February 17, 1972

TO: Distribution

FROM: T. C. Rindfleisch

SUBJECT: LONG TERM DENDRAL PROJECT COMPUTER SUPPORT PLANNING

REFERENCES: 1) Hundley, Smith, and Stefik, "Report on Alternatives in Meeting the Mass Spectroscopy Computing Requirements", December 1971.

2) Buchanan, "Notes on DENDRAL Computing Requirements", November 16, 1971.

3) Stefik and Reynolds, "A Plan for Computations for Mass Spectroscopy Using Mini Computers", September 1971.

This memorandum summarizes long term DENDRAL project computing requirements as currently understood and outlines a number of alternatives for meeting these needs. Motivated by the immediate need to plan a redirection of the ACME grant (to take effect within 18 months), as well as by the lead time required to design and implement an adequate developmental closed loop mass spectrometer system, it is necessary that the following attempt at overall planning be reviewed and iterated soon with the DENDRAL community. Agreement must be reached on the design objectives to be achieved as well as on a general approach for implementation in time for a May 1972 grant application deadline.

This DENDRAL planning effort is coupled to the parallel computer planning study within the Medical Center attempting to define overall long term medical computing needs including ACME alternatives. The DENDRAL project has provided a sizable portion of the ACME support base on the one hand but has growing computing needs on the other which will soon, if they have not already, outgrow ACME capabilities. In the context of the Medical Center study, this plan addresses only the technical question of DENDRAL requirements and the necessary computing capacity to meet these requirements. The technical implications of embedding these capabilities within a larger Medical Center computing solution are discussed briefly but the administrative problems of considering Medical School computing alternatives with and without DENDRAL financial support are left to a separate discussion.

The discussion of DENDRAL computing support is divided into five sections as follows: 1) Overall Support Requirements, 2) Computing Requirements, 3) Design Philosophy, 4) Possible Machine Configurations, and 5) Conclusions and Required Action. The time scale considered in this planning effort assumes the design and implementation in one to two years of an extendable computing capability which will meet DENDRAL needs over the two to five year period. It is clear that projected requirements over such a time scale for a research project like DENDRAL are subject to considerable error. This fact compels as a major design criterion the ability to modify

the DENDRAL computing support base flexibly and with minimal impact to on-going activities as requirements dictate.

I. OVERALL SUPPORT REQUIREMENTS

The overall structure of DENDRAL computing needs derives from the project objectives to 1) conceive, design, and implement an automated, closed loop mass spectrum interpretation system and 2) provide a reliable mass spectrometer instrument control and data analysis system for chemical experimentation, incorporating state-of-the-art spectrum interpretation capabilities. These two objectives are mutually necessary and complementary since the improvement of spectrum interpretation capabilities benefits on-going chemistry experimentation and the results from these experiments provide data for the heuristic development of spectrum analysis algorithms. The various subsystem elements necessary for such a capability are shown in the diagram of Figure 1.

The basic function of the mass spectrometer is to fragment and ionize sample molecules and, through electromagnetic separation, to measure the abundance of fragments with different masses. These abundances are strongly related to the molecular structure of the sample material and these relationships can be used by inference to derive the structures for unknown sample materials from their mass spectra. There are numerous modes of operation of the instrument which allow the measurement of ion abundances with varying time, mass, resolution, and ionization energy, as well as enable the observation of delayed or metastable ion fragmentation pathways. Not all information in all modes of operation can be collected because of data rate, instrument sensitivity, and sample volume limitations and indeed not all collectable information is necessary for the unique interpretation of source structure. The optimum set of information producing the most unambiguous interpretation in the shortest time is not predictable however for an unknown material. Thus closed loop computer analysis of spectrometer output data with subsequent feedback control of spectrometer operation is necessary to maximize collected data quality and to ensure the collection of needed information for structure interpretation.

The various elements in such a closed loop system (see Figure 1) are as follows.

1. DATA ACQUISITION AND DETECTION: This loop element accepts the high rate raw data output of the mass spectrometer, extracts the significant peak information above a background threshold, and encodes the resulting peak profile information for subsequent processing. Since the ratio of peak profile sample points to background sample points is typically low (< several percent), this extraction process can be used to direct the instrument scan to concentrate on the peak portions of the spectrum thereby decreasing the overall spectrum read-out time or improving the ion count statistics (shot noise variations) by longer integration times.

2. INFORMATION EXTRACTION AND REDUCTION: This loop element accepts peak profile data and through calibrated peak shape information, separates overlapping peaks, measures their amplitudes and positions in time, and attaches uncertainties to these measurements based on instrument performance and ion statistics. The resulting peak locations in time are converted to equivalent mass values by applying an instrument calibration model derived from observing the locations in time of peaks of known masses from an appropriate reference compound. This level of analysis verifies the physical performance of the mass spectrometer and can feed back control information to optimize such parameters as resolution, sensitivity, and reference compound pressure.

3. INFORMATION ANALYSIS AND INTERPRETATION: This loop element converts the physical measurements of fragment abundance as a function of mass into chemical composition or structure information. For high resolution spectrum data, the possible combinations of chemical elements and their isotopes which produce the observed peak masses are enumerated. At this point, a higher level search for a structural explanation of the observed spectrum is begun. Based on whatever a priori information is available, a library search through spectra of compounds of known structure may be performed or a more fundamental "theoretical" explanation of the observed spectrum may be constructed based on heuristic rules for the behavior of various molecular structural elements in different chemical situations. A given approach for explaining the spectrum in terms of structure is evaluated in terms of such parameters as time consumed, ambiguity, likelihood of convergence, and the accuracy with which the library or theory for a given class of compounds can explain the observed spectrum.

4. ANALYSIS PERFORMANCE EVALUATION AND CONTROL: This loop element directs the search for an explanation of the observed spectrum by using the available a priori chemistry information and evaluations of the on-going library searches and theoretical constructions. Based on this information, the currently pursued analysis approach is continued or another approach initiated. When ambiguities arise, control information is directed to the preceding system elements to obtain appropriate additional data to resolve the problem. If no solution is found within reasonable bounds, external help is obtained and used to extend the system capabilities by incorporating the new solution and its generalizations.

5. ANALYSIS UPGRADE AND EXTENSION: When new solutions outside of existing library or theory capabilities are encountered, this loop element incorporates these data into the system thereby continually improving system performance. This element draws information from the existing spectrum information store as well as using the new data to abstract theory rules or to expand the library. Help may be obtained from chemists to properly assess the significance and validity of inferred system extentions.

6. RESULT AND SYSTEM STATUS DISPLAY: This loop element provides the chemist user of the system with rapid volatile plots and displays of the various experiment results and system status

information and on demand produces hardcopy displays. Also displays of previously obtained data may be redisplayed for comparison with on-going analyses.

