INTERNATIONAL GEOLOGICAL CORRELATION PROGRAM

December 9, 1980

Circular 80-8
IGCP-163-IGBA

On Whether Derived Variables Should be Stored in a Data Base or Computed as Needed: Economic Considerations

In chemical petrology we use many 'derived' variables, variables computed from part or all of the original data. At the recent Madrid meeting of the project there was lively discussion about whether such derived variables should be generated as required at operation time or retrieved directly from--and therefore stored in--the base.

With regard to relative costs, what is at issue is essentially a trade off between storage and computation charges. The economic argument for generation at operation time is based on the conviction that active mass storage is more expensive than computation. In the early days of the 3d generation computer this was certainly the case. In the last few years, however, the cost of storage has fallen abruptly, much more rapidly than computation costs, and the end of this development evidently is not yet in sight. Proponents of the storage of derived variables argue that it may already be cheaper, and in a few years will be much cheaper, to store large numbers of derived variables, however rarely used, than to compute them on demand.

At the Madrid conference there was no meeting of minds on this issue. The matter was not on the agenda, the whole discussion arose spontaneously, and firm cost estimates supporting neither view were available. They are readily obtained for the system RKNFSYS (Chayes, 1976; Chayes et al., 1977) which is in this respect probably fairly typical of petrographic information systems operating in batch mode on 3d generation computers.

In the structure of RKNFSYS the data base is treated like a reference work in a public library, and is available to users only in read-only mode. The place of the conventional reader is taken by a program module, FLBLDR, that scans the base, extracting or generating the desired information from its contents, and entering it in a file that is the analogue of the reader's notebook. Cost estimates given below are based on a comparison of charges for active mass storage of the base with charges for operating FLBLDR.

In the data base of RKNFSYS the only stored variables are the weight percentages of the essential oxides. Derived variables are computed at operation time, norms and molar amounts by subroutine CIPW, all others by

subroutine TRNEVL. (Before the scan of the data base starts, TRNEVL can interpret up to 20 linear combinations received in character format, and store the sequence of operators, operands and separators for each. During the actual scan, entry EVAL of TRNEVL generates the value of each such combination by processing the current specimen vector in accord with these stored instructions and specifications.)

The Current Costs of Computation

The work is done on a Univac 11040 in the University of Maryland computation facility. Under the present charge schedule, computer time is billed at $198 per hour, but in this "memory time" the actual or clock time is biased upward by an amount that depends on the core memory requirement of the operating program.

The cost estimates reported here are based on a series of executions in each of which the entire base of 16123 analyses was scanned. In each run a particular variable or set of variables was copied or calculated from each analysis and stored in a work file. The basis of each estimate of computation cost is the number of seconds of "memory time" the execution required, shown in column 2 of Table 1. Multiplication of each entry in column 2 by the ratio M/N, where M is the charge per memory second (in mills) and N is the number of analyses in the base, gives the entry in column 3, the estimated charge for performing the operation described in column 1 on one analysis. The relevant value of M/N is 55/16123 = 3.4113E-3.

The unit of column 3 is of course mills, or thousandths of a dollar, but for the present discussion may be thought of simply as a pointer on an arbitrary linear scale that permits meaningful comparison of computation charges, based on time biased upward for space, with storage charges, based on space per unit of time.

From Table 1 one may determine the cost of generating derived variables as opposed to retrieving them from storage. If, for instance, norms were stored in the base, the cost of retrieving one normative component would be .1812 mills. If the norm were generated rather than stored, this retrieval cost would be the same, so the cost of the actual run-time calculation, the expense that would be avoided if the norm were stored in the base, is .2672-.1812 = .086 mills. Numerical entries in Table 2 result from similar manipulations of the other rows in Table 1.

