of about 25% is required for the NOVA 1200 times.

Table 2 Comparison of Time for Least-Squares Inner Loop

	time
CDC 6600	0.93 s (60-bit words)
IBM 370/155	7.5 s (32-bit words)
HP 2100 A (Hardware multiply/divide Software floating-point)	150 s (32-bit words)
HP 2100 A (Hardware floating-point)	29 s (32-bit words)
Nova 800 (Software floating-point)	206 s (32-bit words)

Table 2 continued

Table 2 Comparison of Time for Least-Squares Inner Loop

	time
Nova 800 (Hardware floating-point)	
FORTRAN-like code generation	16.8 s (32-bit words)
Hand-optimized code	13.2 s (32-bit words)
Nova 1200 (Software floating-point)	360 s (32-bit words)
Nova 1200 (Hardware floating-point)	
FORTRAN-like code generation	24.2 s (32-bit words)*
Hand-optimized code	17.5 s (32-bit words)*

*Calculated from Nova 800 performance.

Typically, we at Syntex use about one hour of CDC 6600 computer time for a structure with 40-50 non-hydrogen atoms in the asymmetric unit. If the FORTRAN test in Figure 1 is typical, the "FORTRAN-like" time on the NOVA 1200 would be 26 hours for this same structure. This amount of time is small compared to typical data collection times of one to two weeks. Even without overlap of data collection and data processing there would not be a serious deterioration of diffractometer usage. With simultaneous least-squares calculations and data collection, diffractometer servicing will be negligibly affected.

```
      N = 64
      NREF = 100
      M = N + 1
      MM = M + 1
      DO 6001 IP = 1, NREF
      DO 20 I = 1, M
   20 DV(I) = I **IP**0.9
      K = 1
      DO  5001 J = 1, N
      B = DV (J)
      IF (B.NE.0)  GO TO 5002
      K = K + MM - J
```

Figure 1 FORTRAN test program (continued)

```
          GO TO 5001
5002      DO 5003 L = J, M
          A (K) = A (K) + DV (L) *B
5003      K = K + 1
5001      CONTINUE
6001      CONTINUE
```

Figure 1 FORTRAN test program (continued)

Even though the NOVA 1200 with floating point hardware is 26 times slower than the CDC 6600 and 3.2 times slower than the IBM 370/155, turn-around time will in many cases favor the dedicated computer because it is located in the crystallographer's laboratory.

Another important feature of the small dedicated system compared to the large very fast computer is that it is impossible on the former system to find out one day that an error made by a student has exhausted the year's computer budget.

Because of the above arguments, Syntex has decided to make available to customers a Structure Determination Package which would consist of a 131 000-word fixed head disc, 12 000-word core, and floating point hardware for those who already have a $P\bar{1}$ Autodiffractometer or AD-1 Autodensitometer, and a stand-alone unit consisting of a NOVA 1200, a 131 000-word fixed head disc, 16 000-word core, floating point hardware, and a magnetic tape drive for those who do not have the Syntex instruments. Software will consist of a FORTRAN IV compiler modified to make efficient use of the floating point hardware, a Disc Operating System modified to allow time-sharing of data collection with data processing, machine language subroutines for the diffractometer, FORTRAN versions of the current diffractometer programs, and FORTRAN programs properly broken up into overlays for the basic crystallographic programs. The user will be able to add his own FORTRAN or assembly language programs to the library. At this early stage, it looks quite probably that the selling price would be $30 000 for hardware and software for the attachment to existing instruments, and about $45 000 for the stand-alone option. First deliveries are scheduled for the second quarter of 1973.

The comparison of the cost of the system proposed here compared with existing costs at centralized computing facilities is difficult to make. University computing centers may charge the scientist anywhere from nothing up to the actual cost of the computing service, depending on what other sources of funds are available to the centers. Commercial rates are set to provide a profit for the company providing the service, but are usually complex functions of CPU time, amount of core used, amount of

input and output, and job priority. In Palo Alto the Control Data Center provides the most economical service for crystallographic type problems. Syntex pays about $1000 per structure for their service. Clearly, for us, the break-even point would be 30 structures for the $30 000 attachment or 45 structures for the $45 000 stand-alone configuration.

In conclusion, whether crystallographers would be inclined to buy the Syntex package or whether they would wish to buy directly from the computer manufacturers and provide their own software, I believe that serious consideration should be given to the small dedicated computer. Not only does it provide the desirable features of FORTRAN data-collection programs and the convenience of having one's own computer, but it also provides, in many cases, a substantial cost saving compared to the centralized computer approach used by most crystallographers today.

DISCUSSION

Young: If you are doing full-matrix least-squares, what is the maximum number of parameters you can handle with this sort of adorned mini-system?

Sparks: It turns out to be the same figure Jim Ibers quoted, 240, because the disc size is 131 000 words. A good suggestion by Mike Murphy is that instead of using a fixed-head disc as we are here, we ought to be using a movable-head disc which costs quite a bit less for the amount of disc space that would be available. Then the capacity would be something like two million words.

Young: If you put all those core packages and discs plus an extra arithmetic unit on the mini-computer, why do you bother with putting a diffractometer on it?

Sparks: I've given you the choice. $45 000 or $30 000.

Young: No. My point is that what you've done is build a separate computer system, and the fact that the diffractometer is hooked on is incidental.

Sparks: It does give some capability for the collection programs that we do not now have. A couple of years ago you made a strong point that we ought to be writing these collection programs in FORTRAN.

Lowrey: Professor S. H. Bauer at Cornell University has an extensive system for electron diffraction that is built around the PDP-8 and he has made extensive use of cathode-ray-tube display. He is able to search his electron diffraction data and his radial distributions and look at very fine portions. With respect to Fourier maps, instead of having to print them out you can set up a graphic interaction display for picking out the things you want. Bauer is able to do a great deal of electron diffraction using solely the small computer. He considers the advantage is that not only is it cheap but it is under his direct control so that he can run all night and have a guarantee of getting his programs back, and not have the problems of priorities on commercial computing systems.

Sparks: We also sell a three-dimensional display.

Ibers: Two points might be kept in mind. (1) It is easier to get computing money in a grant than it is to get $45 000 to buy a small computer. (2) It may be possible to sneak small computers into laboratories throughout a campus by claiming that these computers are controlling experiments, but their presence makes computer center directors very nervous, for good reason. If the small computer proliferates throughout the campus you are in trouble. Suppose we have 20 computers of the type you have discussed. In effect a million dollars has been spent and it has not benefited the central computing facility at all. For the good of the university community it might have been more reasonable to put that million dollars into the central facility. In any event there are obviously political problems that are by no means negligible.

Sparks: Yes. I am aware of this. My feeling is that the instrument ought to be treated as having a very special application. It is not by any means a general purpose computer.

Fritchie: Do you have any idea what the annual maintenance costs are on this $45 000 system? Computer alone perhaps?

Sparks: I do not have that figure. What is it on the diffractometer?

Dewar: It will be around 7%.

Coppens: What is the capacity of the system? In other words, how many crystallographers can it handle?

Sparks: It depends on how productive those crystallographers are. It's

better to say, how many structures could you reasonably hope to do on a system like this. We think that for a 40-50 atom structure it would take twenty-six hours for the structure determination. It certainly takes quite a bit longer to collect the data. So really, you are still limited by the amount of time it takes to collect the data.

Coppens: So the system has over-capacity for one crystallographic group.

Sparks: Yes, it has indeed.

Corfield: I think this system is not totally unreasonable, but what makes it reasonable is the availability of inexpensive hardware floating-point arithmetic units. We've had at Ohio State University for the past two or three years a system rather more sophisticated than this but that does not have hardware floating-point arithmetic. Presently we do all our least-squares and all our Fourier summations in-house, but once we get up to a couple of hundred variables, it would be worth our while to use a larger computer because of the limitations of the software floating-point arithmetic on our in-house machine.

Medrud: If this kind of approach is attractive to other crystallographers, there is another encouraging factor in the change in attitude of some of the minicomputer manufacturers. Our first contacts with them, with regard to our application, were met with disdain. The most recent contacts other people in our group have had with them indicate much more interest in systems development. They formerly wanted to hand you a computer and a bag of hardware for interfacing and say "go to it", but now they are willing to discuss a system comparable to yours.

Session III

What Are the Funding Agencies Doing, and What Are Their Plans for the Future?

Session Chairman: Allan Zalkin

Computing and Crystallography: The National Science Foundation and the National Academy of Sciences-National Research Council*

Peter G. Lykos

The National Science Foundation (NSF) makes grants to support innovative projects in the sciences, designed to create or discover new knowledge and, to a lesser extent, to teach and/or disseminate new knowledge. The Research Directorate, one of five comprising NSF, accounts for half the NSF budget. The total NSF budget is surprisingly modest, about the same as that of the Chicago Public School System or about half a billion dollars per year. The Chemistry Section in the NSF Research Directorate, accounts for some of the NSF support for chemistry. Other parts of the NSF such as the Offices of Computing Activities and Science Information Service account for a nontrivial portion.

Within the NSF, crystallography is not singled out programmatically and explicitly as such. At least three NSF Sections support crystallography, namely, Earth Sciences, Molecular Biology, and Chemistry. In Fiscal Year 71 (July 1, 1970 through June 30, 1971), the grants with crystallography as a principal component were as follows:

	No. of Grants	Total Grants	Computer Cost Included
Chemistry	42	$1 080 000	$143 000
Earth Sciences	17	480 000	32 000
Molecular Biology	23	780 000	98 000
Totals (prorated for 1 yr.)	82	$2 340 000	$273 000

This does not include crystallographic work done in the Material Sciences Research Laboratories established by the Advanced Research Projects Agency (ARPA) of the Department of Defense, for which responsibility is in the process of being transferred to NSF. There is increasing pressure on NSF to recognize that certain crystallographic techniques have become routine and an established part of science and technology, the conduct of which is no longer eligible for NSF support in and for and by itself.

The Office of Computing Activities, part of NSF's Directorate of National and International Programs, has three Sections, namely, Computer Science and Engineering (CSE), Computer Innovations in Education (CIE), and Computer Applications in Research (CAR). I came to the NSF last

---

* The opinions expressed here do not necessarily reflect the policies of the National Science Foundation or the National Research Council.

summer to design a fourth Section, Computer Impact on Society (CIS). While doing that, I am functioning as Program Director for Special Research Resources in CAR. In that capacity I shaped two programmatic thrusts, namely, hierarchical computing for laboratory automation, and computer networking to support research. The networking effort is being split off and expanded into a trial National Science (Computer) Network, described at the April 13, 1972 EDUCOM Conference on Networks and Higher Education. Who will administer that network? What computer/communication network technology will be used? When will significant funding be available? Answers to these and other important operational questions need to be known before the deeper problems of computer resource sharing via networking can be addressed.

The Office of Computing Activities was the vehicle through which the NSF subsidized the creation and expansion of campus computing centers. It initially had a primary focus on supporting research, but added an increasing, albeit modest,component of educational support during the life of the Institutional Computing Services (ICS) Program. The ICS Program was terminated abruptly and without warning about two years ago, leaving many university computing centers in serious financial trouble. The following table reveals the annual grant totals for the ICS Program, and for the line items in basic research grants for computation:

	FY '68	'69	'70	'71
No. Basic Research Grants	3917	4146	4041	4679
No. with Computational Support	870	1475	1544	1682
Total Computational Support in Basic Research Grants	$4.5 M	5.6	5.5	6.5
Total ICS Grants	$10.6 M	6.5	6.5	1.6

As grants are made for varying numbers of years, as ICS grants include purchase of equipment which may be operated for as long as five years, and as the computational support in research grants is as listed in the grant and not the audited total, it would be difficult to interpret these numbers on a year by year basis. However, certain qualitative conclusions are clear, namely, that total NSF expenditures for computational support of research (including ICS) has decreased steadily, perhaps by a factor of 2, since 1967, while the number of research grants containing computational support has increased steadily, by a factor of 2, since 1967.

The recent survey of computing activities at 2800 colleges and universities by John Hamblen, covering FY 70, reveals that the total expenditures for all computing in higher education was about $472 000 000

and that the data collected suggest $512 000 000 for FY 71. Thus of all expenditures for computing in higher education, NSF support accounts for about 2%.

Increasingly the true power of the computer, namely as an information processing machine, is coming to be utilized in support of scientific research. This is reflected in the NSF in the recent reorganization of the Office of Science Information Service (OSIS) which is increasingly concerned with machine-based systems of information storage, handling, and retrieval. Until recently, OSIS concerned itself primarily with scientific literature, and with references thereto. Now the support of projects dealing with storage, handling, and retrieval of scientific data has become a major concern. Indeed those computer applications considered as typical "number crunching" problems include data-base management and file handling as an important, if not as the most important, component. And a large increase in "data pressure", such as availability of the Census data, and NASA's ERTS-A (scheduled to go up in June 72) and ERTS-B (scheduled for June 73) earth-surveillance remote sensors producing 10^{11} bits of information per day, emphasize the need for considerable research in the hardware/software problems of the collection, reduction, storage, and retrieval of massive amounts of information on a scale hitherto not seriously contemplated.

Leaving the role of the National Science Foundation in computing in science, as this Symposium was initiated by the National Research Council's Committee on Computers in Chemistry, it is relevant to examine briefly how the National Academy of Sciences-National Research Council is addressing the subject. About the same time the National Research Council's Division of Chemistry and Chemical Technology established the Committee on Computers in Chemistry, the National Academy of Sciences formed jointly with the National Academy of Engineering a Computer Science and Engineering Board (CSEB). This Board has concerned itself with studying and making recommendations on the more global problems posed by the large enhancement in informational technology brought on by invention and proliferation of the information processing machine concomitantly with developments in on-line mass storage devices and ease of interface to our telecommunications systems. Congressman Jack Brooks of Texas has requested $100 000 000 (H.R. 13 200) to enable the National Bureau of Standards to expand its research and development of standards in all areas of informational technology. The proposed legislation provides for utilization of the CSEB as an advisory board.

On the other hand, the National Research Council in its Committee on Computers in Chemistry has also a discipline-oriented approach to the problem. This Committee was responsible for initiating in 1970 with support from NSF a study of Computational Resources for Theoretical Chemistry. A report of the first stage of this study was published a year ago. A more comprehensive study of the feasibility and desirability of a national laboratory for computation in chemistry is now in progress with

joint support of the Office of Computing Activities and the Chemistry Section of NSF.

A similar dual situation is developing within the National Science Foundation. Should the disciplines assume more of the responsibility for computer support in their fields through, say, the support of discipline-oriented regional or national facilities, or should a pluralistic approach be adopted where regional or national computer resources are developed, without regard to the particular disciplines or sub-disciplines served? Discipline-oriented centers of national scope are already in existence with NSF support. Among these are the National Center for Atmospheric Research, a facility providing massive computer and other large-scale specialized equipment support to university researchers as well as to its in-house staff, and the Computer Research Center for Economics and Management Science, developing under a five-year grant from NSF to the National Bureau for Economic Research, which is purchasing its needed computer services actually from an off-site supplier.