7. INSTRUMENT CONTROL: This loop element locally coordinates and implements the various control requests on instrument performance such as parameter adjustment or mode change by planning and issuing the appropriate electronic commands. Conflicting requests from the various system elements are resolved through the system coordination element.

8. SYSTEM COORDINATION AND CONTROL: This loop element receives and maintains the operational status and performance data relating to various system elements and guarantees the appropriate sequencing of interdependent operations. This element also arbitrates conflicting system and instrument control requests through a priority hierarchy designed into the system and coordinates system operation changes commanded from the outside.

9. COMMAND INTERFACE: This loop element decodes commands and information received through the instrument operator or chemist user interface. The decoder is designed to make the human interface with the system highly flexible and convenient.

10. INFORMATION STORAGE AND MANAGEMENT: This element includes the organization and storage of large cumulative amounts of spectral information and the ability to access this data on demand. Access is provided to other on-line system elements requiring calibration data or library information as well as to external users. Facilities are provided for the retrieval and correlation of data store elements based on varied controllable descriptors.

There are four major groups of computer users within the DENDRAL project involved in the development and operation of such a mass spectrum analysis system. Each requires varying types of computing support which the overall facility must provide. These classes of users are summarized as follows.

1. ARTIFICIAL INTELLIGENCE RESEARCH: This class of user requires computing support for the design and implementation of evolving mass spectrum interpretation software. Efforts cover a range of activities from the development of new programs such as for theory formation, problem solving strategy planning, and cyclic structure generation to the extention of existing program capabilities and their incorporation into experimental closed loop mass spectrum interpretation systems. Computing needs are characterized by:

a) Fast turn-around interactive services for program coding, debugging, and testing.

b) Fast turn-around batch execution of programs for testing and experimentation.

c) High quality list processing language, test editting and other system service support.

d) Large core storage, large secondary storage, and high processor speed for program execution.

e) Large data base access and management facilities for spectrum file correlation.

2. INSTRUMENTATION RESEARCH: This class of user requires computing support for the development of an integrated closed loop spectrometer system as well as new instrument capabilities. Activities include the development of reliable computer methods for spectrometer interface and control, and for spectral information extraction and reduction. Additional activities include the incorporation of versions of the artificial intelligence programs suitable for real time closed loop operation, and the extention of instrument facilities such as for scan control around mass peaks and metastable ion decay analysis. Computing requirements are characterized by:

a) Flexibility in hardware and software interface capabilities between the computer and mass spectrometer.

b) Fast turn-around interactive services for program coding, debugging, and testing.

c) Fast turn-around batch execution of programs for testing, simulation, and experimentation.

d) Real time computer support of instrument data acquisition and control feedback testing.

e) High quality FORTRAN or PL-1 and Assembly languages, text editting, and other system utility support.

f) Large core storage, large secondary storage, and high processor speed for program execution.

3. CHEMISTRY RESEARCH: This class of user will require computer support for accessing and utilizing the cumulative data base from mass spectrum experimentation for subsequent chemical analyses. Activities will include the development and application of programs for analyses such as the correlation of molecular structure with other chemical, physical, and biological properties as well as the planning of mass spectrometer experiments and new compound synthesis strategies. Computing requirements will be characterized by:

a) Fast turn-around interactive and batch services for program coding, debugging, testing, and operation.

b) Large core storage and large secondary storage.

c) High quality FORTRAN or PL-1 and list processing language, text editting, and other system utility support.

d) Large data base management and access facilities.

4. ROUTINE MASS SPECTROMETER OPERATIONS: This class of user requires computing support for the day to day operation of the various mass spectrometers. Activities include real time acquisition, reduction, and analysis of mass spectral data utilizing state-of-the-art analysis system capabilities on a large daily volume basis. Closed loop on-line operation of the mass spectrometers will become standard even before automatic spectrum interpretation programs broadly rival human performance, in order to maximize instrument data quality. Computing requirements are characterized by:

a) Fast turn-around real time computer support of instrument data acquisition and analysis at high data rates and subsequent control feedback.

b) Large core storage, large secondary storage, and high processor speed.

c) Highly reliable hardware and software operation with the ability to service several instruments simultaneously.

d) Fast turn-around system status and spectrum analysis result display in volatile and hardcopy form.

e) Convenient and efficient interfaces for the instrument operator and chemist to control the computer/spectrometer system as well as experiment parameters.

II. COMPUTING REQUIREMENTS

The four classes of DENDRAL users mentioned above, place demands on a computer system which can be grouped into two large categories: 1) fluctuating developmental and experimental activities and 2) on-line operational support of mass spectrometer experiments. The characteristics which distinguish these two types of support are in the first instance the need for extensive individual program debugging and text editting facilities in an environment allowing rapid program experimentation turn-around, and in the second instance the real time commitment of computing resources to operate the fully integrated software system in coordination with the mass spectrometer instrumentation. In either case, the individual program requirements in terms of machine resources are comparable. The overall machine resource and response time requirements differ significantly.

Because of the complexity of software and hardware elements in each of these categories, one of the best guides to projecting computing needs is on the basis of performance of existing programs on existing computing hardware coupled with estimates of the effects of

anticipated modifications. The benchmarks which have been performed to date and previous summaries of requirements appear in references 1 through 3. The following outline of computing requirements extracts from these memos as well as projects toward longer term needs. The overall relationship between these requirements is based on the development of the various loop elements shown in Figure 1.

For existing mass spectrometer instrumentation, the required data acquisition rates, result display rates, and control loop time constants are shown in Table 1. It is emphasized that the experiment objectives leading to the numbers in Table 1 are not the final interpretation of output data in real time but rather the ability to guarantee in real time the collection of the information essential to that interpretation. Subsequent completion of the interpretation is assumed to occur within a time scale on the order of or several times longer than the overall experiment duration. The utility of complete data interpretation in real time has not been demonstrated at this time and would place unreasonably great requirements on computer throughput capacity.

These performance requirements are based on the typical duration of effluent uniformity from gas chromatograph driven experiments and the duration of single samples where the source is other than the gas chromatograph. An additional operating mode used to observe the decay pathways of metastable ions will also be used but required data rates and other control parameters are not presently well known. The metastable mode will require much lower data rates than those shown in Table 1 however, so that this mode is not expected to be a determining factor in system throughput requirements. At any one time it can be expected that two instruments will be generating data simultaneously; one with a gas chromatograph source and one without. Thus the most severe set of constraints occurs when the system must support simultaneous high resolution spectroscopy in the two instruments.