Much of the cost shown in each line of Table 2 is attributable to the subroutine call itself rather than to the calculations the subroutine performs. (In this connection, it is to be noted that calculation of the function 'DC', defined in the next to last row of the table, requires 2 subroutine calls per analysis, those in the rows above only one.) In rather similar fashion, most of the charge for copying a variable from core to work file is incurred in preparing to execute a Fortran "WRITE" order; the number of variables copied out by such an order matters very little, as may be seen by comparing the first two

rows of Table 1. Again, the norm subroutine is called only once (if at all) for each analysis; the costs of copying all or only one of the norm components to the work file differ very little. This holds also if normative components are used in other derived variables. For example, the cost of obtaining a norm component and the variable 'DC' jointly is very little more than the cost of obtaining 'DC' alone, as may be seen by comparing the last two lines of either table.

The Current Cost of Storage

At the installation in which RKNFSYS is housed the current storage charge is 2¢ per day per track of 1792 words. Derived petrographic variables are real and usually irrational, so each would be conveniently stored in a separate word. The charge per word per day is 20/1792 = .0112 mills, a very small fee. In each specimen vector, however, a place would have to be reserved for each derived variable, so that for the base of RKNFSYS the unit storage charge per day would be (16123)(.0112) = 180 mills. Bases more than twice as long as this are already in use in petrology, and the base under development by IGCP project 163 is expected to be at least 6 times as long. For the purposes of this discussion, however, the unit daily charge for storage of a derived variable will be taken as 180 mills.

Should Norms be Computed at Operation Time or Stored in the Base?

In the standard CIPW norm there are 29 possible components (see for instance Holmes, 1921, p. 411) and if simplicity of file structure is to be the ruling factor, each specimen vector in the base must include 29 words, one for each potential norm component. At current rates the daily charge for storing norms in a base the length of that attached to RKNFSYS would thus be (29)(180) = 5220 mills, an amount for which, also at current rates, more than 60,000 norms could be computed at operation time.

In oversaturated rocks there are rarely more than 8 non-null normative parameters per analysis. In undersaturated rocks, and especially in alkaline ones, there may sometimes be twice as many. If file structure were adjusted to permit exclusion of null components, the average number of normative parameters to be stored per analysis would probably not much exceed 10. This would reduce the daily storage charge to about 1800 mills, an amount for which one could still compute nearly 21,000 norms, several times more than have ever been requested in a single day of routine activity of the system.

Under current operating conditions, storage of norm components in the base of information system RKNFSYS would thus be far more costly than computing them as needed.

Should Other Derived Variables be Computed as Needed or Stored in the Base?

At first glance, the argument for storage of an individual derived variable seems much more favorable. The daily storage charge for extending the specimen vector by one word for each specimen in the current base of RKNFSYS would be only 180 mills per day, the equivalent, for example, of less than 900 calculations of the function 'DC'. Even if this many values of 'DC' were never required in the course of a day's work, it could surely be argued that 18¢ daily is an affordable extravagance, considering that keeping the base in active storage for a day already costs $2.30.

There are so many derived variables in use in petrology, however, that in practice the balance tips sharply in the other direction. If we are going to store CIPW normative components--whether 10 or 29 of them--in the base, what justification is there for excluding the Niggli numbers si, al, fm, c, alk, ti, p, k, mg and w? Or the somewhat different ACF coordinates of Osann, Eskola and Tilley? Or the LFM coordinates of von Wolff? Or the oxygen equivalents of Barth? Or the 'differentiation' indices of Yoder-Tilley and Thornton-Tuttle? Or the 'characteristic numbers' of Zavaritski? Or the τ and δ indices of Gottini and Rittmann? Or the Larsen variable? Obviously, there is none.

Every petrologist will realize that this list of derived variables is far from exhaustive. Yet if only these were included, the space required for a specimen vector would be increased by a factor of more than 5. The daily charge for storing derived variables would then be nearly $13, enough to pay for the object time computation of 150,000 norms or 63,000 values of 'DC', many times more of each than has ever been required in a day of routine operation.