Two computer-communication network developments should be noted. Firstly, public institutions of higher learning in many states are organizing themselves on a state-wide basis for computer service and resource sharing. Secondly, commercial computer-communication networks of national scope already exist (Tym-net), or are well along (DATRAN, partial operation scheduled to begin in Fall 1973), or will shortly become available (DoD's ARPA Network already in existence to be transferred within the next two years to a commercial operation).

I hope this brief overview of selected NSF and NAS-NRC computer-in-science activities is informative and will aid the crystallographic community in determining what collective action, if any, it ought to take.

The National Institutes of Health and Computational Needs and Resources in Crystallography

Michael A. Oxman

The mission of the National Institutes of Health is to improve the health of the Nation through research, education, and the exchange of knowledge. Because of its mission orientation, there are no formal programs dedicated to further developing technology itself in the basic sciences. Instead, NIH provides financial support for research projects designed to increase our medical knowledge base or to solve specific health-relevant problems.

With respect to the field of X-ray crystallography, primary NIH support is through research projects on a problem by problem basis. In Fiscal Year 1972, for example, 207 projects that involve crystallography are being supported (Table 1). Of these, 61 have crystallography as a primary emphasis term with the majority being funded by the National Institute of General Medical Sciences and the National Institute of Arthritis and Metabolic Diseases (Table 2).

One exception to NIH's emphasis on research projects, which relates to the purpose of today's meeting, is the Biotechnology Resources Program of the Biotechnology Resources Branch (BRB), Division of Research Resources. The BRB establishes and supports regional resources that make available sophisticated technologies on a broad base to the biomedical research community. In addition to providing a service function, each resource is responsible for promoting strong core research and development in its technology, engaging in collaborative efforts between resource core scientists and members of the user community unsophisticated in the use of the technology, and providing an arena for training in that technology for both future technologists and biomedical researchers. Areas included in the BRB program are biomedical computing, mass spectrometry, nuclear magnetic resonance spectroscopy, and electron microscopy. The BRB is the NIH focus for such activities, particularly in the area of biomedical computing. Initially, the BRB program supported batch processing systems primarily (Figure 1). However, over the past few years, the trend had been to establish more general-purpose multiaccess systems and process-control systems in order to meet the needs of the maximum number of biomedical investigators in the most effective way.

A few applications have been received over the past few years to establish regional instrumentation centers for X-ray crystallographic studies. For various reasons, none were approved for funding. It appears that there is insufficient demand for the analysis of small, biomedically important molecules that could be performed on a routine service basis. In addition, using presently available technology, large molecules require such an enormous amount of time and effort that the number of analyses that could be performed per year, for example, would be quite small. Since, very few scientists could be served, support from NIH for such a

center appears to be unjustified at this time. Finally, the major bottleneck in the area of X-ray crystallography seems to be the availability of appropriate crystals of macromolecules and not the acquisition or processing of data.

None of the computer resources funded by the BRB provide any significant computational support to X-ray crystallographic studies except the centers at Columbia University and Princeton University. Both, however, are rather specialized in that they are oriented toward computer graphics and model building, not large-scale number-crunching operations. Since BRB support bridges many areas of science and technology that relate to health research, it would be outside the mission of the program to develop a center limited to one specific field of research. On the other hand, as part of their overall service responsibilities, a few centers supported by BRB currently provide limited computational support to crystallographers. Although other centers are not providing such services at present, many have the necessary capacities and capability for doing so. Certainly many general-purpose university systems also can accommodate the crystallographer's computational requirements.

In summary, it seems most prudent to take advantage of the tremendous computational capabilities that are now available. Many computation centers, especially those based on large computers, are underutilized to various degrees. Therefore, selective upgrading of existing systems and developing an appropriate communications network between crystallography laboratories and appropriate computer facilities could result in a national system adequate to meet the special computing needs of X-ray crystallography.

The BRB will continue its efforts to reduce the research overhead of NIH-supported investigators by supporting computers, and is at present gathering a knowledge base to determine what shared biomedical resources will best meet the needs of NIH's clientele.

Table 1 Grants Funded by PHS During FY 1972 to Support Research Involving Crystallography, by Dollars

PRIMARY EMPHASIS INDEX TERM*	NUMBER OF GRANTS	TOTAL FY 1972 DOLLARS
Crystallography		
Computer Analysis	6	$ 266,000
Computer Image Processing & Display	1	153,000
Computer Programming	4	116,000
Computer, Man-Computer Interaction	1	-0-
Computer Simulation	2	71,000
No Secondary Computer Term	28	801,000
X-Ray Structure Analysis	17	557,000
X-Ray Diffraction	2	148,000
	61	$2,112,000

INDEX TERM*, NOT PRIMARY	NUMBER OF GRANTS	TOTAL FY 1972 DOLLARS
Crystallography	35	$ 684,000
X-Ray Structure Analysis	23	1,113,000
X-Ray Diffraction	85	2,035,000
Neutron Diffraction	3	37,000
	146	$3,869,000

TOTAL NUMBER OF GRANTS = 207

TOTAL FY 1972 DOLLARS = $5,981,000

*The Division of Research Grants, NIH, maintains a data system on funded programs. Each grant application is coded according to keywords chosen to describe the project. Keywords that describe the basic research effort are listed as PRIMARY EMPHASIS INDEX TERM. Other keywords listed as INDEX TERM.

Table 2 Grants Funded by PHS During FY 1972 to Support Research Involving Crystallography, by Funding Unit

AWARDING ORGANIZATION		NUMBER OF GRANTS FUNDED
National Institutes of Health		
Allergy and Infectious Diseases		11
Arthritis and Metabolic Diseases		41
Cancer		16
Child Health and Human Development		1
Dental Research		17
Eye		3
General Medical Sciences		83
Heart and Lung		18
Neurological Diseases and Stroke		12
Research Resources		1
	TOTAL	203
Environmental Health Service		
Air Pollution Control		4
	TOTAL	4

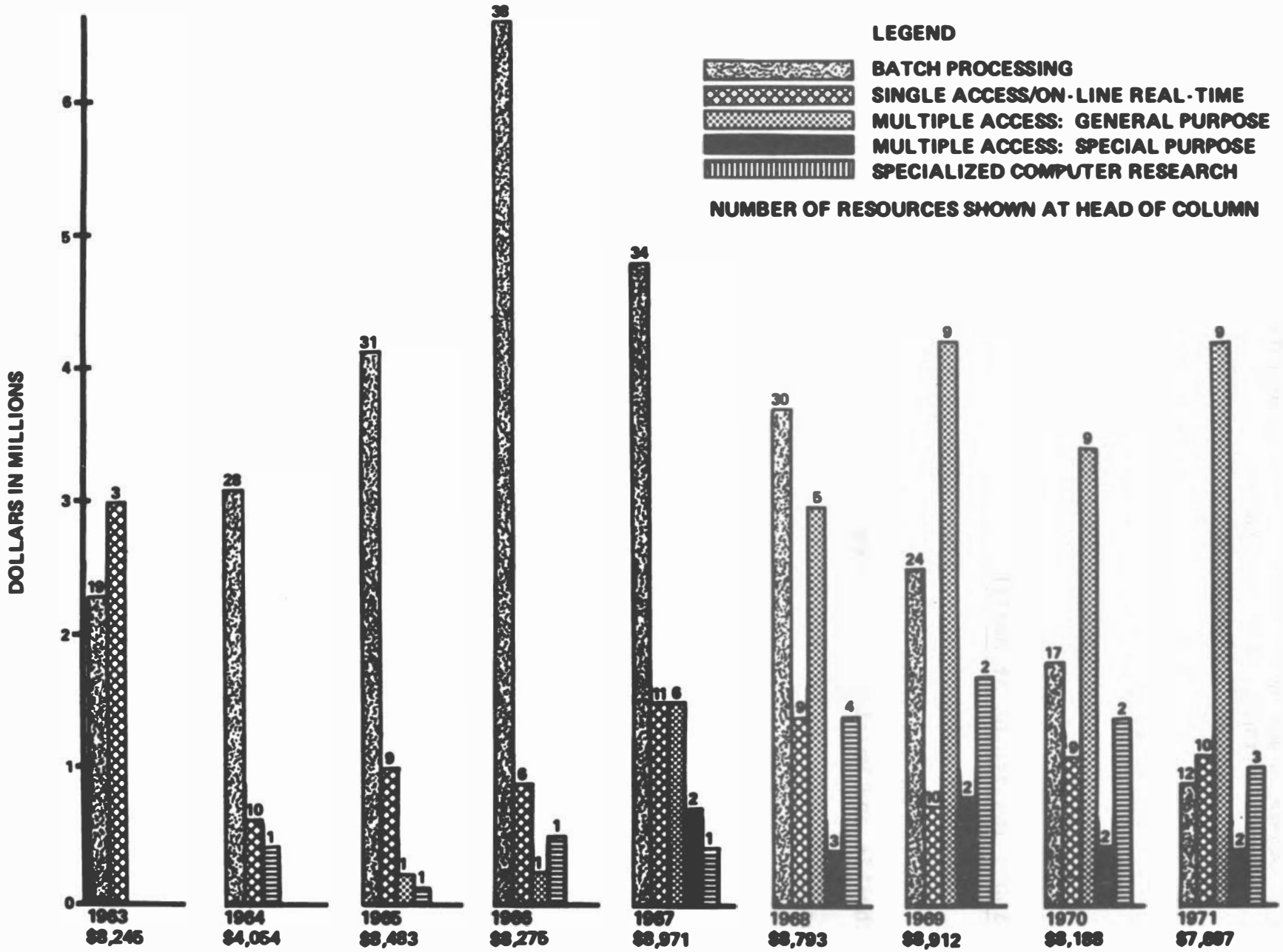

Figure 1 Biotechnology Computer Grants

Computer Support by The Atomic Energy Commission

James F. Wagner

Let me set the stage for my discussion by presenting a few general statistics on the automated data processing equipment in the AEC. As of June 30, 1971, there was a total of 954 "computers" in AEC's inventory. However, you have to be careful in analyzing this figure. Of the 954, over 750 are smaller computers used for the most part in on-line control and data collection. As a matter of fact, only 201 cost more than $200 000. This level of cost is significant to AEC because most of our management and control effort is spent on those above the $200 000 cost level.

AEC's budget for major computers has been about $35 - $40 million per fiscal year. If one looks at the increasing cost for computers, it becomes obvious that if our available dollar were to remain relatively constant the number of computers we could buy would be less. This is one reason for seeking new approaches to acquiring computers, such as the multiple-computer deferred-payment approach which I describe later.

As of June 30, 1971, our total investment in automated data processing equipment was about $327 million. Of this amount, $250 million or almost 75% was located at the following 13 major installations:

Location	Dollar Investment (in millions)
Defense Establishments	
Los Alamos Scientific Laboratory	$39
Lawrence Livermore Laboratory	37
Sandia Laboratories (Livermore and Albuquerque)	36
Bettis Atomic Power	16
Knolls Atomic Power Laboratory (Westinghouse)	15
Civil Establishments	
Argonne National Laboratory	21
Brookhaven National Laboratory	15
Oak Ridge National Laboratory	15
Lawrence Berkeley Laboratory	14
Stanford Linear Accelerator Center	9
Bendix Corp.	12
Oak Ridge Computer Technology Center	12
Savannah River Laboratory (DuPont)	8

How Much Money Does the AEC Spend in Support In Military-Related Research?

Table 1 gives a summary of AEC's operating costs for R&D. A total of $1374.6 million will be spent on R&D in FY 1973. Of this amount, 55.1% is for nondefense-related R&D, representing an increase in the nondefense-related R&D over FY 1971 of about 10% and over FY 1972 of 6.7%. While the defense-related support from FY 1971 to FY 1973 remains relatively constant, the nondefense-related support shows an increase. Total R&D plant construction in FY 1973 will be $1.6 million.

Table 1 U.S. Atomic Energy Commission - Summary of Operating R&D Costs (Based on FY 1973 Budget to Congress) (In millions)

	FY 1971		FY 1972[a]		FY 1973		Percent
	Amount	% of Total	Amount	% of Total	Amount	% of Total	Change over FY 1972
Total R&D	$1302.9	100.0	$1307.8	100.0	$1374.6	100.0	+5.1
Defense-Related	615.2	47.2	597.7	45.7	616.6	44.9	+3.2
Non-Defense	687.7	52.8	710.1	54.3	758.0	55.1	+6.7
Total Research	429.0	100.0	427.2	100.0	454.1	100.0	+6.3
Defense-Related	70.7	16.5	69.5	16.3	77.8	17.1	+11.9
Non-Defense	358.3	83.5	357.7	83.7	376.3	82.9	+5.2
Total Development	873.9	100.0	880.6	100.0	920.5	100.0	+4.5
Defense-Related	544.4	62.3	528.2	60.0	538.9	58.5	+2.0
Non-Defense	329.5	37.7	352.4	40.0	381.6	41.5	+8.3

a/ Includes proposed supplemental totalling $10.0 million.

Will the Planned Acquisition of New-Generation Computers at the National Laboratories Provide Excess Computing Power?

That can be made available to research workers in universities on a large-scale basis? If so, does the AEC have any specific plans for accomplishing this or will this be left up to the individual laboratory?

The installation of "new"-generation computers at the national

laboratories will be the result of what is known as the MCP (multiple computer procurement). Some time back (Fall 1970), one manufacturer proposed that the AEC buy several of the large computers, all to be installed within a six-month period but payment to be spread over four or five years. It was decided that the matter should be pursued further but on a competitive basis. A review of current needs of the various installations was made and it was determined that if the money were available, seven locations could justify the acquisition of a new large-scale computer. Contact was made with the General Services Administration, a request-for-proposal was prepared and issued, and responses from manufacturers were received. A special task force is currently (April 1972) evaluating the proposals. It is hoped that a selection will be made soon and contract negotiations can begin. The seven locations involved are:

1. Lawrence Livermore Laboratory (LLL)
2. Stanford Linear Accelerator Center (SLAC)
3. Los Alamos Scientific Laboratory (LASL)
4. Argonne National Laboratory (ANL)
5. Union Carbide Corp. - Oak Ridge Computer Technology Center (CTC)
6. E. I. du Pont de Nemours & Co., Inc. - Savannah River Laboratory (SR)
7. Brookhaven National Laboratory (BNL)

First of all, if there will be any substantial amount of time on these machines that will be available for use by others, we are not aware of it. The inclusion of these locations in the MCP was based on the need for each of them to increase their computing capability to meet their in-house programmatic requirements. Average monthly utilization figures at these locations for their major computers are as follows:

Location	Computers	Avg. Hours Per Month
LLL	CDC 7600 (2)	696
	CDC 6600 (3)	681
SLAC	IBM 360/91	560
LASL	CDC 7600	585
	CDC 7600	603
	CDC 6600	630
	CDC 6600	624
	CDC 6600	518
ANL	IBM 360/50/75	668
CTC	IBM 360/50/65	553
SR	IBM 360/65/ MP	476
BNL	CDC 660 (2)	657

As can be seen, these machines are used at near capacity. The acquisition of the new computers will provide needed additional capacity. In some cases, the new machine will replace some existing equipment and in others the new machine will supplement existing capability. If it turned out that for a short period of time in the beginning there was some available capacity, we would leave it up to each of the facilities to make the time available to others. There would be no central direction from AEC - at least from the controller's office.