The mass spectrometer data system that exists today does not meet the requirements in Table 1 for a variety of reasons. The existing programs do not yet support real time closed loop operation, do not perform all of the data stream processing requirements , and do not operate reliably at the indicated high data rates. The existing programs run on a variety of machines including a PDP-11/20 for data acquisition; a 360/50 for information extraction and reduction as well as for elemental composition analyses and data display; and a 360/67 for the developmental spectrum interpretation software. Table 2 shows a gross comparison of existing capabilities against long term requirements expressed in terms of computer throughput. These numbers are of course oversimplifications but give approximate measures of where improvements are required. It must be noted in all of this discussion that the projected computing needs for meeting long term project goals are subject to considerable error since in most cases the necessary algorithms are not designed, coded, or benchmarked. Considerable work in this area is necessary, particularly as regards the throughput improvements which can be realized in the existing LISP coded artificial intelligence programs. Little effort has been spent to make these programs efficient in the LISP language or to examine the utility of some other language. It should also be noted that some of these processing steps can be overlapped and others must be

serialized so that care must be exercized in straightforwardly adding the times shown in Table 2. A more explicit discussion of process sequencing appears later.

The following gives more specific measures of current throughput and operating parameters as well as estimates of improvements required to meet the overall system goals.

1. DATA ACQUISITION AND DETECTION

A. Current or anticipated functions

i) Raw data stream acquisition and buffering

ii) Partially adaptive threshold peak detection

iii) Run length compression of data stream for subsequent processing

B. Required additional functions

i) Adaptive threshold peak detection accommodating variable instrument background and broad metastable background peaks

ii) Active scan control to force data collection around spectral peaks with superposition of multiple local scans

iii) Failsafe raw data stream processing and storage so that data is not lost if downstream processing fails

C. Current performance parameters

i) These programs currently run on a PDP-11/20 computer with 4K words (16 bit) of core, no disk, and are written in Assembly language. The ACME 360/50 is used to file the compressed data and provides assembly and text editting support for the PDP-11 programs. The filing programs on the 360/50 require 100K bytes of core, 50K bytes of disk per spectrum and are written in PL/ACME.

ii) This system currently processes a 10KC raw data stream containing less than 5% significant data with occasional overruns. This rate is limited in part by the time share environment and data channel limitations between the PDP-11 and ACME as well as by the lack of a direct memory access (DMA) device for transferring data in and out of the PDP-11. Ignoring the ACME limitations and with improved program coding, the existing system could process as an upper bound up to a 20KC raw data stream containing no more than 5% significant data. With DMA input/output, this upper bound would approach 30-50KC. These latter bounds, with and without DMA capability, assume nearly 100% machine cycle utilization and no statistical fluctuations.

D. Required upgrading

i) The reliable processing of 30-50KC raw data streams requires increased core buffer space, faster processor speed, extended hardware arithmetic capabilities, and overflow external storage.

ii) The addition of the functions in (1.B) require increased core, significantly faster processor speed, and extended arithmetic capability.

iii) The minimization of core requirements and the flexibility to process a data stream inherently of unknown length in the time-shared downstream processor (currently 360/50), requires more efficient input/output programs allowing interrupt controlled buffer manipulation and overlapped processing.

iv) Estimated computing addition requirements

a) Increase core by a factor of 2 to 4

b) Increase processing speed by a factor of 4 to 8

c) Add local disk storage

d) Add extended hardware arithmetic capability

e) Guarantee a continuous high rate data path between data acquisition and downstream processing

f) Extend the small machine programming facilities to higher level languages using associated large machine resources

2. INFORMATION EXTRACTION AND REDUCTION

A. Current or anticipated functions

i) Measure thresholded peak amplitudes and positions

ii) Determine instrument scan calibration from reference compound peak locations in high resolution spectra or from mass defect amplitudes in low resolution spectra

iii) Convert sample peak positions in time to equivalent mass

B. Required additional functions

i) Resolve adjacent peak multiplets (above threshold) into component amplitudes and locations

ii)Assign uncertainty estimates to measured peak amplitudes and positions

iii) Derive instrument performance measurements to set up and maintain optimum parameter settings (resolution, sensitivity, etc.)

C. Current performance parameters

i) These programs currently run on the ACME 360/50 in approximately 25-50K bytes of core, using approximately 50K bytes of disk per spectrum, and are written in PL/ACME

ii) The current programs perform functions (2.A) for high resolution spectra in approximately 36 seconds wall clock time on a dry machine (no other users). Besides system overhead, this time is equivalent to CPU time since no input/output processing time is included. The location and identification of reference compound peaks required for overall mass calibration is somewhat unreliable. With other users in a time shared (equal priority) mode, this time increases by a factor of 5 to 10. A comparable processing time is anticipated for low resolution spectra although this capability does not presently exist except in a very old version. By improving the coding of the algorithms in non-interactive form and making suitable use of Assembly language subroutines, this time could be reduced overall by approximately a factor of 4. Part of this improvement will be offset by necessary increased coding complexity to improve reliability yielding a net short term improvement factor of 2 to 3. This net improvement in performance for this process on the 360/50 (single user) would only bring the running time down to between 15 and 20 seconds.

D. Required upgrading

i) The processor speed must be increased significantly even with more efficient coding and assuming no other users on the machine.

ii) In a large machine environment which meets the processor speed requirements, the allocation of resources to multiple users must be on a priority basis for the support of a real time operation.

iii) Flexibility of language choice and object module configuration must be present to allow easy debugging in an interactive environment and also efficient execution after debugging in an experimental or operational environment.

iv) Estimated computing addition requirements

a) Increase the processor speed by a factor of 4 to 8

b) Provide increased system program flexibility allowing interchange between time share, batch, and real time standards of machine resource allocation

3. INFORMATION ANALYSIS AND INTERPRETATION - ELEMENTAL COMPOSITION

A. Current or anticipated functions

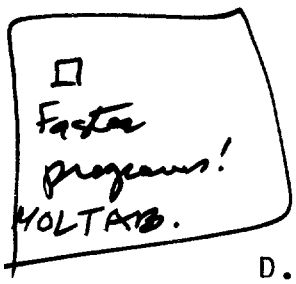
i) For high resolution spectral data, enumerate possible elemental combinations resulting in the observed peak mass within a given fixed error and within prescribed element abundance limits

B. Required additional functions

i) Utilize confidence estimates based on instrument performance to assign data dependent error limits in determining elemental compositions

C. Current performance parameters

i) These programs currently run on the ACME 360/50 in approximately 50K bytes of core, using approximately 10 to 20K bytes of disk per spectrum, and are written in PL/ACME



ii) The current programs perform functions (3.A) in approximately 5 seconds wall clock time on a dry machine. Improved coding in non-interactive form and making suitable use of Assembly language subroutines can result in an improvement by a factor of 1 to 2.