Cost of Storing the Capability to Compute Derived Variables at Operation Time

If derived variables are to be computed at operation time, then whenever the base is accessible to a user he must also have access to the programs that compute them. In system RKNFSYS, as noted above, these are the subroutines CIPW and TRNEVL, attached to program FLBLDR. The first occupies 1066 and the second 2465 words of storage. The current daily storage charge for these 3531 words is 39.5 mills, less than a fourth the charge that would be incurred by adding one word--the capacity to store just one derived variable--to each specimen vector of the current base.

Conclusion

At the present charge schedule it would be grossly uneconomic to store derived variables in the base of system RKNFSYS. The margin in favor of computing only such derived variables as are specifically requested at object time is so broad as to suggest that the (economic)

preference for object time calculation will persist until the ratio of storage cost to computation cost undergoes further drastic decrease or the level of usage increases by orders of magnitude. Although detailed cost estimates will no doubt vary somewhat with time, hardware and monitor, the same overall relation between storage and run-time calculation of derived variables probably holds for work done by systems like RKNFSYS on any 3d generation computer.

What bearing does this result have on the design of the IGBA base? Even though the IGBA information system will have to be far more complex and sophisticated than RKNFSYS, it will be subject to the same trade off between computation and storage charges; any requested derived variable not stored in its base will have to be computed at operation time. For two reasons the margin in favor of computation at operation time will probably be even more pronounced for IGBA than for RKNFSYS:

(1) The IGBA base will ultimately contain several times as many analyses as that of RKNFSYS and the cost of storing derived variables varies directly with, but the cost of computing them is unaffected by, the number of analyses in the base.

(2) The program modules of RKNFSYS that compute derived variables-namely, subroutines CIPW and TRNEVL--are not particularly efficient. There is every reason to suppose that the analogous IGBA modules will be considerably superior in performance.

To justify storage of derived variables, the decrease of storage versus computation charges would have to be several times greater for the IGBA information system than for RKNFSYS. For the near- and mid-future, I believe we should plan to compute derived variables at operation time rather than store them in the IGBA base.

* * *

If you have experimental data bearing on this subject that you would like to bring to the attention of the project, send in a note that can be distributed as a project circular.

Felix Chayes, Chairman
IGCP-163-IGBA

References

Chayes, F. (1976). Version NTRM2 of System RKNFSYS. Unpublished, available from the author.

Chayes, F., McCammon, K., Trochimczyk, J., and Velde, D., (1977). A new edition (RKOC76) of the data base of the rock information system RKNFSYS. CIW Yearbook 76, p. 635.

Holmes, A. (1921). Petrographic methods and calculations. T. Murby & Co., p. 411.

TABLE 1. Total Time and Unit Costs for Some Retrievals and Generations Based on Complete Scans.

Operation	Memory seconds for 16123 replications	Overall unit cost*, mills/analysis
Retrieve a stored variable	53.115	.1812
Retrieve 10 stored variables	62.014	.2115
Generate a norm and retrieve one norm component	78.339	.2672
Generate and retrieve the sum of 2 stored variables	69.247	.2362
Generate and retrieve the sum of 9 stored variables	91.149	.3109
Generate and retrieve DC=100(HY+.130L)/(HY+OL+DI)-26.14**	113.373	.3867
Generate and retrieve DC and a norm component	118.123	.4030

*For rows 3-7, inc., the cost estimate includes reading of the base, subroutine call(s), calculation of variable(s), and copying of calculated variable(s) to the work file. For rows 1 and 2 it includes only reading of the base and copying of stored variables to the work file.

**'DC' is used here because it provides a rather severe test of the function generator; following a call to CIPW, TRNEVL must perform 3 additions, a subtraction, 2 multiplications and a division, as well as evaluate two parenthetic phrases.

TABLE 2 - Estimated cost of generating derived variables of Table 1 at run time

Variables requested	*Cost in mills per analysis
CIPW norm	.0860
Sum of 2 stored variables	.0550
Sum of 9 stored variables	.1297
DC=100(HY+.130OL)/(HY+DI+OL)-26.14	.2005
DC and a norm component	.2218

*Includes only call to and calculation within subroutines CIPW and/or TRNEVL.