How Much Outside Computing is Presently Being Done at the AEC Large Computer Installations?

In discussing the "outside computing presently being done at the AEC large computer installations," there are two ways of looking at the question. One is to what extent are we (AEC) using computers located outside the AEC, and the second is how much computing are we providing to other Government users. It must be understood that it is not AEC's policy to provide computer time to non-Government users or for non-Government use. We can provide some dollar figures on the outside time purchase and hourly figures on time provided to other Government users, but not without qualifying them.

The number of hours purchased by AEC installations is not practical to provide. Much of this time is for time-sharing and the "hours" are not consistent. Dollars, however, can be provided. These dollars are for time acquired from "commercial" sources. "Commercial" does not include time acquired by universities and other contractors where the computer is used as a part of the performance of some other task. For the six-month period ended June 30, 1971, the following amounts were spent for outside computer services by the various installations:

Cost of Computer Time Acquired from Commercial Sources for the Six-Month Period Ended June 30, 1971

Location	Amount
Sandia, Albuquerque	$36 686
Sandia, Livermore	11 639
LASL*	45 983
BNL	937
ANL*	26 294
NAL	33 407
NYU	None
ORNL	94 890
WADCO	41 925
LBL	1452
LLL*	5428
SLAC*	9679
OR-CTC*	15 248

SR *	1673
Others	354 604
Total	$679 935

*Installations included in the Multiple Computer Procurement

And the following figures show the amount of time (in hours) provided to other Government users for the six-month period ended June 30, 1971:

Location	Equipment	Number of hours
Lovelace	B5500	113
LASL	6600 1401 7600	1205
Sandia, AL	3600	14
BNL	6600	297
INC	360/75	133
NYU	6600	1005
Carbide, CTC	7090 360/50-65	15
Carbide, ORNL	360/75-91	770
LBL	(2)6600	964
Total		4516

Economics and Efficiencies of Large Regional Computer Centers

The controller's office has encouraged the establishment of "regional" centers or, more accurately, the centralization of computer capability to serve several users. This effort has been acknowledged by members of Congress. Evidence of locations where the type of operation exists in AEC is the IBM 360/75 at the National Reactor Testing Station, Idaho Falls, the Univac 1108 (CSC) and IBM 7090 at the Hanford Laboratory, the CDC 6400 at the Nevada Test Site, the Computer Technology Center at Oak Ridge National Laboratory, the computing center serving Argonne National Laboratory and the National Accelerator Laboratory, etc. All of these "centralized" installations exist as the result of a study made to determine the best way to meet the data-processing needs of a given number of users.

Generally speaking, it is probably more economical and more efficient to have one or more larger computers than it is to have several smaller ones. You get a greater capacity and it costs you proportionally less. Some people say that for an increase of two in dollars you can get up to eight in increased capability. There are the advantages of

larger memory, faster memory, and larger storage. The larger computers have generally better and more efficient operating systems.

The controller's office encourages centralized computer facilities for basically two reasons. Overall costs should be less and it is easier to justify the need for capital funds for a computer if several users are to be served.

DISCUSSION

Lowrey: Dr. Lykos was talking about the establishment of a computer network, and obviously you're going to need people to use it. Will a certain amount of funds, or percentage of funds, be available essentially to try using this kind of computer network? Will one be able to justify computer support for the purpose of simply utilizing and experimenting with the system you're setting up?

Lykos: By way of background to your question, the National Science Foundation for something like a dozen years gave money to universities to help build up the university computation centers for research and education. That was the computer facilities program (ICS). When abruptly terminated, its annual budget had decreased to somewhat over 6 million dollars a year. That activity, primarily, became the Computer Applications and Research Section (CAR) of the Office of Computing Activities. Because the user community hadn't really assessed its use of the computer for research, and the costs associated with it, there was really no case that could be made for rechanneling those funds in specific ways through the disciplines. Accordingly the CAR budget is substantially smaller than the sum of the budgets of the programs out of which it was shaped.

Perhaps I didn't emphasize sufficiently that the big problem is not the adequacy of computer network technology. It's very clear through the ARPA Network, for example, that the technical workability of computer networking is being demonstrated. The main problem facing us is the finding of specific applications that reveal how research can be enhanced through networking. To what extent this can be done

within the National Science Foundation's new program is somewhat up in the air at the moment. This new thrust requested impact money similar to what was used last year for the demonstration of computer assisted instruction. As far as I know, the requested impact money for this coming fiscal year is not forthcoming. Conceivably in the succeeding years laid out for this project, it will be there.

Certainly, some CAR support will have to be provided for the purpose of demonstrating the use of computer networks, and that has already been done in the current fiscal year, in three-level hierarchical computing for laboratory automation, for example, and in certain experiments in computer-network-assisted research. But this cannot be done in an extensive way by the Office of Computing Activities alone in the National Science Foundation. Within the NSF, the Office of Computing Activities is a channel between the world of computers and the other disciplines through its Computer Applications in Research Section. Unless some other parts of the National Science Foundation (in particular the Research Directorate) become alerted to the computing needs of the scientific community, and unless they reexamine their priorities along this line and act accordingly, then the experiment may very well fail. So, computer-using researchers, the ball is in your court.

Williams: Mr. Wagner, I wonder if you could communicate to us the progress being made on the multiple computer procurement with respect to Argonne and NAL?

Wagner: I'm not begging off your question, but I'm not privileged with much information. I know just about as little as you do about it although I was in on it at the beginning. But right now the AEC is in the process of reviewing the proposals from the manufacturers on the basis of the request for proposals that went out which includes the stipulation on the non-conversion cost consideration. By May 1st a selection will be made, Argonne's included with the other six, and I cannot tell you much more than that. One thing that I should mention. It's essential that we do all we can in that multiple procurement to get competition. Although we're accused of being bureaucrats, people don't realize the number of pressures we have. For instance, some vendor will run to Congress claiming that we're not considering his tape units when we buy an initial IBM system. Well, that's a great thing to say but how do you do it? They go to Congress and they go to GAO, and we get complaints right and left. So in the multiple procurement we do our level best to assure competition. As far as the status of the procurement, May 1st is the date for selection and June 15th for the negotiation of contracts. The first delivery, I think, would probably be sometime in the fall, but not at Argonne. The first delivery would probably be at Savannah River.

Donnay: If I understood you properly, you told us that a manufacturer who had six machines for sale went to you and made this offer. Is it really the manufacturer who should ask for a computer? Here we have seven outstanding research centers and, from what you say, they didn't ask for the computer--it was the manufacturer who figured out they needed one. Is that the way it should go?

Wagner: I didn't cover this point but I said that at all seven laboratories we reviewed their needs and saw that they needed computers. But we only get 35-40 million dollars a year to buy them. With six or seven laboratories in AEC and on a 35-40 million dollar level a year, we can't buy them all a major computer. What the manufacturer did was say, "Okay, we know you need them but you can't get them because of the budget process. We'll show you how you can." Now, for instance, the program divisions in the Atomic Energy Commission in 1973 asked for 121 million dollars for what they needed in the computer area. We know we can only get between 30 and 40 million dollars. We have never had a computer item cut out of the AEC budget by Congress because we're realistic when we go in with the request. But we know we're not getting what we need. All the laboratories had requested them in the budgets but they didn't succeed in surviving the budget cut. When the manufacturer came in and gave us this opportunity, we jumped at it. We already knew we needed them; the laboratories had told us that. The manufacturer just gave us a means of getting them quicker.

Freer: Do you anticipate that the 7600 at Lawrence Berkeley Laboratory will be going on the ARPA Network soon?

Wagner: I don't know anything about it. There is an internal AEC communications network. To be honest with you, the first time I heard about the ARPA Network was today.

Sayre: A few days ago I received a letter stating that the National Research Council, Division of Chemistry and Chemical Technology, is sponsoring a study of the feasibility and desirability of a National Laboratory for Computation in Chemistry. The study is divided into 7 sections, of which one concerns chemical structures. Has this study come to the attention of the speakers, and if so can they tell us a little more about it?

Lykos: There is a feasibility study under way to look at the question of a national laboratory for theoretical chemistry. That's a sequel to a two-day conference that was held in Bethesda in May 1970, where the general question of computational support for theoretical chemistry was considered. Even though that was a short conference of two days duration, as a result there emerged a clear sense of direction that this question merited an in-depth study. The feasibility study

was put in the form of a proposal from the National Academy of Sciences to the National Science Foundation, was funded, and is in fact under way.

The feasibility study has been made more complicated, however, because although it was conceived by chemists, in terms of the problems that chemists are facing today, the Academy's Committee on Science and Public Policy in reviewing the report of the Bethesda conference charged the follow-up study with all the problems of scientific computing in research in higher education. Those same questions that were posed in Dr. Harvey Brooks's covering letter, by the way, were also sent to the NSF Office of Computing Activities for reaction, and the OCA response is on record.

So, the feasibility study is under way with the awareness that some of the questions to be addressed transcend chemistry. There seems to be a general feeling that discipline-oriented centers ought to come into being; accordingly this particular study will be followed with interest by a large audience.

The feasibility study group is being led by Professor Kenneth B. Wiberg, a physical-organic chemist at Yale University and a member of the National Academy of Sciences. The study group involves theoretical chemists, experimentalists, computer scientists, and representatives from academia, government and industry. There will be a first general meeting of the study group on May 5th and 6th to get a first cut at the position papers now in preparation. The final report should be in hand by October.

Sayre: The panel on structural studies is due to meet, I understand, on April 24. It would seem appropriate for any conclusions we may arrive at today to be reported to that meeting.

Hamilton: I am on that Subpanel, Dave, and my intention is to present a full report of this meeting at that time.

Dewar: I must say I have a reaction to the report from the AEC. I find a severe discrepancy between claims to competence in the procurement of large computers and ridiculous claims to Congress of 99% utilization. I'd like to ask if AEC, at least internally, has more accurate utilization figures, because this type of estimate makes no sense and is clearly unrealistic.

Wagner: The Controller of AEC, Mr. John P. Abbadessa, feels that utilization is the key to what we get. What we do about inflated utilization figures, I do not know. The way the utilization is reported to us is determined by requirements set by the Office of Management and Budget. If you're talking about the figures I quoted from the test-

imony to Congress by Livermore, I am not going to argue with you about what they are running on the machine, because I am not in that kind of capacity. If the kinds of things run on the machines should not be run, then somebody's going to have to say that, and I am not about to, because I cannot go to Brookhaven or Oak Ridge or Livermore and look at what they're running and tell them they shouldn't be running it, if that's what you're telling me.

Dewar: My concern is that if AEC is in a position where it feels it has to say there is a hundred percent utilization, it may not be in a position to know for itself whether there is time available for possible distribution elsewhere.

Wagner: I do not think places like Livermore and Sandia and Los Alamos were pretending that there's a hundred percent utilization, because it's a 24-hour day, 7-day week operation. The only other thing you can say is that some usage is redundant, and that's what I'm telling you I'm not going to argue about. But they are using 24-hours a day, 7 days a week, and those are the ones that were quoted, Los Alamos and Livermore.

Larson: We do have at Los Alamos what we'll call a zero priority situation, where if there is free time on the machine a job will be run free of charge. Approximately the first of February I placed a one-hour job on the shelf, and it has not been run yet. Does that answer what you're asking about free time?

Williams: We have the same thing at Argonne, zero priority.

Session IV

Computing Centers and Networks

Session Chairman: S.C. Abrahams

Some Experiences with Crystallographic Systems

James M. Stewart

The word "system" especially as used in the context of computers is a very broad one. The factors to be "systematized" must be defined in each instance. Since the introduction of computers, most crystallographers have made an effort to systematize their computing usage and from this effort there have been written a great number of useful programs. Some of these programs are used as a collection with common mass-storage definition, common core-storage rules, common data-input conventions, common output conventions and common documentation formalism. These are the characteristics of an "operating system" as usually supplied with any computer from a PDP-8 to an ILLIAC IV. The complexity of any given system is dictated by the complexity of the computer and the user demands to be allowed on this computer.

In 1961 at the University of Washington, D. High, L. H. Jensen, E. C. Lingafelter, B. W. Brown, and others, including myself, reviewed the advent of the IBM 709 and the loss of the IBM 650 and considered the possibility of producing a collection of programs that would have the following features in common:

1. Programs that produce accurate and rapid diffraction analysis.
2. Stylized input formats for characteristic crystallographic data and operations.
3. Careful and detailed documentation.
4. Mass storage files defined.
5. Independence of the local computer operating system, and compiler.
6. Space group and setting universality.
7. Provision for large range of data set size, independent of high-speed memory (20 000-word minimum for data).
8. Highest efficiency possible consistent with the preceding criteria.

The current "X-ray" system then is a result of these design criteria. The authors and implementors of the system are now given as an appendix in the documentation for the system (see p. 95 at the end of this paper). Over the years many authors have contributed codes to the system and in each case we have endeavored to bring these codes into conformity with our particular criteria. Also during this time many others have written other systems with different or similar criteria and features. Most of these efforts have been rewarded with the production of many interesting crystal structure analyses.

For purposes of this discussion, I will confine my attention to the criteria that have been most frustrated by the immaturity of the

computer field, and to the advantages and disadvantages of a "system" approach.

The greatest difficulties have been and are being encountered in our effort to be computer-, operating system-, and compiler-independent. We have tried to introduce a concept of using a subset of FORTRAN which we call Pidgin FORTRAN in a partially successful attempt to achieve this objective. I must say, however, that one tends to become more and more paranoid with each successive FORTRAN compiler that is released. I will not dwell here on all the incredible examples I have seen but will simply give three, in order to give the flavor of the problem.

1. One manufacturer makes the presence of more than "N" comments in succession a fatal compiler error (where N is a number small at the pleasure of the compiler writers).
2. One manufacturer has one compiler in which a RETURN statement is required and another in which its presence is a fatal error.
3. One manufacturer has made the statement

```
I = J
```

cause the movement of only one half a word, while

```
A = B
```

always causes "normalization" of B before it is stored in A.

Consider the movement of alphabetical or "packed" information. The list is, of course, much longer.

There are also frustrations concerning word size, actual hardware configuration, etc. But these have turned out to be minor compared to the problems of operating-system software. It must be emphasized that this problem for us has been reduced to the non-equivalence of the meaning of identical looking FORTRAN statements. We believe we have succeeded in achieving reasonable interchangeability in spite of these problems, so that we are now able to make blocked magnetic tapes on the UNIVAC 1108 capable of being read, compiled, loaded, libraried, and executed to give the same crystallographic results on a variety of other computers. We also believe that given one of the others we could carry out the same operation.