D. Required upgrading

i) Upgrading similar to (2.D) is required except that processor speed increase requirements are in the range of a factor of 1 to 2.

4. INFORMATION ANALYSIS AND INTERPRETATION - ARTIFICIAL INTELLIGENCE (including Spectrum Interpretation, Performance Evaluation and Control, and Analysis Function Extention)

A. Current or anticipated functions

i) Enumerate possible topological structures for a given molecular formula. Current capabilities are limited to acyclic structures but on-going modifications will include cyclic configurations.

ii) Within restricted classes of compounds (eg alkanes and recently estrogens) use heuristic molecular fragmentation rules applied to enumerated possible structures to obtain the best explanation of the observed spectrum. The problem solving strategy is guided by spectral content.

iii) Preliminary machine abstraction of theory rules from sets of spectral data

B. Required additional functions

i) Evaluate on-going analysis performance and prognosis for solution to guide additional information collection, to allow selection of the most effective problem solving strategy, and to recognize failure in order to take corrective action.

ii) Sophisticated machine extention of problem solving strategy planning, heuristic theory rules relating molecular structure to mass spectrum composition, and library search capabilities.

C. Current performance parameters

i) The current programs run on the SCC 360/67 in approximately 300K bytes of core, using approximately 1M bytes of disk, and are written in LISP.

ii) Existing benchmarks on the 360/67 for the system of programs used for estrogen structure analysis indicate that up to a total of 1 to 5 minutes and as little as 10 to 20 seconds of CPU time are required, depending on the complexity of the analysis. These times include all phases of the processes involved. No attempt has been made at this time to code the LISP programs efficiently so it can be expected that these times could be reduced by a factor of 2 or more by more careful coding. It is estimated that the parameters which are necessary for instrument control and guiding additional information collection can be available in from 5 to 15 seconds after beginning the interpretation processing. This processing cannot start however until the high mass peaks of the spectrum are available to determine the molecular ion mass.

iii) Early versions of the structure generator programs were written to run on the Artificial Intelligence Project PDP-10 computer under a time sharing environment. These programs required approximately 50K words (36 bits) of core and variable running times depending on machine usage and the complexity of the run.

iv) Benchmarks have also been attempted on the ACME 360/50 under an interactive version of LISP with a very great increase in running time (> 10 times). More realistic benchmarks using the equivalent batch version of LISP run on the 360/67 will be attempted. Based on a comparison of functional characteristics, one should expect at least a factor of 4 degradation in performance.

v) No reliable estimates exist on the running efficiency improvements possible by coding in another language such as Assembly language.

D. Required upgrading

i) The above running times (CPU times) approximate those required for the control aspects of closed loop operation,

at least within the restricted class of compounds now considered. The total analysis time however can easily exceed the time between gas chromatograph peaks (approximately 30 seconds) and thus final results would not be available in near real time. Such near real time completion of sample interpretation is not reasonably required at present. As the generality of the programs increases, the requirements for computing speed and core size will increase easily by several factors of 2 in order to maintain reliable control feedback within experiment time constraints.

ii) Same as (2.D.ii)

iii) Same as (2.D.iii)

iv) A processor of at least comparable speed to the 360/67 is required in the near term to service real time operations.

D – new hard wired programs for control loop purposes.

5. COORDINATION, CONTROL, INTERFACE, DISPLAY, AND DATA MANAGEMENT

A. Current or anticipated functions

i) Preliminary printout and graphical displays

ii) Preliminary large volume archival spectrum storage capability

B. Required functions

i) Real time loop coordination and control

ii) Instrument function control

iii) Flexible high speed result display and interaction

iv) On-line large volume spectrum storage and accession by user-defined descriptors.

C. Current performance parameters

i) The current programs run on the ACME 360/50 computer in 20 to 50K bytes of core, use 10 to 100M bytes of disk total, and are written in PL/ACME and Assembly language. The IBM 1800 computer is used to drive on-line graphical displays.

ii) The currently available display and file management programs essential to the real time loop and have satisfactory running times. CRT plots can be generated in a few seconds on a dry machine. This performance is degraded to 5 to 10 seconds in a busy time sharing environment. The use and support of volatile displays in general are fairly primitive however, in terms of facility

of user interaction and must be improved.

D. Required upgrading

i) A more flexible capability for using volatile displays for printed as well as graphical material including subsequent convenient user interactions via light pen or function keys must be developed.

ii) The development of coordination, control, and other interface programs is yet to be done. It does not appear that these programs will approach the computing requirements of the other loop elements described earlier. Thus given reasonable resources meeting those needs, it is expected that the additional control and coordination function requirements will be met given good system software support of real time operations.

The above elements in the mass spectrum analysis loop operate interdependently since one element cannot process the output of a previous element until at least a part of it exists and certain elements cannot make significant progress until a sizable fraction of the overall spectrum data is available. This interdependence is shown approximately in Figure 2 for an overall scan time which is between the low resolution and high resolution requirements shown in Table 1. The first three operations; data acquisition, information extraction and reduction, and elemental composition determination (for high resolution data only), can proceed nearly in parallel since they perform operations on local portions of the spectrum only. The interpretation aspects of spectral analysis, however, require operations on larger portions of the spectrum and may in fact (such as currently implemented) be dependent on information available only toward the end of a scan such as the molecular ion mass. The above relationships assume that the current practice of scanning from high mass to low mass is reversed and data are available starting at low mass values. This is necessary because the instrument scan calibration and associated data reduction processes can only be performed starting at the low mass end of the spectrum and working up. The essential point to be made is that processing times must be added taking into account the delays inherent in beginning some of the processing functions. This forces the overlap relationships shown in Figure 2 where the control information coming out of the early steps in spectrum interpretation allow the collection of additional information on a succeeding scan in parallel with the further interpretive processing of the scan data just completed. Note that the times shown marginally allow several scans in a gas chromatograph peak time (approximately 30 seconds) and assume no appreciable delay beyond normal magnet retrace time in setting up the instrument for a different mode such as converting from high resolution mode to metastable mode.

The expected overall loading of the DENDRAL computing resources including developmental and operational aspects are summarized below.

1. DEVELOPMENTAL COMPUTING: Over the near term the number of active programmers includes

Computer Science	5
Instrumentation Research	3
Chemistry	2
TOTAL	10

This number can be expected to grow slowly over the next few years to approximately 15.

Usage will be primarily by terminal and will include a full spectrum of program coding, debugging, testing, and experimentation, both in and out of the real time environment. Based on gross estimates of current usage one might expect up to the equivalent of 5 people continuously using terminals. This load will occur during the 8 hour prime shift primarily. In addition, several hours of overnight batch processing can be expected on a regualr basis as is currently the practice.