When one speaks of computing costs, the development costs should probably be quoted separately from the structure-solution costs with checked-out programs. The development costs for the X-ray system have been very high. And much of the cost has been in trying to beat the problem of the differences in the various FORTRANS. This problem may become worse because of the rapidly changing computer technology. With

terminal operation to remote terminals this feature of mutual compatibility may lead to more and more frustrations for crystallographers.

Now to advantages and disadvantages of a "system" approach to crystallographic computing:

The advantages are mainly those of convenience of use, interchangeability from computer to computer, slow but steady improvement in reliability and function, and the many active users aiding in the check out of the codes.

The disadvantages are mainly those of inflexibility of modification, "black box" effect on users, and some sacrifice in speed for generality (although for many potential users it will usually take a long time to recover the development time due to this sacrifice).

Another problem area is that, despite our best efforts to understand operating systems of the various manufacturers, there is often an initial delay in implementation due to the size of the system as it is presently (1971) constituted.

In summary, I believe that the system approach has something to recommend it for crystallographers who do not have a large enough group to develop and maintain their own set of programs. They are well advised to consider one of the essentially checked out systems in use (1) and adapt it to their computer or use it by a remote terminal. This is especially true for those groups whose main interest is in routine structure analysis. On the other hand for groups with good access to funding and computers and a few in-house programming crystallographers, they too may wish to begin to collect their own libraries. If they are not keen to distribute their "system" they may be greatly aided by the in-house operating-system software. This is, in my opinion, an extravagent way to do crystallographic computing. I believe that groups like this will find, as we have, that when the computer system changes they will have to invest a large amount of time and effort into the conversion of their libraries.

I would prefer to see the efforts of this group of talented men directed to the production of better standard codes or new methods of crystallographic computing in an interchangeable form. The funding agencies might notice that much of the programming effort of crystallographers in the past has gone into development and check out of manufacturers software. This item in most cases is charged to "crystallographic" computing rather than computer software development.

---

(1) On this continent, Busing, Hamilton, Johnson, Larson, Ahmed, Ibers, Marsh, Sparks, and many others have collections of programs or systems in operation or potential operation.

In the development of the X-ray system, we must recognize direct and indirect support by many different sources, including NSF, NASA, ARPA, the Army, the Air Force, NIH, AEC, the Computer Science Center of the University of Maryland, the Research Computer Center of the University of Washington, the U.S. Geological Survey, the National Bureau of Standards, and the Science Research Council of the United Kingdom. These sources of support have now become less accessible in a direct way because of the prevailing economic conditions. It is still my hope to be able to maintain, improve, and distribute the X-ray system by whatever means we can find.

Appendix

Contributors to the X-Ray System

The X-ray system has been developed over a number of years with contributions from a large number of people. This effort has fallen into three main categories -

1. System Editing - i.e. the writing of the nucleus, maintenance of the programs, the write-up, general organization, and system philosophy decisions

2. Program writing - without which there would be no need for a system

3. System implementation - i.e. the responsibility for providing information for making the system run on specific machines and for checkout of new system releases.

Obviously, some program authors have actively contributed in other respects and due acknowledgement of their authorship is given within the program descriptions in Section 1 of this write-up.

The affiliation given for each contributor is that appropriate at the time the contribution was made and should not necessarily be considered as current.

System Editors

Baldwin Dr. J.C.	Atlas Computer Lab., U.K.
Chastain Dr. R.V.	Univ. of Washington, Seattle
High Dr. D.F.	Univ. of Washington, Seattle
Kruger Dr. G.J.	CSIR, Pretoria, S. Afr.
Kundell Dr. F.A.	Univ. of Maryland
Stewart Prof. J.M.	Univ. of Maryland

Program Authors

Ammon Prof H.	Univ. of Washington, Seattle
Alden Dr R.A.	Univ. of Washington, Seattle
Boonstra Dr E.G.	Univ. of Orange Free State
Brown Dr B.W.	Portland State College
Braun Dr R.L.	Univ. of Washington, Seattle
Busing Dr W.R.	Oak Ridge National Laboratory
De Camp Dr W.H.	Univ. of Maryland
Dickinson Mr C.W.	U.S. Naval Ordnance Lab.
Dayhoff Dr Margaret	Natl. Biomedical Res. Foundation Inc.
Freer Dr S.T.	Univ. of Washington, Seattle
Hall Dr S.	Mineral Sci. Div., E.M.R., Ottawa
Holden Dr J.R.	U.S. Naval Ordnance Lab.
Jarski Mrs Mary A.	Univ. of Washington, Seattle
Jensen Prof L.	Univ. of Washington, Seattle
Keefe Dr W.	Medical Coll. of Virginia
Kerr Dr Ann	Cambridge Univ., England
Kraut Prof J.	Univ. of California, La Jolla
Lingafelter Prof E.	Univ. of Washington, Seattle
Levy Dr H.A.	Oak Ridge National Laboratory
Mauer Mr F.A.	National Bureau of Standards
Mighell Mr A.	National Bureau of Standards
Martin Dr K.O.	Oak Ridge National Laboratory
Plastas Mrs Linda	Univ. of Maryland
Santoro Dr A.	National Bureau of Standards
Schneider Dr. M.L.	Univ. of Maryland
Takeda Dr H.	Johns Hopkins Univ.
Zocchi Dr M.	National Bureau of Standards

System Implementors

Appleman Dr D.	U.S. Geological Survey	IBM /360 series
Kirchner Dr R.	Univ. of Washington, Seattle	CDC 6600
Lenhert Prof P.G.	Vanderbilt Univ.	XDS SIGMA 7
Morosin Dr B.	Sandia Corporation	CDC 6600
Protherough Mr M.	I.C.L./ Univ. of Surrey	ICL 1900 series
Snyder Dr R.	M.I.T.	IBM /360 series
Thomas Mrs Judith M.	Atlas Computer Lab., U.K.	Atlas
Watenpaugh Dr K.	Univ. of Washington, Seattle	CDC 6600
Wolten Dr G.	Aerospace Corp.	CDC 6600

Valuable technical assistance has been given by Miss Jean Willis and Miss Stefanie Nucci, both of the University of Maryland.

DISCUSSION

Dewar: The three examples of the three difficulties you gave with FORTRAN should all have been covered by the ANSI standard. Is it in fact a case that if the ANSI standard properly adhered to by manufacturers that almost all of your problems would go away?

Stewart: This is the problem that I aluded to, namely, when you write to a set of specifications for a code, and a programmer takes this up without any real care for what actually happens, he writes exactly to the specification so that the contract is fulfilled. He delivers the contract on time and now you've got this "code." I have seem this many times around the Washington area and in the university where they send out a contract for a program. The programmer meets the letter of ANSI specifications in every instance. But because there is nothing in the ANSI standards about the word structure, or the idea of what the meaning of an i or a j is, trouble develops. They have defined it to be a 24-bit binary "thing" for that particular machine and so therefore they meet the spirit of the ANSI specification. When challenged the manufacturer stands on the fact that he has completely met the ANSI specifications. We just happen to be so stupid that our Pidgin FORTRAN wasn't pidgin enough to recognize that anyone would ever make a non-equivalence between "words", in a machine.

Ibers: Some of us, particularly the younger members of the crystallographic community, forget the disproportionate contribution that AEC-sponsored people have made to our various program libraries. We have the various Oak Ridge programs of Busing, Levy, Johnson and others; we have the Fourier program of Zalkin; we have a variety of programs of Hamilton. These tend to be parts of many program libraries, albeit in highly modified form. The AEC has already made invaluable contributions to crystallography.

Thomas: This may be a little facetious but I think it should be said that one of the problems is that some manufacturers are crooks.

Stewart: Yes I did. I don't care to repeat it.

Lykos: Regarding computer program packages, their standardization, certification, and dissemination, three NSF-supported activities are in progress that may not be generally known and may be of interest here. First, the Quantum Chemistry Program Exchange which flourished at Indiana University with support from AFOSR for many years serving chemists on an international scale. Its range has expanded so that a more descriptive name might be the Computational Chemistry Program Exchange. Now with NSF support, it is in transition toward becoming self-supporting and has been approached to extend its services to the

world of crystallography.

Second, a software certification thrust led by Professor L. Fosdick, Chairman, Department of Computer Science, University of Colorado, Boulder. He organized and conducted a small conference at Boulder in order to obtain better coordination of several NSF-supported projects in software development and certification. The need for coordinated efforts along those lines became even more evident as a consequence of that Spring 1972 Boulder Conference, and a structured approach is evolving. He could be approached regarding criteria and procedures for software certification.

Third, Professor Frank Harris who is well established as a quantum chemist, an applied mathematician, and a sophisticated user of large-scale scientific computers, is developing a set of computer programs for users of quantum chemical techniques. He is capitalizing on the fact that the University of Utah has a node in brilliantly conceived ARPA Network by testing the machine independence, reliability, and accuracy of the programs on physically and logically different computer systems via the ARPA Network. Additionally he will recruit a small number of users to access the University of Utah UNIVAC 1108 via terminal in order to test the viability of remote terminal - local "hot phone" augmented access as well.

Thus there are specific models and techniques available for computer-software resource sharing should the crystallographers wish to explore them.

Interactive Graphics and Remote Computing

Edgar F. Meyer

At a time when computer technology has been advanced to the point where selected computers can communicate at a rate of a million bits per second (the ARPA Network), we who are concerned with crystallographic computing could well consider the conditions under which remote computing would be advantageous or necessary.

The case of a laboratory without access to a local, large computer is an evident candidate for remote computing. An equally good case can be made for operations such as those of Dr. Ibers at Northwestern University, where he needs the large memory available in one of the latest generation computers to refine 400-500 variables simultaneously. Perhaps Professor Jensen at the University of Washington would care to comment on the advantages of having a computer currently 10 to 30 times faster than his own to reduce the 10 hours per cycle needed for one refinement operation on his protein structure.

The above three cases may generate some discussion but I would like to turn attention especially to the broad, intermediate area of routine crystallographic computing in a laboratory with access to several local computers, from the 4000-word minicomputer running the diffractometer to the local computing center. Campus politics aside, what arguments can be raised for remote computing on a latest generation computer with a special support facility for crystallographic computing? And what pitfalls can be forseen?

First, let me clarify the usage of "remote" by indicating that some type of terminal to the distant computer is implied, rather than mailing off your deck to a good friend (with a large computing budget) at X University. The quality and cost of service of the telephone system in this country is a topic of much discussion in the popular computer magazines these days, with Ma Bell on the defense. You can rent an ASR 33 Teletype with an acoustic coupler for $65/month (maintenance included). This produces printed text at a rate of 10 characters per second, which is suitable for printing R factors, slow for coordinates, and unthinkable for Structure Factor tables. Telephone rates at night are about 20-25¢ a minute.

Several types of terminals are available for roughly similar prices that will operate at 30 characters per second. These include both hard copy and modified television master displays for alphanumeric text.

I propose that a case can be made for a considerably more efficient crystallographic computing system from dial-up terminals than is currently provided by the average campus computing center. People who buy time-sharing services may buy better service, but my first supporting

argument is that instead of all the effort required for each group to develop and maintain its own computer library, this library could be maintained on a regional or national basis, provided I: Specialized software support.

The trend to larger computers with multi-user capabilities may make it hard for the computing center to replace their 360 or 6600 with a 7600, ASC or 370, both for reasons of funding and limited demand. Yet these larger computers can handle Professor Ibers' 500 variable matrix and reduce Professor Jensen's time per refinement cycle by a factor of 30 (10 hours to 20 minutes), provided II: Greatly increased capacity.

One of the useful results to come out of a conversational time-sharing system like Project MAC has been the ability of users to borrow routines and, in general, to interact with each other through the computer. Thus, III: the "Critical Mass" required for smaller laboratories to become viable could be reduced through increased interaction of routines and crystallographers.

Finally, a subject of some interest to me involves IV: Storage, retrieval, and three-dimensional display of structural information. Mrs. Kennard in Cambridge is continually adding to a library of over 4000 structures taken from the scientific literature. The "Protein Data Bank" at Brookhaven National Laboratory is gathering coordinates of macromolecules at various levels of resolution as they are submitted. The first set of protein structure factors has been submitted. A method of referencing protein Fourier maps needs to be devised. I feel strongly that a low cost, three-dimensional graphics display with interactive capability would be an ideal component to a remote terminal, especially.

Now I shall raise some counter-questions:

I. How can one adequately gesticulate over the phone when the software has been changed and one's program no longer works?

II. a. Will disc storage be available for each user?
b. How will large listings be handled?
c. How can listings be obtained rapidly?
d. What will the response be at peak times?
e. Is an interactive, conversational system practical?
f. What is a workable upper limit to the number of crystallographers and groups using a given system?
g. What reasonable usage limits can be assigned to groups: (1) solving and refining structures, and (2) developing new techniques?

III. a. Will sufficient safeguards be provided to protect privileged files?
b. Since card decks will not be the usual form of data and programs, will a long-term file retrieval system be available?
c. Currently, many groups doing crystallography on a low-keyed level have been able to find support locally. How accessible will support

be for the "gentleman" crystallographer?

d. What provision will be made for marginal cases; that is, who qualifies as a crystallographic user?

e. Having recently experienced a funding upheaval, wouldn't it be safer to hammer along locally than to face a potential discontinuity in the support curve?

IV. a. Where do I sign up for my own display terminal?

b. Who will fix it for me when it goes down?

c. What terminal configuration has the optimum capability/cost ratio?

d. How flexible can the configuration be for the requirements of individual laboratories?

e. How will the transfer of huge (megaword) files be handled?

The purpose of these remarks is to point out some of the technical possibilities available today. Beyond this, some of the pitfalls mentioned can serve as a starter for the creative pessimist.

I conclude with a comment on a criticism I have heard in Europe that American crystallographers are too many and too disperse: better a few good groups than one everywhere. I suggest firstly that even synthetic organic chemists are doing crystallographic analyses; its use is not elitist, but its advancement may be. Secondly, of the million-plus known organic compounds, the 4000-plus in Mrs. Kennard's library leave some significant work to be done. And finally, with a link to an available computing service (plus provision for diffractometer data), practically every chemistry department could reference and contribute to the growing library of structural data. Then national meetings could be reduced to the afternoon outing and banquet; we could all keep in "touch" over our terminals.

DISCUSSION

Xuong (referring to a stereoscopic display of structures demonstrated by Prof. Meyer on a cathode-ray terminal): What is the display in color for and how much would it cost to buy a duplicate?

Meyer: The display is colored for the reason that it gives you the

three-dimensional effect in two colors and you get the stereo separation by viewing through a colored screen for each eye. You could also use the color for getting color tonality in the molecule. The second question of cost might better be answered by Dr. Sparks.

Sparks: $37 500.

Anonymous: One could use the Tektronix or some other storage scope. There are a lot of new displays coming on the market.

Meyer: That point is well taken. The technology is moving along quite rapidly. The devices we're using now might well be out of date in a few years, but the point is that in my laboratory I have to do with what I have right now, and what I have shown you is a usable device. There is, for example, under consideration quite a reduction in cost and we hope to take advantage of this. The price will ultimately drop. One does not have to use the disc to drive the display, for example. Among questions that of course have to be held uppermost are the quality, the utility, and even the eye strain. You get fatigue if you sit there in front of the tube all day, but the fact is that the system works.

Some Thoughts on the Role of Hierarchical Computing and National Networks in Protein Crystallography

Steven T. Freer and Nguyen Huu Xuong

Protein crystallographers are caught in the bind between reduced or stagnant computing budgets and ever-increasing computational needs. Indeed, many of us find our research slowed and restricted by lack of computing funds. This situation is somewhat paradoxical because many computers throughout the nation are now utilized at only a fraction of their capacity. It is apparent that our computational needs cannot be fully satisfied without a new approach that will enable us to utilize more effectively existing computational resources. We believe that hierarchical computing is such an approach.

A hierarchical computing system is a network of special-purpose computers linked together so that several ascending levels of interconnected hardware and software can be effectively shared among many users. At the lowest level of a typical hierarchy are several minimal minicomputers each of which is dedicated to controlling a specific experiment. The minicomputers are linked to a traffic control computer that provides sophisticated input/output and large bulk storage. A large amount of money is saved by sharing I/O and bulk storage equipment, which can be extremely expensive and is usually much under-used. At the next level of the hierarchy, the traffic control computer is linked to a more powerful computer that can satisfy the computational requirements of the users. Ultimately, the traffic controller would be connected to a national network that could provide instant access to any desired computer anywhere in the United States, thereby allowing the user to select the computer best suited for each facet of his research project. At UCSD, hierarchical computing is playing an increasingly vital role in protein-structure determination. The reason for this is that all facets of protein crystallography, from data collection through display of the solved structure, require extensive use of different types of computers: dedicated minicomputers for control of data-acquisition systems, large and powerful number crunchers for the calculations associated with structure determination and refinement, and special computers for dynamic display and manipulation of molecular models. The protein crystallographic computing system that we are trying to develop is shown in Figure 1. The lowest level of the hierarchy will consist of three data-collecting devices: an automatic diffractometer, precession cameras in conjunction with an automatic film scanner, and a multireflection diffractometer (Xuong and Vernon, 1972), plus an interactive model-building and coordinate-measuring device. Each of these instruments will be controlled by a minimal minicomputer interfaced to an IBM 1800 computer, using standard CAMAC modules (EURATOM, 1969). Peripheral equipment associated with the IBM 1800 traffic controller includes high-speed disc drives with thirty megabytes of storage, a line

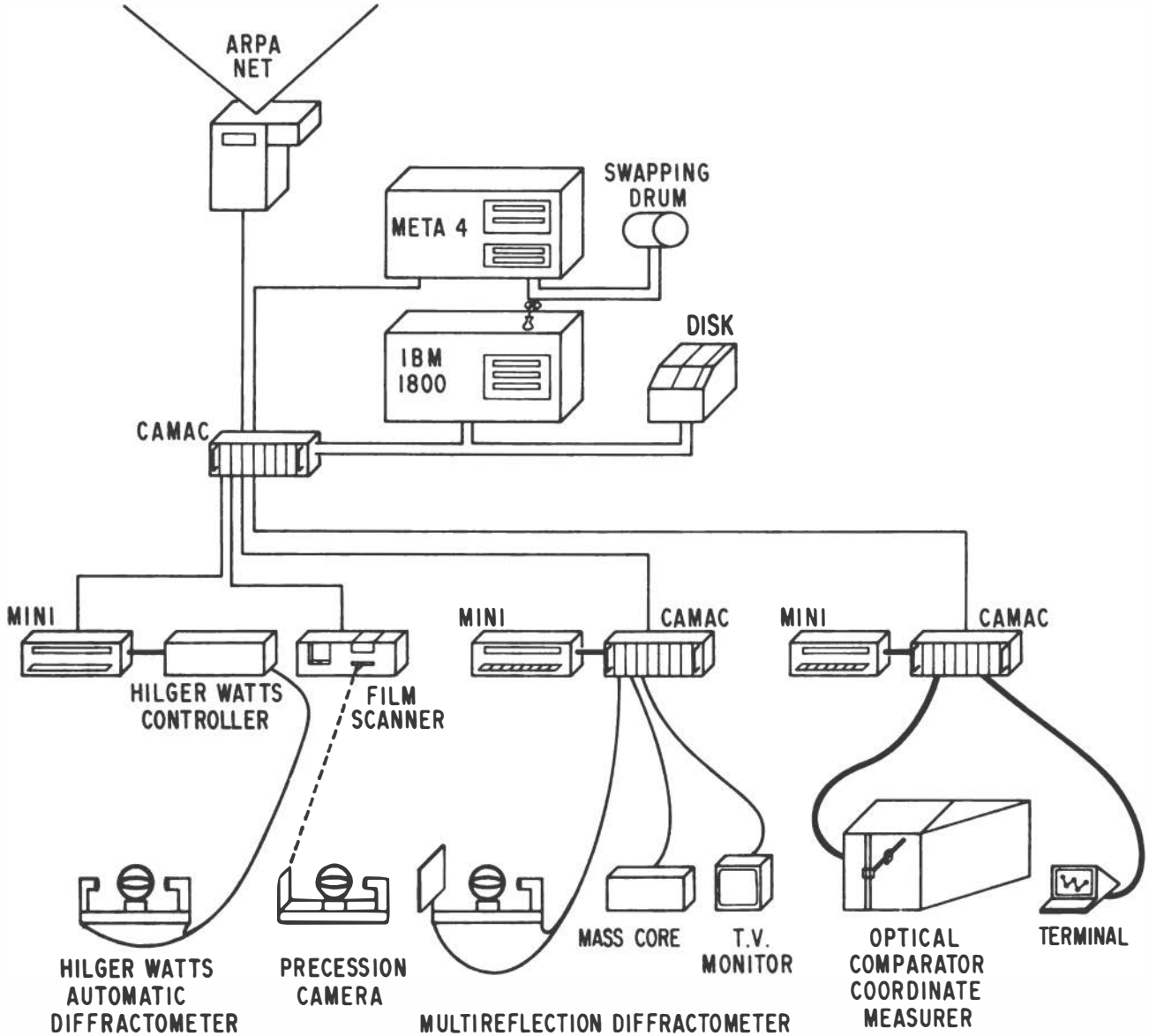

Figure 1 Computing System for Protein Crystallography

printer, a card reader/punch, a magnetic-tape drive, 2 keyboard/typewriters, a storage tube with interactive device, and a CALCOMP plotter. A Meta-4 computer, also linked to the IBM 1800, will be used as a local data-reduction processor for jobs that contain too much data or require too fast response time to be economically transmitted to a remote computer. At a higher level of the hierarchy, the IBM 1800 will be linked to a regional or national computing network.

The capital investment, as well as operational and system programming expenses, for all equipment except that dedicated to protein crystallography will be shared among seven research groups within the department of chemistry; this makes our local operation very cost-effective. A block diagram of the entire proposed system is shown in Figure 2. A significant portion of this system is already in existence.

To illustrate the practicality and advantages of hierarchical computing, we shall describe the portion of our system dedicated to collecting protein intensity data with an automatic diffractometer. The diffractometer is controlled by a dedicated PDP-8 computer which, through a fast communication link that can transmit data at a maximum rate of 100 000 words per second, is connected to the IBM 1800 situated 1000 feet away in another building. All of the PDP-8 programs are stored on the 2311 disc drives associated with the IBM 1800. The PDP-8 monitor can request the loading of various programs as they are needed during data collection, and these requests are usually satisfied within a fraction of a second since the IBM 1800 is operating under the MPX time-share system. Raw intensity data are screened by the PDP-8 in order to detect any slippage of the crystal and, after some preliminary data reduction (done in parallel with data collection of the next sequential reflection), the intensity measurements are passed in blocks to the IBM 1800 to be stored on the 2311 disc. This simple example of hierarchical computing allows a small PDP-8 with but 4000 words of core and no magnetic-tape or disc drive to operate the diffractometer as it it were a considerably more sophisticated computer with a large memory and disc storage.

Once a crystal is mounted and aligned, the system is capable of measuring data for days at a time without operator intervention and the output of intensity data is handled easily by the IBM 1800 peripheral I/O equipment. We were happy to find that the IBM 1800 time required is very small: less than 15 minutes per day. Our successful construction and use of the PDP-8 to IBM 1800 hierarchy has convinced us of the power, convenience and economy of hierarchical systems.

In the proposed scheme, our local hierarchical system will provide reliable rapid data collection and preliminary data processing, while access to a national computer network would be the best way to get the necessary computing power for protein structure determination and refinement. There is little doubt that within this decade a viable national computing network will revolutionize computing activity in the U.S. At the present time, the ARPA Network is the most highly developed network.

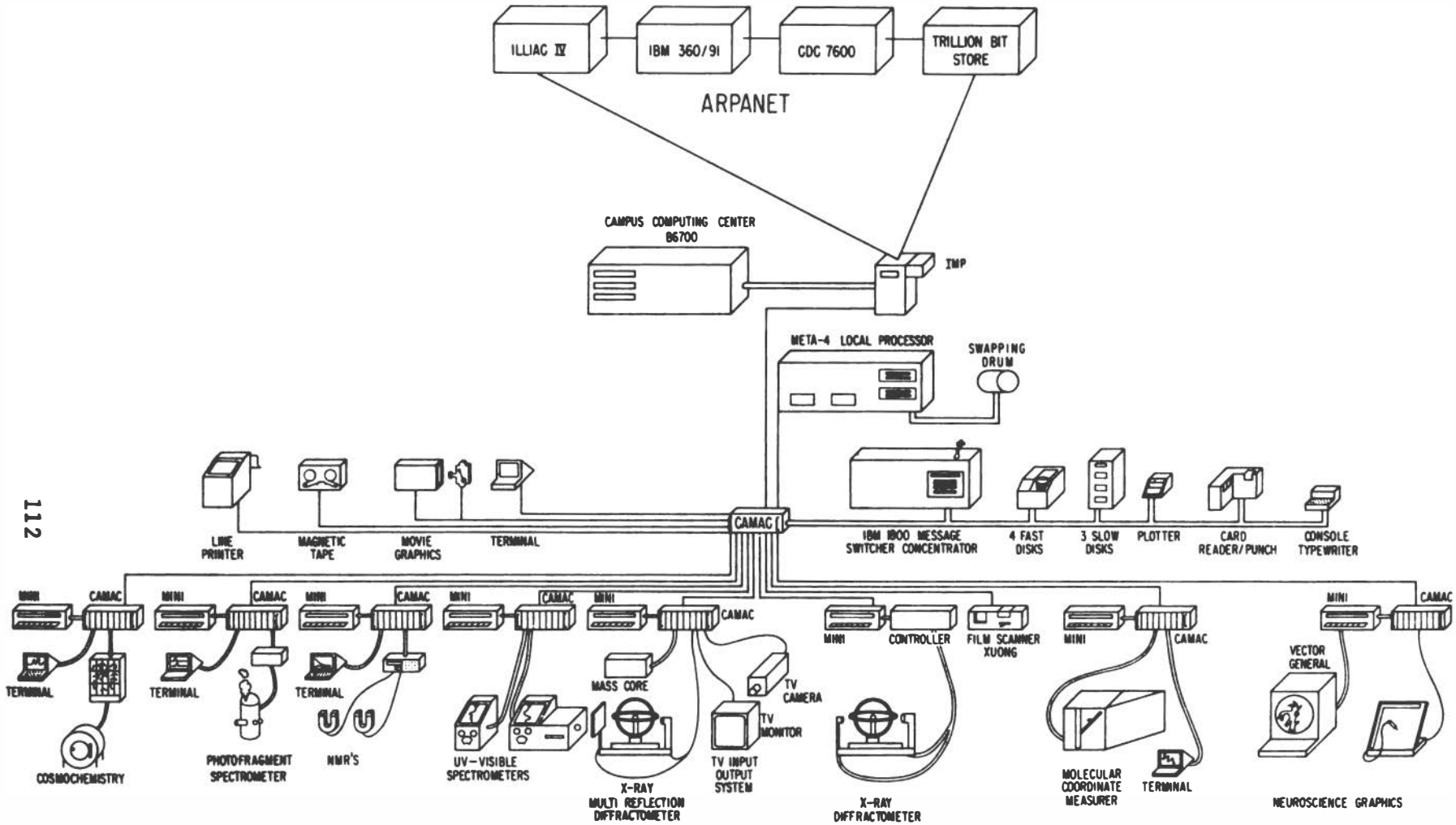

Figure 2 Block Diagram of System for Hierarchical Computing

The NET itself consists of small message store-and-forward computers called IMP's (Interface Message Processors), and wide-band AT&T leased telephone lines. HOST computers are connected to the NET through their local IMP. Each IMP is connected to at least two other IMP's, which insures the existence of alternate transmission routes between any two HOST computers. Communication between two HOST computers is handled automatically by these message processors which also select the optimum transmission route, depending on existing traffic.

The ARPA Network now contains about 24 sites with 37 HOST computers. These computers include representatives from all the major U.S. companies and range in size from a small PDP-11 to the huge CDC 7600 with the addition, in the near future, of the ILLIAC IV. In short, the ARPA Network links a group of heterogeneous computers distributed nationwide, in such a way that every local resource is available to any computer in the network. A geographical representation of the network is shown in Figure 3, which is taken from an article by Roberts (1971). The reader is referred to both this article and an article by LeGates (1971) for a comprehensive description of the philosophy and operational details of the ARPA Network. The five characteristics of the ARPA Network that are meaningful to users are: (1) easy access to a wide variety of computers, (there are now more than 15 different types of computers on the net); (2) negligible communication error rates (less than the error rates within a local computer); (3) rapid end-to-end response time (within 0.1 second); (4) fast data-transmission rate (about 80 000 bits per second); and (5) low cost of data transmission (less than $1 per megabit).

A national computing network will help the protein crystallographer in three ways: (1) it will enable him to handle new and exciting research problems by providing him access to the most advanced hardware, software, and data bases available, (2) it will decrease his computing costs by providing the optimum computer for each job and also by giving him access to computers subsidized by agencies sponsoring his research, and (3) it will make his life easier by eliminating the traumatic upheavals that occur with the periodic change of computers at his institution. As a matter of fact, since a network would bring about competition between computer centers, service in general should be upgraded. In addition, the redundancy of hardware within the network should considerably reduce research delays caused by extended computer down time.

In conclusion, we would emphasize that hierarchical computing, both at the local level, through sharing resources among many research groups, and at the national level, through connection to a continental computer network, is a practical way for protein crystallographers to satisfy their ever-increasing computational needs while at the same time maintaining a realistic computing budget. The necessary technology is already developed; what remains is the psychological and political acceptance of such interdependent resource sharing by the funding agencies, the research institutions, and by the scientists themselves.