Terminal response should allow text editting and interactive job entry conversations with characteristic turn-around times of approximately 1 second. Experimentation activities require that program performance in time-share and batch modes not be degraded by more than 50 to 75 percent.

2. OPERATIONAL COMPUTING: There currently exist three mass spectrometers which are expected to interface the DENDRAL computing system. Of these, one (the MAT-711) can run in the gas chromatograph driven mode which requires the analysis of a time sequence of different materials, each one lasting from 20 to 40 seconds and the entire experiment lasting from 1 to 2 hours. The other two instruments (the MS-9 and CH-4) operate with a single compound or simple mixture source which can last for several minutes.

Based on the predicted requirements for running chemistry experiments on the mass spectrometers, a peak load capability to support two spectrometers, the MAT-711 plus either the MS-9 or the CH-4, simultaneously is necessary. This load will be somewhat sporadic depending on instrument down time and experiment loading. In routine operation up to 3 gas chromatograph experiments and 20 single sample experiments will be run per day.

In the short term (approximately 2 years) the DENDRAL spectrum interpretation programs will be highly experimental with limited performance compared to human chemists. In this period it is expected that on-line computing support will be required for real time data acquisition, data reduction, elemental composition determination, and primitive instrument control, with subsequent non-real time computer aided human interpretation of results. The development and extension of on-line artificial intelligence and sophisticated instrument control capabilities require provision for fully automated operation of only one machine at a time during this era, with the remaining instrument able to operate simultaneously in the semi-automated mode. Much of the on-line artificial intelligence and instrumentation

experimentation can operate without real time turn-around commitment of computing resources.

As DENDRAL capabilities develop to rival human performance over the 2 to 5 year period, the capability for simultaneous fully automated DENDRAL support of more than one instrument should be provided.

III. DESIGN PHILOSOPHY

In addition to the obvious constraints of planning the necessary amount of computing support for the least money, the combination of research and operational aspects of the DENDRAL project make desirable certain additional design goals. These derive from the fluctuating nature of computing needs during various developmental and experimental phases of the work, the fact that DENDRAL progress will occur over a period of years during which time developments in complementary computer hardware and software fields will occur, and the fact that long term computing needs are at best a gross approximation since algorithms and system design elements will evolve in unforeseen ways. These factors lead to the following constraints.

1. The selected approach for implementing DENDRAL computing support should draw as much as possible upon externally supported technology both in computer hardware and system software areas.

2. The DENDRAL computing support should be expandable in terms of computer hardware and with upward software compatibility as the need arises.

3. As relevant future developments in processors, peripherals, languages, and system support capabilities take place, both within industry and within related academic projects (such as artificial intelligence work), these should be readily incorporatable into the DENDRAL computing support system.

4. It must be expected that DENDRAL computing needs will overflow the DENDRAL specific computing capacity from time to time. It should be possible to obtain overflow computing support from hardware and software compatible facilities, either on campus or through a network such as the ARPA network.

5. The hardware and software system should be designed emphasizing modularity so that as system elements evolve, they can be modified and reincorporated into the system without redesigning large portions of the overall system.

IV. POSSIBLE MACHINE CONFIGURATIONS

In considering the possible ways of meeting the DENDRAL computing requirements within the above philosophy, it is clear that the dominant factors in scaling the overall system derive from the real

time experimentation and operation of the integrated mass spectrum analysis software system. Development activities on individual programs and subsystem elements require neither comparable overall system size nor response time commitments. The following discussion is therefore organized in terms of first defining a configuration which meets the fully integrated system requirements and then examines the impact of including development requirements as well.

A spectrum of machine configurations is possible ranging from performing the entire real time task with a single processor to performing the task with a series of very small machines each performing some small specialized aspect of the job. Both of these extremes appear undesirable while a compromise approach combining many of the benefits of large machine capacity and facility with small machine speed and economics, provides the flexibility and performance required within a reasonable cost.

The large central processor approach forces the effective serialization of all operations (even though in practice they may be interleaved) since only one instruction at a time can be executed. Based on the processing times summarized in Table 2 for existing programs and the estimated net improvements required, the central processor would have to be on the order of 1 to 2 times as capable as the 360/67 to meet short term requirements. Note that these times in Table 2 are all essentially measures of CPU time requirements. This estimate is very optimistic in assuming that current artificial intelligence programs will not increase their machine requirements significantly. Using the gross relationship between processing speeds of typical existing computing hardware shown in Figure 3, a machine in the range of an IBM 370/155 to an IBM 360/75 would be required. Although no explicit benchmarks have been run, it is assumed that the performance of the 370/155 will be on the low end of the range indicated in Figure 3 since it derives much of its speed from a relatively large and fast cache memory which will be far less useful for LISP programs which have unpredictable addressing sequences.

This approach results in a very expensive piece of hardware (approximately $2 to 3M) used almost completely during real time operations and used sporadically during non-operations and off-shift hours. It is difficult to find additional compatible users of such a facility under conditions granting DENDRAL usage needed high priority, unless batch computing with no guarantee of turn-around can be sold. Furthermore if DENDRAL requirements outgrow the existing machine, an increase in capability necessitates a major change in hardware and significantly increases the cost.

These aspects of DENDRAL computing requirements make it undesirable to attempt to merge the needed high load real time computing capability with a central facility providing more casual interactive, batch, or low rate real time services to a large number of users. This affects in particular the desirability of merging DENDRAL computing support with general Medical Center computing for example.

Similarly the extreme of fragmenting the problem into many small subsets, each using a small inexpensive computer is currently

undesirable because of excessive overhead in intermachine communication and coordination and the difficulty, inconvenience, and high cost of initial programming as well as subsequent program modifications. The incorporation of results of on-going research on DENDRAL programs may easily affect the structure and interactions of significant numbers of small machines in such a network. Such modifications must be implementable with system support for the automated parsing and delegation of problem elements to individual machines. This kind of system does not presently exist.

Rather than these two extremes, a more flexible, convenient, and economical solution is possible which combines the advantages of each approach. The facility for parallel processing of moderately sized and well defined problem subsets can be implemented on relatively small machines which on the one hand, are inexpensive but on the other are large enough to allow easy design modification. This capability can be coupled to the advantages of a relatively large central machine which provides broad high performance for coordinating the small machines and for running programs in the system which are still under early development. The central machine also provides currently available sophisticated facilities for program development and modification on both the large and small machines.

The basic problem of real time mass spectrum analysis breaks into a number of natural elements as indicated in Figure 1. The criterion for parsing the problem into such elements is that of determining elements whose interactions (inputs and outputs) remain as independent as possible of the method chosen to implement each particular element. For example, the input to "Information Extraction and Reduction" is the raw peak data without background and its output is a set of peak amplitudes, masses, and associated errors. These data remain independent of the pattern recognition and instrument scan calibration techniques used inside the element. Feedback and control information does in part depend on this implementation but even here various functions are definable independent of implementation.