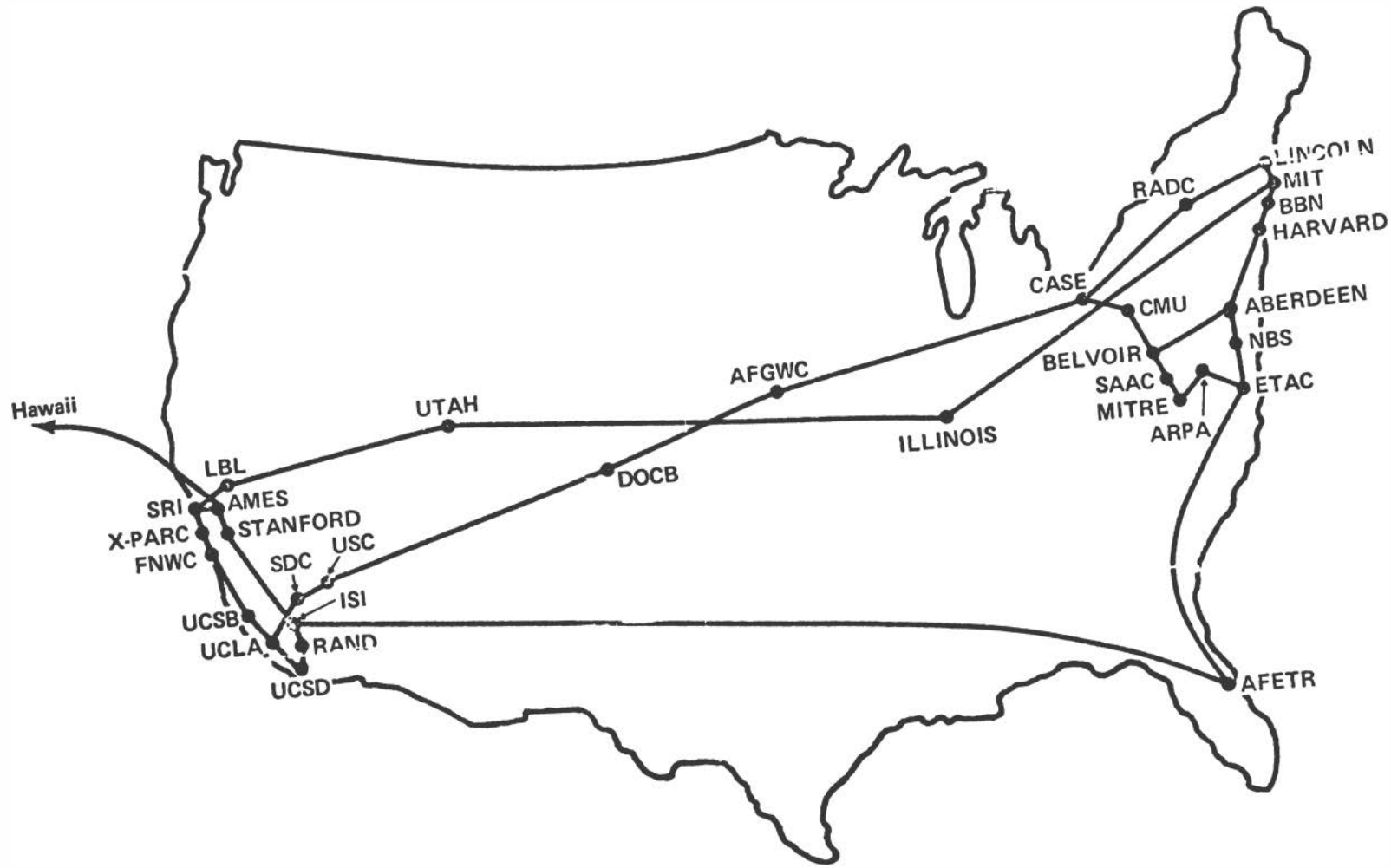

Figure 3 ARPA Network as of February 1973

We gratefully acknowledge the help of John Cornelius, Richard Alden, Kent Wilson, Wayne Vernon, and Joe Kraut in the formulation of the ideas presented here.

References

EURATOM, "CAMAC", a Modular Instrumentation System for Data Handling", European Atomic Energy Community Report No. EUR 4100e, (1969).

LeGates, J.C. "The ARPA Network Technical Aspects in Nontechnical Language", EDUCOM Interuniversity Communications Council (1971).

Roberts, L.G. "A Forward Look", Conference on Computers in Chemical Education and Research, DeKalb, Illinois, p. 7-5, (1971).

"A Rapid Data Acquisition System for Protein Crystallography", Xuong, Ng. H. and W. Vernon, Abstract J10, ACA Albuquerque Winter Meeting, (1972).

DISCUSSION

Berman: What does one have to do to get on the ARPA Network?

Freer: It depends on whether the ARPA Network will let you on. You first have to get the ARPA Network to sell you an IMP and they sell for fifty thousand dollars. Then you have to arrange for the cost of the computers with the various host computers as well. I don't think your question is a realistic one quite yet. The question is, "what has to be done now to make the net truly operational on a national level?"

Anonymous: In your local computing hierarchy wouldn't it have been just as convenient and a lot cheaper to connect the PDP-8 to a magnetic-tape drive rather than go to the expense of interfacing it to the IBM 1800?

Freer: No. We gave a lot of thought to this problem. We debated for a long time whether to buy a magnetic-tape drive or a disc drive, or a reader/punch for the PDP-8. We were also considering the possibility of buying more core storage. It turned out to cost about $2000 for the PDP-8-to-IBM 1800 interface whereas a magnetic-tape drive and interface would have cost us about $8000. So, for about a quarter of the price of a magnetic-tape drive alone, we gained the use of all these I/O devices when we hooked into the hierarchy. Of course we also gained an effective increase in core size and computing power as well.

Xuong: I hear rumors that the ARPA Network is looking for another agency to take it over. Is NSF or NIH going to take it?

Dewar: I've heard talks on the ARPA Network many times now and I've always asked the question, what is actually going on in the net, and the answer as far as I can gather is so far nothing. Is that still the situation or is there an appreciable amount of real traffic?

Lykos: I thought I had spoken to that point earlier. Computer networking technology has far outstripped the realization of its potential for resource sharing supportive to research. As far back as the 1970 ACM Meeting, Sidney Fernbach (in charge of the massive Lawrence Livermore Laboratory computer complex interfaced via an intra-site network) remarked in response to a direct question about the ARPA Network that one does not create a network just to have a network. The expression "The ARPA Network is a solution looking for a problem" has become a cliche amongst its detractors. As a matter of fact the brilliantly conceived ARPA Network, which was designed to be an experiment and a demonstration in computer networking, constitutes a major challenge to researchers to discover how computer networking can enhance the conduct of research.

Three researchers present at this meeting have just embarked on a highly relevant project. Walter Hamilton at Brookhaven is generating a protein-structure data bank, to be accessed from remote terminals which generate 3-D images. The terminals are being designed by Edgar Meyers at Texas A&M, with the cooperation of Helen Berman at Philadelphia who will play the devil's advocate through critical use of the evolving system from one of the prototype terminals.

The ARPA Network was discussed briefly earlier today. Some cost/performance figures and related considerations may be of interest. The ARPA Network is:

1. Telephone 'lines' leased from the telephone company 24 hours a day, 7 days a week, spanning the country with a bandwidth of 50

kilobits per second currently linking 29 nodes, each with at least two lines connected, with an annual telephone bill of $800 000. To place that in perspective, one voice-grade (2400 bits per second) line leased around the clock with cross-country dial-up access costs about $25 000 per year. Thus the cost of 32 single WATS (Wide Area Telephone Service) dial-up lines (from one point to anywhere in the continental USA, or, from any point to a given point) is about the same as the telephone bill for the much larger data-carrying capacity of the ARPA Network.

2. Interface Message Processors or IMPs which are small computers constituting the network nodes. Depending on how many host computers and/or local terminals need to be connected, IMP's cost $53 000 to $117 000. It is essential to the integrity of the ARPA Network that the IMP, hardware and software, be modifiable by the Network Manager only. (IMPs can be rented for about $1500 per month).

3. The on-line Network Information Center which compiles a listing of resources and facilities available on the network and maintains a journal facility enabling transfer of messages among various user terminals. Currently it is based at the Stanford Research Institute.

4. Network management and maintenance currently being handled by Bolt, Beranek and Newman.

5. Any computer intended to serve as a host interfaced to an IMP needs interface hardware, about $10 000 purchase, and software modification to conform to host protocol.

6. Users of the net establish accounts with the host computers of interest. The net facilitates remote access to a variety of hosts.

The Office of Science and Technology has taken the attitude that the ARPA Network is no longer an experiment but must be considered operational and therefore no longer appropriate as an ARPA project. Although the ARPA contractors who use the ARPA Network disagree with that position, bid specifications are being prepared such that some outside agency can take over and operate it. It would seem that a consortium of universities similar to consortia operating National Laboratories would be an appropriate agency but, so far, none has come forward.

On a pay-as-you-go basis (rental of IMP, prorata share of telephone bill), participation in the ARPA Network costs about $30 000 (1) per year exclusive of host computer use costs.

(1) "Networks for Higher Education", EDUCOM, 1972, pp. 7-12 and pp. 63-64

At the moment a three-fold load increase could be accommodated by the ARPA Network without a noticeable degradation in service.

Calvert: What's the capital tied up in this network right now?

Lykos: I have heard as an estimate that $10 000 000 has been spent for design, development, and implementation to create the experiment called "ARPANET". Capital in the IMPs is about $3 000 000.

Session V

Summing Up

Robert B.K. Dewar
Allen C. Larson
R.A. Young

SUMMING UP

Dewar: I will start with a comment on the qualitative estimates Walter Hamilton presented in his opening remarks of what expansion was expected in the field over the next five or ten years. Well, it depends on the number of graduate students, number of postdoctorals, and other unknowns, but nowhere was there a suggestion that it might depend on a number of worthwhile problems. I think that significant because I think a sociological change has occurred forcing us to look at our research programs in a different way. The way we shall have to follow is to determine that there are certain problems to be solved, to assess the costs associated with their solution and then estimate whether the solution is worth the associated costs. We have seen very little of that kind of reasoning in the last decade, because when one's expanding rapidly one doesn't need to answer these questions. But the age of rapid expansion in some ways is certainly slowing up, and I think the impact of this on the computer question in particular is that we must be very much concerned with using computer resources efficiently. My entire impression from many things said today is that we are not doing this now and we do not have a means for doing it.

For instance, we do not have rational systems for spending money. In most situations, the computer is sort of cream off the top which has to be scrounged from money in a different category from that provided for other expenses. I have made an estimate, which may be disputed and not entirely accepted though I believe it to be approximately correct, that a professional research scientist costs about $40 000 a year. That is the kind of figure non-profit and industrial research organizations assess. Then it is worth spending quite a few dollars of computer time to optimize utilization of personnel resources, whereas at the moment we have no mechanism for rational division of funds among the various costs. It really doesn't make much sense to talk about computer costs per structure. What we want to get at is the cost per structure and a rational way of minimizing that.

I have a further specific observation. I have a feeling information interchange is missing. There seems to be a lot of information about possibilities of computer hardware and computer software, and that information is not being transferred effectively within the crystallographic computing community. One of the things I hope would come out of this meeting is perhaps not a specific recommendation, say some network to use or some gigantic computer to be built, but that continuing committees should look into the whole question of efficient utilization of resources. I do not currently see any mechanism for doing this in an effective way.

Larson: Mr. Wagner of the AEC has informed us of the possibility that the AEC will acquire several new computers in the near future. Although he has suggested that it may be possible to buy time on these facilities, it may be very difficult. From what has been said, computing networks like the ARPA Network will come into being, but from what I have heard listening to Division of Military Applications personnel, I seriously doubt that many of the government laboratories, particularly the two with the largest computing arrays, will get into providing external services. If people are interested in trying to use the computing power that is available at the AEC laboratories, it might be worthwhile to contact the staff at these laboratories and see what can be worked out. But I would not hold out a good chance of success.

Walter Hamilton in his survey attempted to find out how much it costs to do a structure, or what the kinds of costs in doing structures are. I am going to interject some observations of my own from studies in our laboratory. I have written a code that comes moderately close to doing what Bill Busing's AXIØM does. I can set up and do the data reduction, go through the symbolic reduction without having given to the code anything more than the numbers necessary to control the processes. This comes out at the end of a short run with the structure-factor table printed and ready for publication, along with a stereo drawing of the structure. On one simple structure, in fact one of the two that D.T. Cromer talked about (at the ACA meeting preceding the conference), I believe the computing cost was about five dollars. That has nothing to do with the costs of collecting the data, but gives an idea of what happens with a code like AXIØM at a large computer center with a fairly sophisticated set of programs.

It is quite obvious that we are going to have to go to some kind of network. We are going to have to use large computers for many of the problems, particularly in the area of refinement which Dr. Johnson brought up. I recently asked our local systems people to set up so that I could easily do seven-hundred parameter least-squares on our equipment. They asked, "Are you sure this system is non-singular? Are you sure this system will not blow up?" I performed the tests they asked me to, and they then said, "This really is a stable system." Then they said, "We are proposing not to solve the least-squares problem for you in the normal procedure, but rather to solve it without formation of the normal equations." And this, of course, is a fairly reasonable procedure. You do not end up with all of the sums of products. It will probably not speed up the solution but we will not be playing with all of these squares. This procedure is some form of the Householder method; I don't know much about it but it is something our better theoreticians might take a close look at.

With regard to trying to do structures with mini-computers, I think this is obviously the way for a lot of small laboratories to go, or at least seriously to think about. Because as Sparks said, Vitamin B_{12} was solved on a computer that does not have the power of a PDP-8 or a NOVA. We are leaving a lot of our computing power sitting there doing nothing. I have been unhappy about that sort of situation in my laboratory for a long time, but have been too busy trying to make the big computer do all my work to go down and use the little one.

Young: What was the $5 structure? Sodium chloride?

Larson: It's the azobiscarbamide Cromer talked about. It's a pair of nitrogens sitting across a center of symmetry with a carbamide group attached to the nitrogen at each end of the azo chain. This is a simple structure.

Hamilton: You mentioned AXIØM. There may be a number of people here who weren't at the ACA banquet the other night who don't know what AXIØM is.

Larson: AXIØM is a procedure where you put the crystal on the diffractometer, mount it and center it, and then go away. The device selects a good single crystal, orients it on the diffractometer, collects the diffractometer data, sends it over to the computing center which processes the data, makes the assumptions I had to make in my code, performs the symbolic addition, and what have you. You have to tell it in some way how much you're expecting, I guess. Maybe it can do that too, why not? Anyway, you end having all the information worked up automatically and the structure poured out automatically at the other end. We haven't started on the paper-writing code yet. That seems to be all we need.

Young: I thought one of the nicest things about AXIØM was the reason Bill Busing pointed out for the choice of its name: because you can take it for granted.

I will structure my remarks by first stating what I now think to be the question we were addressing (for it is not the question I thought it was when we came in), by then giving some pros, cons, and alternatives, and by winding up with a recommendation. It seems now that the question may be stated as follows: With two demonstration projects and a feasibility study now under way with NSF money, with the ARPA Network now in operation and about to go commercial, with all three agencies we've heard from expressing strong interest in networks involving fewer major computer centers, the question is not whether we think network and remote user systems with capabil-

ities suitable for crystallography should be established but, rather, whether we want crystallography to be counted in on developments that are obviously occurring anyway. The question might even be raised of possibly increased future difficulty of getting computer dollars from the agencies if crystallographers failed to take what the agencies think to be the advantage of these developments. We should at least look at them carefully. If we decide not to take advantage, we should certainly know why in a rational way.

The crystallographic community has a strong history of cooperative effort, both national and international. For example, the data compilations in the crystallographic field are widely recognized as outstanding examples of cooperative effort. So crystallographers would be a good group to consider cooperative efforts in computing, certainly. It also rather seems to me that this group is in fact interested and rather positively inclined toward doing so.

Let us consider this question of remote users. I would like to summarize some of the main pros and cons I've heard stated at this symposium.

On the pro side I have selected about five points. It seems clear to me that protein crystallographers need larger computers than are now available in any one place. Second, the possibility has been raised of a quantum jump resulting from almost a new dimension of computing capacity that a network might make available (a network backed up, of course, by the most advanced computers and systems). Third, it appears that sufficiently reliable, high-speed data transmission is available, now or in the near future, to obviate the need for physical transport of data or results. I was much impressed with this possibility of 80 000 bits per second, and envious of the turn around time that Helen Berman quoted. My turn around time is nothing like that, especially not near the end of the quarter when all the students are crowding around the computer. Fourth, for the largest component of the cost of computing, which is refinement, most of us are perfectly happy to use somebody else's well-tested program, and do just that. In our laboratory we'll probably expand our use of other people's programs as systems such as Jim Stewart talked about become available and are adapted to our local computing conditions. Many of you are already using Stewart's X-RAY 70 or the Busing, Martin, Levy least-squares refinement program, to name but two of several widely used programs. As these new systems get developed and checked out, it seems only simple wisdom for most of us to ride along on the work of these good programmers, and use their powerful, cost-effective systems. Presumably these good systems would then be available on a crystallographic computing network. Fifth, Ed Meyer mentioned the possibility of information retrieval from a central library, information retrieval not only

in hard copy but possibly on one's own display terminal. You'd call in to Chemical Abstracts or some center for Crystal Data Compilation (which is obviously going to get on a computer now), ask the necessary coordinates for a structure, and a stereo view would also appear on your display terminal. Finally, a number of other exciting possibilities were mentioned by Dr. Freer.

On the con side, I heard the following problems raised. The first is the computing cost. In our campus environment, real costs are subsidized and hidden to some extent, and the amount of costs charged to the project is often a small part of the real cost. If we are ever faced with having to pay full real costs, we may find that we do not do as much computing. We cannot ride along indefinately on the institution's computing program.

A second point against the network approach is that terminal and telephone costs are not trivial. Helen Berman quoted $917 per month for the terminal costs and about $200 a month for telephone service; to that one must add computer costs, for CPU time primarily. The sum of these charges gets to be fairly substantial. One has a fixed cost of $917 per month for the terminal whether one does anything or not. Clearly, one has to be in the crystallographic computing business deeply before it becomes worthwhile to meet the fixed costs for network use.

A third point against, although I did not hear this said explicitly, is the possibility of weakening local resources. While removal of the main crystallographic computing load from the local computing center might not be disastrous in itself, if the main computing load in several other fields were also transferred to a network, the reduction of support for and demand for services from the local center would certainly affect its development adversely. That would mean, ultimately, less computing capacity available to other faculty members and the professional staff. Furthermore, diversion of computing and its dollar support may ultimately produce a local cost problem that will reduce the amount of free time available to us and to our colleagues at our own institution. Another question I think was raised concerns our contribution to program development. An effect of going to a network might be to decrease the effort put into this area because of (1) the difference in the cost basis of the computer time available, and (2) the possible feeling of awkwardness in treating a program remotely. The result could be less rigorous development of better computing programs and, even, hardware. A further point against whole-hearted committment to networks is the specter of disaster wreaked if, for political or economic reasons, the network failed or stopped rendering services.

An interesting alternative was offered by Dr. Sparks, and Dr. Larson has addressed himself to it. It is the use of an "adorned"

or "festooned" minicomputer in association with the diffractometer.

If it were decided that crystallographers should consider getting onto a network, it seems to me that Ed Meyer has raised a number of specific practical questions that have to be answered. A small committee might be formed to study these questions further, along with the possibilities Steve Freer has raised, and other points, to determine whether it really is worthwhile for crystallography, as such, to opt for network computing as a prime resource. It seems to me we have made a good start, and have a good basis for further informed study of the options.

Larson: A further point I wish to bring up is that the development of this next generation of computers is along the line of parallel processors, which are essentially vector processors. If you think about the protein problem, or for that matter any crystallographic least-squares problem, the whole thing can be quite easily broken down into a vector procedure. Some of the problems we are contemplating may make much better utilization of these machines than the vast majority of other types of projects or other types of number-crunching games that, for instance, I am competing with. We have long vectors; with a thousand data points a vector of hkℓ is a thousand points long. For a protein, you have a vector of x,y,z that is 3n long. These vectors are long enough so that these vector machines can start to move. I think the protein people should seriously address themselves to the possibility of using these machines.

Hamilton: I had hoped to have someone here to tell us about the capabilities of ILLIAC-4. A related point is that once the protein people decide which really are the best algorithms for the refinement of their structures, one might even design special hardware to carry out these algorithms very specifically.

Dewar: Once one has a network one no longer has a drive to move things around to all computers on the network and one can't afford to write highly specialized code for highly specialized hardware. I'm sure that if 10 million dollars was spent on crystallographic computing several million of it could have been saved by rewriting critical parts of least-squares programs in assembly language. I think there's no doubt about that, and it becomes a lot more practical and attractive on a network where you can do a lot of local tailoring and use it in situ in a place where it's going to work.

Larson: This comes back to the question of developing programs. Over a period of many years at Los Alamos, from the time we started to do single crystal work up until three or four years ago, it was hard to talk us into even giving away a program. The major reason was that we felt that in general a person who knows what his computer

program does is more likely to use it correctly than someone who is using one taken on faith from someone else. So, there are many considerations to balance.

Frey: For routine structure determination, first question, is it useful to refine to death? Second question (if not, as I think), is it useful to save the data you collected?

Larson: No, it is not useful to refine something to death. No it is not useful to save the data you've collected.

Frey: But I know no means, no way actually at the moment, to save, say, a factual structure determination. It's difficult to see what I'm going to publish.

Hamilton: There are ways to save the data, as you know, by publishing them in Acta Crystallographica and various data banks being maintained, so that if somebody else wants to come along later and refine them to death, perhaps they can use their computing time to do it. I think that's a good point--by all means collect the data as well as you can, but perhaps only abstract the information you need from it yourself and save it for someone else to carry on with. I think there are people who would argue the point.

Larson: I'll argue with this point. I think you should collect the data to do what you want done. If you want to find the atoms of the structure and you don't care about the thermal motion, collect the data in a mode that will locate you the atoms and give something that neither you nor anybody else will believe as the thermal parameters and then throw the data away when you get through with it. You might publish the numbers, but the cost of recollecting data on moderately small structures is actually very small. How much time does it take for an automated Picker diffractometer to collect 3000 data points? Our standard procedure runs about 600 a day. That's five days. And it runs while I do something else.

Dewar: If it's desirable to maintain data banks, clearly networks and trillion-bits stores are useful.

Larson: The task you're talking about is presumably protection of a local software system, and this is essentially the security problem we are up against. If you put a computer onto a system of this sort, you must have the system designed with safeguards such that each and every individual computer can detect any attempt by the system or network to steal information. There should be no way for computer A to find out anything about the system or data or anything else at computer B without permission. There should be no way to find out about it's system and the possibility of acquiring

data will require knowing sufficient passwords and other things to get through the red tape.

Anonymous: You're saying there will actually not be perfectly free access from my local computer to any other in the net?

Larson: Would you like Walter Hamilton to have free access to your carefully collected structure factors?

Anonymous: I'm not talking about that. I m talking about such things as optimizing a PL-1 compiler, for example.

Dewar: There is a significant question here having to do with whether IBM will be upset when the number of PL-1 compilers on order goes down by two orders of magnitude. I'm not quite sure what one can hypothesize about that situation. IBM will probably work out something that seems reasonable to them and that seems reasonable to people who use the network, and they'll buy it. Otherwise they won't.

Thomas: There is another implication of networking that Ray Young mentioned but that hasn't been made explicit enough. We might all agree that there is developing a dependency on computers, not only by crystallographers but by everybody else, which approaches the dependency that we have on electric power for example. That kind of dependency can be catastrophic unless sufficient redundancy is built into the networks and systems. Thus a network must not be allowed to fail totally, for hardware, software, or almost any other reason. One of the points of this symposium should be to express our concern regarding the required reliability and redundancy that must be built into networks in order to avoid the kind of collapse one can easily foresee.

Hamilton: That's a very good point. If all of the crystal structures are being refined on one computer that has one bit wrong in some position, then all the structures in 1972 will be wrong for the same reason.

Dewar: We seem to be getting swept away by the ARPA Network and for many of us it's a first exposure. But there is another view which is that the computations we're performing are basically trivial. To a computer scientist, for instance, they are general, perhaps in the last analysis even boring, problems just consisting of simple arithmetic, and part of what we are fighting today is that big machines are not designed for simple arithmetic. They're designed for data retrieval, swapping, paging, devouring resources, etc. There is a cogent argument for establishing a big computer designed solely for numerical processing, getting rid of all these problems that come from information handling. From this point of view, elim-

inating completely extraneous uses of the computer, one can certainly get utilizations and efficiencies that are greatly enhanced, and I think we should not neglect the possibility that the most effective way of dealing with crystal structure solution may be to establish one crystallographic computer specifically oriented towards that task. The aims are similar enough to those of the quantum chemistry study that there is a confluence of objectives here.

Sayre: A National Computation Center does not in itself satisfy the desire for rapid transmission of problems from smaller installations. Furthermore, the placing of such a Center in a network would not preclude the development of special-purpose hardware at that Center aimed at providing particularly efficient service for crystallographers and other special groups. Therefore it seems to me that the committee considering such a National Computation Center might stipulate that such a Center, if it is to be created at all, should be created within the framework of a computer network.

Dewar: I just want to point out that the existence of the network today may not be the entire solution. We may still want particularly to look at special-purpose hardware.

Kay: If networks are going to be set up in the next few years, offshore areas such as the Caribbean where both I and my colleagues such as Fletcher work should be remembered. Hawaii, Alaska, or even Montana, have similar problems. Our telephone system is rather poor and the international lines are probably not fast, and are very noisy. I wonder if satellite transmission has been thought about for remote areas? It would be helpful if some provision were made for international usage for the benefit of undeveloped areas. Interested people in Jamaica or Venezuela or other places that have moderately active groups could probably make good use of U.S. data bases. It could aid in their economic and educational development.

Lykos: There already exists a system in Hawaii called ALOHA. They have a version of the IMP called MENEHUNE ('little elf'). What is different about their system is that radio transmission is used rather than copper wire and microwire (i.e., telephone circuits). In fact, Larry Roberts, ARPA director of the ARPA Network, discusses the implications of radio transmission between a handheld transceiver and a computer/communications system at the 1972 Spring Joint Computer Conference, of which Proceedings will be distributed at the Conference (May 16-18, 1972). With regard to your second point, satellite data communication is already happening. In October of 1972 there will be an International Conference on Computers and Communications at Washington, D.C., designed to provide a forum for technologists in Computer Science, Economics, and Law. A demonstration project is planned involving India, England, and the USA. Should the world

of crystallography have a contribution to that demonstration, I would be happy to provide a contact.

Oxman: The National Library of Medicine's Lister Hill Center for Biomedical Communications is currently involved in trial use of a NASA satellite for transmission of voice and data communications between the center, located in Bethesda, the University of Alaska, and designated clinical centers to improve health-care services in Alaska. This is perhaps a basic research effort to develop and test the feasibility of using satellites for the communication of data.

Fritichie: I'd like to ask the representatives of the government agencies, in view of their encouragement of large computing networks and centers, what is the attitude towards scientific data banks, such as for example a depository for crystal structure information. A decade ago or so this sort of thing was discussed, but the technology then essentially was not up to the conception. Has that point been passed and are the agencies now willing to put money into support of construction of such data banks? We seem to be close to the point of eliminating publications. If everyone is hooked into a data circuit and has access to data, once they have been appropriately referred and deposited, why publish at all?

Lykos: You are speaking directly to the programmatic thrust of the NSF Office of Scientific Information Systems. Program and data sharing by network not only provides a medium of communication between scientists with similar interests but also between scientists who are finding channels of communication via problem-solving algorithms implemented as computer programs. Of course this will have an impact on the very process of publication. Crystallographers who want to get into the creation, maintenance, and accessibility of data banks should get in touch with OSIS.

Hall: Crystallographers should not get as involved in systems programming as we have done in the past. I refer particularly to machine and assembly language coding. If this is needed for special applications, our efforts would be a lot better employed if we directed people who are really trained to do it. We all agree that if you replace the much-used program DO-loops with machine code you can speed things up, but with the optimizers available in most compilers today the gains are minimal. Much more time and money have been lost over the years in changing other people's programs from an assembly code back into a form that will suit your machine than has been gained by this operation. With vector computers of the future, again I think it should not be our concern to optimize the vectoring of our programs. Our objectives with a computer such as ILLIAC-IV should be to ensure that the systems analysists supply a compiler that will

vector it efficiently for us. The crystallographer at large should not have to worry about such systems problems. I would like to put in a plug for generalized FORTRAN programming that can be used in a black-box fashion. Allen Larson suggests that all users must understand the workings of the crystallographic programs they are using. I think those days are well and truly past. Already we have chemists, mineralogists, and others using our program who know little about crystallography per se. Such use is likely to increase with time, and I do not view it as a bad thing.

Dewar: A lot of people who are normally crystallographers are actually first assistants to programmers. A lot of people working in crystallographic laboratories are really first-rate programmers who have contributed a tremendous amount of computer technology. One of the ideas back of the proposal for a National Center for Computation in Chemistry is bringing together a group of highly trained personnel, not just hardware. Some of that tremendous expertise is terribly diffused at the moment. I cannot agree that it is not worth looking into specialized programming. There are people who have interests right up that street and, given ideal situations, could achieve tremendous things. Certain chemical crystallographers should not get involved, but there are others very much interested, and they should be given an environment in which they can work effectively.