Furthermore since various loop elements emphasize different specific aspects of machine performance (input/output rate, core size, arithmetic speed, logical operations, addressing facility and speed, etc.), a processor can be selected for each element which on the one hand provides its needs in terms of throughput maximization and on the other is large enough that changes are relatively easily and flexibly made. Furthermore if an element outgrows its processor, up-grading is possible at reasonable cost.

The overall coordination and control of the various loop element processors involved would be performed by a relatively large central processor whose overall requirements are diminished by the processing requirements of loop elements now satisfied by satellite machines. The central machine performs in addition those functions which either inherently require the broad facilities of a large computer or which are not sufficiently developed to warrant the selection of a special satellite processor. The large machine facilities also allow easy development, modification, and operational loading of peripheral processor programs thereby minimizing this cost and allowing access to more sophisticated languages in programming the smaller satellite

machines.

Outgrowing existing computer needs does not necessarily imply modifying the large central processor but may only require increasing the size or number of peripheral machines. For example, it is clear that the broadening of spectrum interpretation capabilities will require more and more computing capacity. The problem of selecting the appropriate problem solving approach at various stages of the interpretation processing (library search, theory construction within some class of compounds, etc.) will be ambiguous to some extent and could be attacked by initiating at any given time several parallel attempts which appear equally reasonable on special processors and selecting the most promising path based on performance. This is contrasted to doubling or trebling the central processor performance capabilities to accomplish the same thing in the same time serially. The continuing effort and success at developing high performance, inexpensive "mini" computers in recent years and the future possibility of software modifiable microprogrammed machines promise increased flexibility in the selection of appropriate satellite machines.

A conceptual configuration of this type is shown in Figure 4. In the configuration shown, the data acquisition, data reduction, and elemental composition functions are shown as satellite processor functions. The current artificial intelligence programs would reside in the central processor along with overall coordination functions. Common access to secondary storage and to some blocks of main storage are provided to eliminate multiple movement of large amounts of data and to provide common access to instrument calibration parameters, etc. It is expected that in the near future (1 to 5 years) various aspects of the artificial intelligence software will be sufficiently developed to allow its operation in one or more satellite processors, thereby making the central processor available for newer developments.

The ability of such a facility to meet both research requirements and operational real time requirements is facilitated on the one hand by the lack, in general, of severe time constraints in development work but made more difficult on the other by the fact that the same programs or subsets thereof requiring relatively large core storage will be run. The reliable prediction of system loading due to development and that due to operations is impossible by nature since these needs fluctuate. A system which can comfortably (with 50 to 100 percent reserve capacity) meet anticipated operational needs in the near term will certainly have resources to allocate to development activities both from the reserve capability and from the fact that the operational usage will not be continuous. It is recalled that expected near term operational needs include processing data from two mass spectrometers in real time to the point of data reduction and preliminary data interpretation with interactive result display. Only occasionally will the full DENDRAL interpretation software system be run in the real time mode. It must be expected however that the joint requirements of the two activities will overflow such a facility and a backup source of computing must be found. This backup should be transparent to the user in that software will run in either environment without change and comparable program development and experimentation facilities will be available in either place. This

source of overflow computing would either be other Stanford facilities or a network such as that initiated by ARPA.

The choice of machine hardware is coupled both to the primary system design and to the question of overflow. On the one hand most work in artificial intelligence which would benefit the DENDRAL project is done in facilities which use DEC (PDP-10) computers. The best implementation of LISP exists on the PDP-10 machine and this machine is currently the primary source of computing on the ARPA network. On the other hand Stanford in general, including SLAC, is currently committed to IBM hardware and would appear to remain with IBM to avoid major conversion transients. IBM has an excellent hierarchical selection of tested large processors available although very little in the small machine line. An important aspect of system design is architecture homogeneity since the problems of running software in mixed FORTRAN, LISP, and Assembly language on differing sets of hardware or even differing system software on the same set of hardware are formidable.

Other manufacturers such as XDS, CDC, UNIVAC, etc. offer hardware which has advertized performance comparable to that of IBM and DEC equipment. In general, however, the status of hardware and software systems of the class meeting DENDRAL requirements lacks demonstrated reliability and experience as well as a large user community contributing to system extensions and improvements. There is serious doubt that any increased hardware performance or cost effectiveness exists to offset these gaps in system development status or potential cross fertilization from related efforts. Thus it appears that the choice of hardware for the DENDRAL computer facility must be between IBM and DEC.

This problem of hardware choice is in part technical in the sense of performance, available software, and cost, and in part administrative in the sense of commitments to internal compatibility within Stanford and dependence on outside facilities (ARPA or AMES) for overflow capability. A summary of currently identified pros and cons for IBM versus DEC hardware and software systems is shown in Table 3.

The estimated hardware cost of a facility such as shown in Figure 4 is very roughly in the range of $1 to 2M. This assumes a central processor of the class of a PDP-10 (KI-10) or a 370/155 with 100 to 200K words of core (approximately $1-2M) and several satellite processors of the class of the PDP-11/45 with 16-32K of high speed memory (each approximately $50-75K). In addition such a facility would require administrative, systems programming, and operations support personnel.

V. CONCLUSIONS AND REQUIRED ACTION

The following general conclusions are drawn from this review of long term DENDRAL requirements and existing program performance data.

1. The DENDRAL computing requirements place severe constraints on a facility in terms of developmental and real time loading which can be met only by dedicating a sizable facility to DENDRAL programming and operations support.

2. It is feasible to meet DENDRAL computing requirements within anticipated hardware and software system capabilities drawing largely upon existing technology.

3. The nature of the projected DENDRAL loading of such a facility makes it undesirable to embed the DENDRAL requirements in a larger general purpose facility such as the Medical Center, or SCC.

4. The best way of meeting DENDRAL needs is through a moderately sized central computer (of the order of an IBM 360/65 or 370/155, or a DEC KI-10) with multiple satellite processors performing subsystem functions in parallel and at high speed (requires on the order of DEC PDP-11/45 machines).

5. The choice of hardware for implementation is primarily between IBM and DEC. The choice depends on technical and administrative questions. From the technical point of view, DEC appears to be the best choice based upon currently projected hardware and software capabilities.

The actions which are required to follow up this initial planning effort are:

1. Review the stated requirements and ground rules of this study and incorporate any necessary additions, deletions, and modifications.

2. Examine in greater technical detail and refine the long term requirements in terms of algorithm designs and their impact on machine capabilities. This should include a specific effort to benchmark sample programs written efficiently in LISP and other languages to compare machine and language performance.

3. Examine in greater technical detail the possible hardware and software configurations which meet DENDRAL needs within the central/satellite machine concept and develop a more accurate cost estimate for such a facility.

4. Begin the administrative evaluation of the hardware manufacturer decision from the standpoint of long range Stanford commitments.

DISTRIBUTION:
B. Buchanan
C. Djerassi
A. Duffield
E. Feigenbaum
R. Jamtgaard
J. Lederberg
E. Levinthal
W. Reynolds
D. Smith
M. Stefik

TABLE I

COMPUTING SUPPORT REQUIREMENTS

	NON GAS CHROMATOGRAPH (Long duration Source)			GAS CHROMATOGRAPH (Short duration source)	
	LOW RESOLUTION	HIGH RESOLUTION	VERY HIGH RESOLUTION	LOW RESOLUTION	HIGH RESOLUTION
RESOLUTION (M/ΔM)	1500	10,000	30,000	1500	9000
SCAN TIME /DECADE	10 sec	15 sec	45 sec	3 sec	10 sec
MASS RANGE	1.5 decade	1.5 decade	1.5 decade	1.5 decade	1.5 decade
TOTAL SCAN TIME	15 sec	23 sec	68 sec	5 sec	15 sec
EXPERIMENT DURATION	1-5 min	1-5 min	5-10 min	1-2 hour	1-2 hour
# SPECTRA/EXPERIMENT	3-5	3-5	3-5	~1000	~200
# EXPERIMENTS/DAY	~20	~20	~20	~3	~3
RAW SAMPLE RATE *	5.2 KC	23 KC	23 KC	17 KC	81 KC
NON-BKGD SAMPLE RATE	1 KC	0.7 KC	0.2 KC	3 KC	1 KC
RAW DATA VOLUME	8×10^4 samp	5×10^5 samp	1.5×10^6 samp	8×10^4 samp	5×10^5 samp
NON-BKGD DATA VOLUME	1.5×10^4 samp	1.5×10^4 samp	1.5×10^4 samp	1.5×10^4 samp	1.5×10^4 samp
= 15 Samples 15 sam					
DWELL TIME / PEAK	3 msec	0.7 msec	0.7 msec	0.9 msec	0.5 msec
TIME BETWEEN PEAKS	15 msec	23 msec	68 msec	5 msec	15 msec
RESULT TIME (~~VOLATILE~~ CRT):					
ION CURRENT				<1 sec	<1 sec
MASS BAR PLOT	<30 sec	<1 min	<1 min	<3 sec	<15 sec
ELEMENTAL COMPOSITION		<2 min	<2 min		<15 sec
RESULT TIME (HARDCOPY):					
ION CURRENT				AFTER EXPT	AFTER EXPT
MASS BAR PLOT	<4 min	<4 min	<4 min	<1 min	<1 min
ELEMENTAL COMPOSITION		<5 min	<5 min		<1 min
CONTROL LOOP TIME CONSTANTS:					
INSTRUMENT MODE	~15 sec	~20 sec	~70 sec	~5 sec	~15 sec
INSTRUMENT PERFORM.	~1 sec	~1 sec	~5 sec	~.5 sec	~1 sec
PEAK DWELL	<6 msec	<1.5 msec	<1.5 msec	<2 msec	<1 msec

* Indicated data rates are in general limited by peak dwell time (ion integration time). Directed spectrometer scanning would make this integration much more efficient, thereby increasing the data rates to approximately 50-100 K samples per second and decreasing scan times uniformly to ~3 seconds

TABLE 2

	CURRENT COMPUTER	CURRENT THROUGHPUT	REQUIRED THROUGHPUT	SHORT TERM IMPROVEMENT FACTOR	ADDED REQTS. DEGRADATION FACTOR	NET REQUIRED IMPROVEMENT FACTOR
DATA ACQUISITION AND DETECTION	PDP-11/20	10 KC	30-50 KC	~2-3	~5	~8
INFORMATION EXTRACTION AND REDUCTION	360/50	36 SEC	~5 SEC	~2-3	~2-3	~7
INFORMATION ANALYSIS - ELEMENTAL COMPOSITION ON nx100 peaks	360/50	5 SEC	~5 SEC	~1-2	~1-2	~1-2
INFORMATION ANALYSIS - ARTIFICIAL INTELLIGENCE	360/67	15-300 SEC*	~5 SEC	~2-10?	?	?

$$\left\{\begin{array}{c}\text{NET REQUIRED}\\ \text{IMPROVEMENT}\\ \text{FACTOR}\end{array}\right\} = \frac{\left\{\begin{array}{c}\text{ADDED REQUIREMENTS}\\ \text{DEGRADATION}\\ \text{FACTOR}\end{array}\right\}}{\left\{\begin{array}{c}\text{SHORT TERM}\\ \text{IMPROVEMENT}\\ \text{FACTOR}\end{array}\right\}} \times \left\{\begin{array}{c}\text{GROSS REQUIRED}\\ \text{IMPROVEMENT}\\ \text{FACTOR}\end{array}\right\}$$

$$\left\{\begin{array}{c}\text{GROSS REQUIRED}\\ \text{IMPROVEMENT}\\ \text{FACTOR}\end{array}\right\} = \frac{\left\{\text{CURRENT THROUGHPUT TIME}\right\}}{\left\{\text{REQUIRED THROUGHPUT TIME}\right\}} \text{ or } \frac{\left\{\text{REQUIRED THROUGHPUT RATE}\right\}}{\left\{\text{CURRENT THROUGHPUT RATE}\right\}}$$

* These times are for complete analysis. The parameters necessary for information collection control are estimated to be available in from 5-15 seconds.