King: A crystallographer cannot be expected to design the system, but an appreciation of what the systems can do would avoid binds where some FORTRAN code does not effect the complete transfer of one word. If you appreciate the differences between systems, you can sometimes avoid costly situations. Also, such understanding can help you write your FORTRAN code so as to use whatever is most efficient for the particular system. I have found that, whenever I wanted to write a subroutine in machine code in order to get a tighter loop, a nice way to do this when using batch processing was to get the compiler listing from the FORTRAN code, and then remove the redundant steps. This went about as far toward making the loop tighter as anything one could go, and besides, one learned a little bit about the system, even if only about the practical details of the FORTRAN system that could be safely eliminated. Now that we are using a remote terminal, I regret that we do not get a compiler printout to tell us how to improve our own FORTRAN programs.

Hall: One should not have to worry about whether a half word is transferred in a computer operation, or not. The computer people should be made to fulfill the specifications as defined in ASCI-II, etc., and our efforts would be best used in ensuring that such standardization is adhered to. As users, we should not have to be concerned about such special hardware operations.

Larson: I think it should not be necessary as you say for us really to

worry about it. The one who's doing the systems programming like Jim Stewart or myself has to worry about that. He has to have it done so that if we deliver these systems to you, you can get them on without too much trouble. It's sometimes very difficult to do that.

Anonymous: I would like to echo Dr. Hall's sentiments on this matter. Professor Dewar started out saying that crystallographers should be concerned about the social importance of their problems, but then he turned around to chide us for not putting our programs into specialized machine language at a specialized computing center. The emphasis has been so much on efficiency, as opposed to the thought that the networks and the national computing centers would perhaps enable the protein crystallographer to formulate and actually solve his problems. I would hope that the study of a national center would be much more concerned with the scientific value than with the efficiency that would result.

Dewar: I do not think you can separate the two, There are problems with price tags attached to them, and increasing efficiency is the name of the game, because that is what puts more problems within reach.

Young: Surely we should do whatever we do efficiently, but we should not be so concerned just with efficiency that we fail to look about to see what else we could be doing. Steve Freer touched on that point when he suggested that the availability of this network may produce a new dimension in computing, a quantum jump. We may be able to take on new problems we had not conceived of before.

Stewart: But the quantum jump will not come in crystallography if all of us are continually trying to second-guess the systems people on the network.

Larson: You are, of course, pounding at lack of communication with the software designers. We are not apparently getting to them the message that FORTRAN-IV should be FORTRAN-IV or PL-1 or whatever language.

Young: That message is clear enough.

Larson: But they are not reading it! We should not have to push that message to them. Those of us who are programmers should be able to take a FORTRAN compiler on any machine and presume that the FORTRAN deck will produce the same results on an IBM 360, on a Univac 1108, on a CDC, Honeywell, or GE. That is the reason you have gone to your pidgin FORTRAN, and even your pidgin FORTRAN is not legal on all machines. The biggest trouble I think you are having there on that half word is in the ability to use half-word integers or two byte integers.

Hamilton: Is there a consensus here that the crystallographic community is interested in exploring seriously the possibilities of crystallographic computing networks and centers, and that this message should be carried to the group that is studying the theoretical chemistry center? It seems to me it would be a good idea if the ACA Computing Committee also were to discuss this matter.

Larson: The ACA Computing Committee plan in their report to make a strong suggestion that people submit programs to the Quantum Chemistry Program Exchange.

Hamilton: I was suggesting that the ACA Committee might give a few hours of good discussion to the topics we've raised today, perhaps to see whether the ACA wishes to stimulate any positive action on the part of the funding agencies. Any final comment?

Fritchie: One problem that was mentioned fairly early has been somewhat neglected in our discussion. That is the human problem of maintaining a service attitude in such a center. The advantage of competition from several equivalent centers has enormous potential, and I think this point should be considered carefully by any study group.

Calvert: I would not like a consensus to go forward on the basis of one particular area of interest. As a society we should be careful to remember that a large number of crystallographers are interested in small or medium-sized molecules. The special problems peculiar to very large molecules might better be discussed separately. It is possible that small structures can continue to be done simply and cheaply by existing techniques.

Hamilton: I would agree with that. I certainly wish to thank all of the people who have participated, the principal speakers and members of the audience as well. The symposium is adjourned.

Appendix 1

QUESTIONNAIRE. Please return before 1 March to

Walter C. Hamilton
Chemistry Department
Brookhaven National Laboratory
Upton, New York 11973

To Assist in the preparation of background material for the Symposium on Computing Needs and Resources in Crystallography, the organizers would be especially grateful if each member of the ACA (only one spokesman for each group) would complete and return this questionnaire at his earliest convenience.

1) Is your work primarily ( ) structure determination
( ) other (please specify)

If structure determination, are the structures generally
( ) small ( ) medium ( ) large?

2) What model computer do you use for most of your crystallographic computations? If none, so state.

_____

3) What is your real computing cost per year? $_____

a) How much of this is from individual federal grants? $_____

b) How much from other grants? $_____

c) How much is subsidized by your institution? $_____

4) Please estimate the proportion of your computing costs devoted to

a) Collection & preliminary processing of data _____%

b) Direct methods, Patterson methods _____%

c) Refinement _____%

d) Other (please specify) _____%

5) How much do you expect your computer usage to increase
a) By 1977_____% b) By 1982_____%

6) What is the computing cost of an average structure determination (if any) in your laboratory? $_____

7) Have you ever	Yes	No
a) Carried out computations from a remote terminal?	___	___
b) Made extensive use of computers not in your own institution?	___	___
c) Used the ACA or IUCr program lists?	___	___

8) Do you feel that limitations on the quality or quantity of the computing available to you affect your work

( ) seriously ( ) moderately ( ) Slightly?

Appendix 2

List of Participants

S.C. Abrahams, Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
Gopal Ambady, Roswell Park Memorial Institute, Buffalo, New York
Herman L. Ammon, Department of Chemistry, University of Maryland, College Park, Maryland
Marcia F. Bailey, Department of Physics, Central Michigan University, Mt. Pleasant, Michigan
Werner H. Baur, Department of Geological Sciences, University of Illinois, Chicago, Illinois
John D. Bell, Department of Chemistry, University of California, Los Angeles, California
Helen M. Berman, Institute for Cancer Research, Philadelphia, Pennsylvania
John F. Blount, Hoffmann-La Roche, Inc., Nutley, New Jersey
Robert Blumenthal, ENRAF-NONIUS, Pittsburgh, Pennsylvania
Bruce W. Brown, Department of Chemistry, Portland State University, Portland, Oregon
George M. Brown, Oak Ridge National Laboratory, Oak Ridge, Tennessee
Gilbert Brussell, Sandia Laboratories, Albuquerque, New Mexico
William R. Busing, Chemistry Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee
Sue Byram, Syntex Corporation, Palo Alto, California
Howard H. Cady, Los Alamos Scientific Laboratory, Los Alamos, New Mexico
L.D. Calvert, Chemistry Division, National Research Council, Ottawa, Canada
Horace L. Carrell, Institute for Cancer Research, Philadelphia, Pennsylvania
Charles N. Caughlan, Department of Chemistry, Montana State University, Bozeman, Montana
J.R. Clark, Naval Postgraduate School, Monterey, California
H. Cole, International Business Machines Research Center, Yorktown Heights, New York
Philip Coppens, Department of Chemistry, State University of New York, Buffalo, New York
Peter W.R. Corfield, Department of Chemistry, Ohio State University, Columbus, Ohio
David Crump, University of Western Ontario, London, Ontario, Canada
David L. Cullen, Department of Biochemistry and Biophysics, Texas A & M University, College Station, Texas
Betty Rolfs Davis, Department of Chemistry, Brookhaven National Laboratory, Upton, New York
Victor Day, Massachusetts Institute of Technology, Cambridge, Massachusetts
Robert B.K. Dewar, Department of Computer Science, Illinois Institute of Technology, Chicago, Illinois
Brian Dickens, National Bureau of Standards, Washington, D.C.
Robert J. Doedens, Department of Chemistry, University of California, Irvine, California
Gabrielle Donnay, Department of Geological Sciences, McGill University, Montreal, Canada

J.D.H. Donnay, Université de Montréal, Montreal, Quebec, Canada
James W. Edmonds, Medical Foundation of Buffalo, Buffalo, New York
Earl F. Epstein, Department of Chemistry, Colorado State University, Fort Collins, Colorado
Donald J. Fensom, Noyes Laboratory, California Institute of Technology, Pasadena, California
Alvin Fitzgerald, Department of Chemistry, Montana State University, Bozeman, Montana
Steven T. Freer, Department of Chemistry, University of California, San Diego, La Jolla, California
Bertram A. Frenz, Department of Chemistry, Texas A&M University, College Station, Texas
Michel N. Frey, Department of Chemistry, Brookhaven National Laboratory, Upton, New York
Charles J. Fritchie, Jr., Department of Chemistry, Tulane University, New Orleans, Louisiana
E.J. Gabe, National Research Council, Ottawa, Canada
Richard D. Gilardi, Naval Research Laboratory, Washington
Sydney R. Hall, Mineral Sciences, Department of Energy, Mines, and Resources, Ottawa, Canada
Walter C. Hamilton, Department of Chemistry, Brookhaven National Laboratory, Upton, New York
Jonathan C. Hanson, Department of Bilogical Structure, University cf Washington, Seattle, Washington
Herbert Hauptman, Medical Foundation of Buffalo, Buffalo, New York
Stuart W. Hawkinson, Department of Biochemistry, University of Tennessee, Knoxville, Tennessee
John Hodson, University of Washington, Seattle, Washington
Francis E. Holmes, Bureau of Narcotics & Dangerous Drugs, Department of Justice, Washington, D.C.
Alan E. Hutchings, Kew Corporation, Cleveland Heights, Ohio
James A. Ibers, Department of Chemistry, Northwestern University, Evanston, Illinois
G.A. Jeffrey, Department of Crystallography, University of Pittsburgh, Pittsburgh, Pennsylvania
Carroll K. Johnson, Chemistry Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee
Gopinath Kartha, Roswell Park Memorial Institute, Buffalo, New York
Henry Katz, University of Pennsylvania, Philadelphia, Pennsylvania
K. Ann Kerr, Department of Physics, University of Calgary, Alberta, Canada
Murray Vernon King, Laboratory of Physical Biochemistry, Massachusetts General Hospital, Boston, Massachusetts
Richard M. Kirchner, Department of Chemistry, Northwestern University, Evanston, Illinois
Thomas F. Koetzle, Department of Chemistry, Brookhaven National Laboratory, Upton, New York
Robert Lancaster, ENRAF-NONIUS, Pittsburgh, Pennsylvania
Allen C. Larson, Los Alamos Scientific Laboratory, Los Alamos, New Mexico

E.C. Lingafelter, Department of Chemistry, University of Washington, Seattle, Washington
A.H. Lowrey, Naval Research Laboratory, Washington, D.C.
Peter G. Lykos*, Office of Computing Activities, National Science Foundation, Washington, D.C.
Gretchen Mandel, University of Pennsylvania, Philadelphia, Pennsylvania
Neil Mandel, University of Pennsylvania, Philadelphia, Pennsylvania
R.C. Medrud, Technical Staffs Division, Corning Glass Works, Corning, New York
Edgar F. Meyer, Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas
Alan D. Mighell, National Bureau of Standards, Washington, D.C.
C.M. Mitchell, Department of Energy, Mines and Resources, Ontario, Canada
Michael R. Murphy, University of California, Los Angeles, California
Hugh B. Nicholas, Biophysics Laboratory, University of Wisconsin, Madison, Wisconsin
Yoshi Okaya, Department of Chemistry, State University of New York at Stony Brook, Stony Brook, New York
Michael A. Oxman, Biotechnology Resources Program, National Institutes of Health, Bethesda, Maryland
R. Parthasarathy, Roswell Park Memorial Institute, Buffalo, New York
George T. Pashape, ENRAF-NONIUS, Pittsburgh, Pennsylvania
Martin A. Paul, Division of Chemistry and Chemical Technology, National Research Council, Washington, D.C.
Nicholas C. Payne, Department of Chemistry, University of Western Ontario, London, Ontario, Canada
Carl F. Pihl, International Business Machines, E. Fishkill, New York
Joseph J. Pluth, Department of Geophysical Science, University of Chicago, Chicago, Illinois
Rod Restivo, Department of Chemistry, University of Virginia, Charlottesville, Virginia
Robert Rosenstein, Department of Crystallography, University of Pittsburgh, Pittsburgh, Pennsylvania
Fred Ross, State University of New York at Buffalo, Buffalo, New York
David Sayre, International Business Machines Research Center, Yorktown Heights, New York
Jesse W. Schilling, Yale University, New Haven, Connecticut
Ryonosuke Shiono, Department of Crystallography, University of Pittsburgh, Pittsburgh, Pennsylvania
Verner Schomaker, Department of Chemistry, University of Washington, Seattle, Washington
Nadrian Seeman, Department of Biological Sciences, Columbia University, New York, New York
Douglas L. Smith, Eastman Kodak Company, Webster, New York
Robert A. Sparks, Syntex Analytical Instruments, Palo Alto, California

*On leave from Illinois Institute of Technology

Richard H. Stanford, Jr., Department of Chemistry, California Institute of Technology, Pasadena, California
James M. Stewart, Department of Chemistry, University of Maryland, College Park, Maryland
F.L. Suddath, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts
J.T. Szymanski, Mines Branch, Department of Energy, Mines and Resources, Ottawa, Canada
David H. Templeton, Department of Chemistry, University of California, Berkeley, California
Robert Thomas, Department of Chemistry, Brookhaven National Laboratories, Upton, New York
Louis P. Torre, Bell Telephone Laboratories, Murray Hill, New Jersey
Robert Vardeman, Sandia Laboratories, Albuquerque, New Mexico
Donald Voet, Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania
J.F. Wagner, Data Processing Evaluation & Control Branch, Atomic Energy Commission, Washington, D.C.
Keith Watenpaugh, Department of Biological Structure, University of Washington, Seattle, Washington
Charles M. Weeks, Medical Foundation of Buffalo, Buffalo, New York
Chin Hsuan Wei, Oak Ridge National Laboratory, Oak Ridge, Tennessee
Johnnie Marie Whitfield, Department of Chemistry, Louisiana State University, Baton Rouge, Louisiana
Jack M. Williams, Chemistry Division, Argonne National Laboratory, Argonne, Illinois
Richard M. Wing, Department of Chemistry, University of California, Riverside, California
Robert D. Witters, Department of Chemistry, Colorado School of Mines, Golden, Colorado
Gerard M. Wolten, Aerospace Corporation, Woodland Hills, California
John Woods, Picker Corporation, Cleveland, Ohio
Nguyen-huu Xuong, Department of Chemistry, University of California, San Diego, California
R.A. Young, School of Physics & Engineering Experiment Station, Georgia Institute of Technology, Atlanta, Georgia
Allan Zalkin, Department of Chemistry, Lawrence Berkeley Laboratory, University of California, Berkeley, California