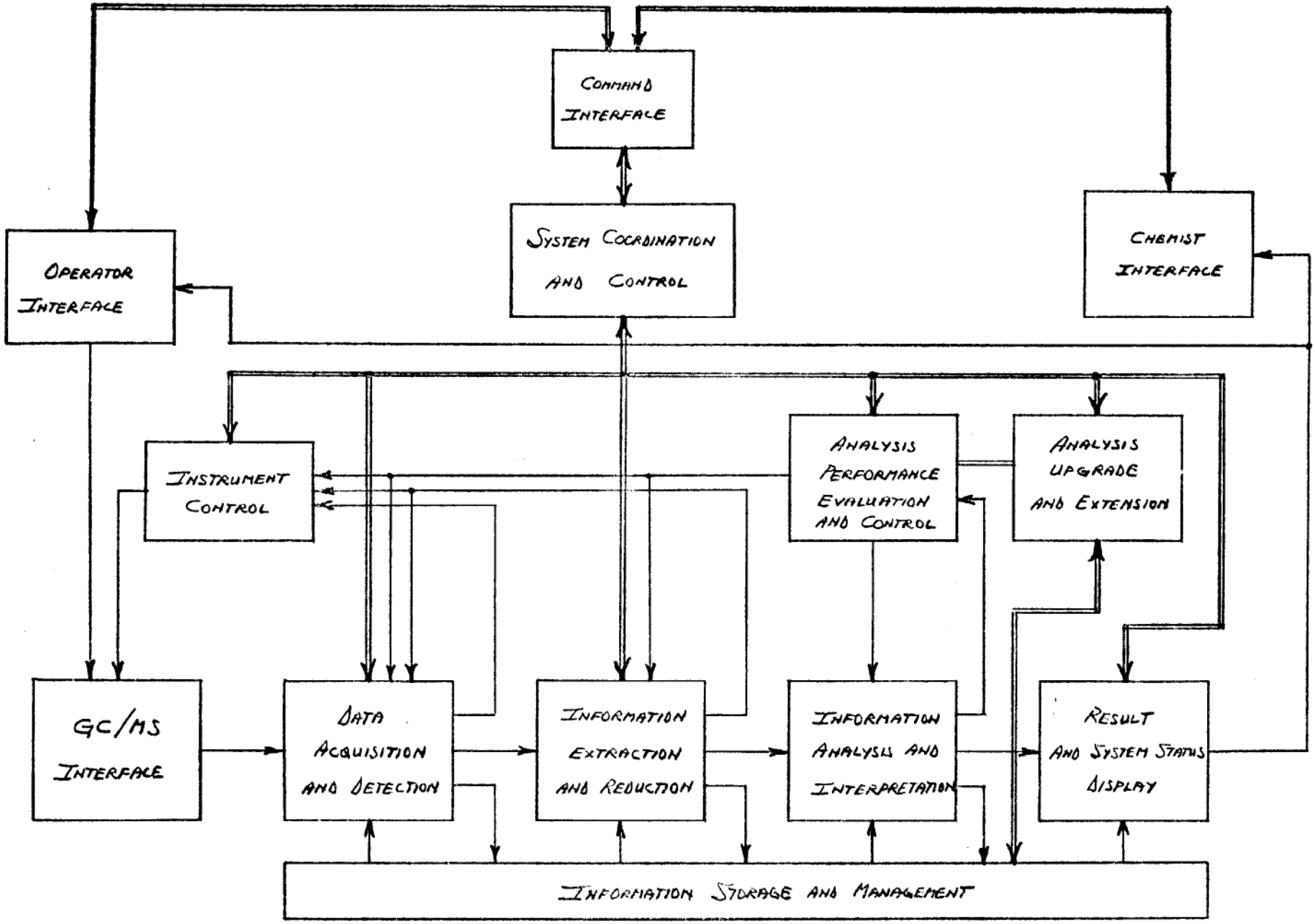
TABLE 3.

I. IBM

PRO	CON
1. Good hierarchy of large operational machines.	1. Relatively inefficient system software just now focussing on time sharing and with little real time capability.
2. Good service facilities and reliability.	2. Limited interrupt architecture for real time support.
3. Current data reduction and artificial intelligence software written for System 360.	3. Relatively little artificial intelligence research work on IBM machines.
4. Stanford appears committed to IBM enhancing local overflow support.	4. Relatively expensive equipment.
5. IBM will likely produce reliable new hardware incorporating state-of-the-art but constrained to support existing software.	5. IBM developments are geared largely to the non-scientific market.
6. DENDRAL facility could draw from the Stanford systems group without duplicating effort.	6. Very limited small computer capability in the IBM line now.

II. DEC

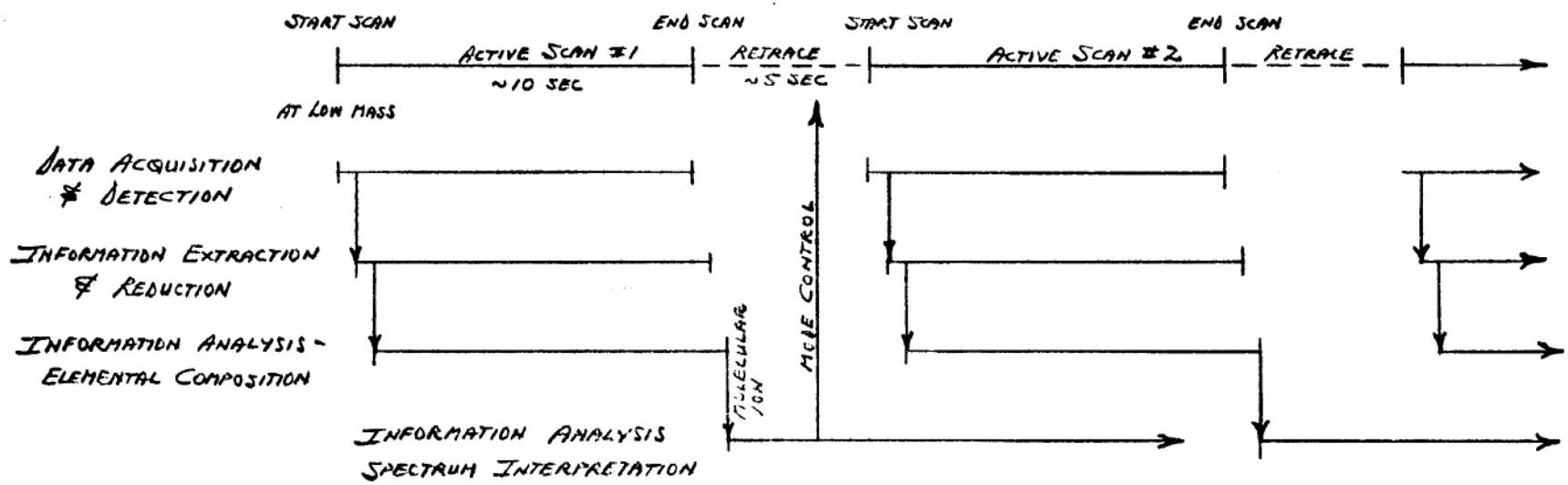
PRO	CON
1. Most artificial intelligence work is done on DEC machines.	1. Very large machines either just being delivered or under development.
2. ARPA network is built around DEC hardware at the nodes.	2. Stanford facilities are committed to IBM so limited local overflow is possible.
3. Existing reliable machines of moderate capacity.	3. Only partial common systems effort possible with other local PDP-10 users - in general requires a separate systems group.
4. Existing mini computer line of very high speed and good capabilities.	
5. Excellent system software supporting time share, batch, and real time processing as well as small machine programming.	
6. Excellent architecture for real time support in terms of interrupt structure, etc.	



REAL TIME MASS SPECTRUM INTERPRETATION SYSTEM

FIGURE 1

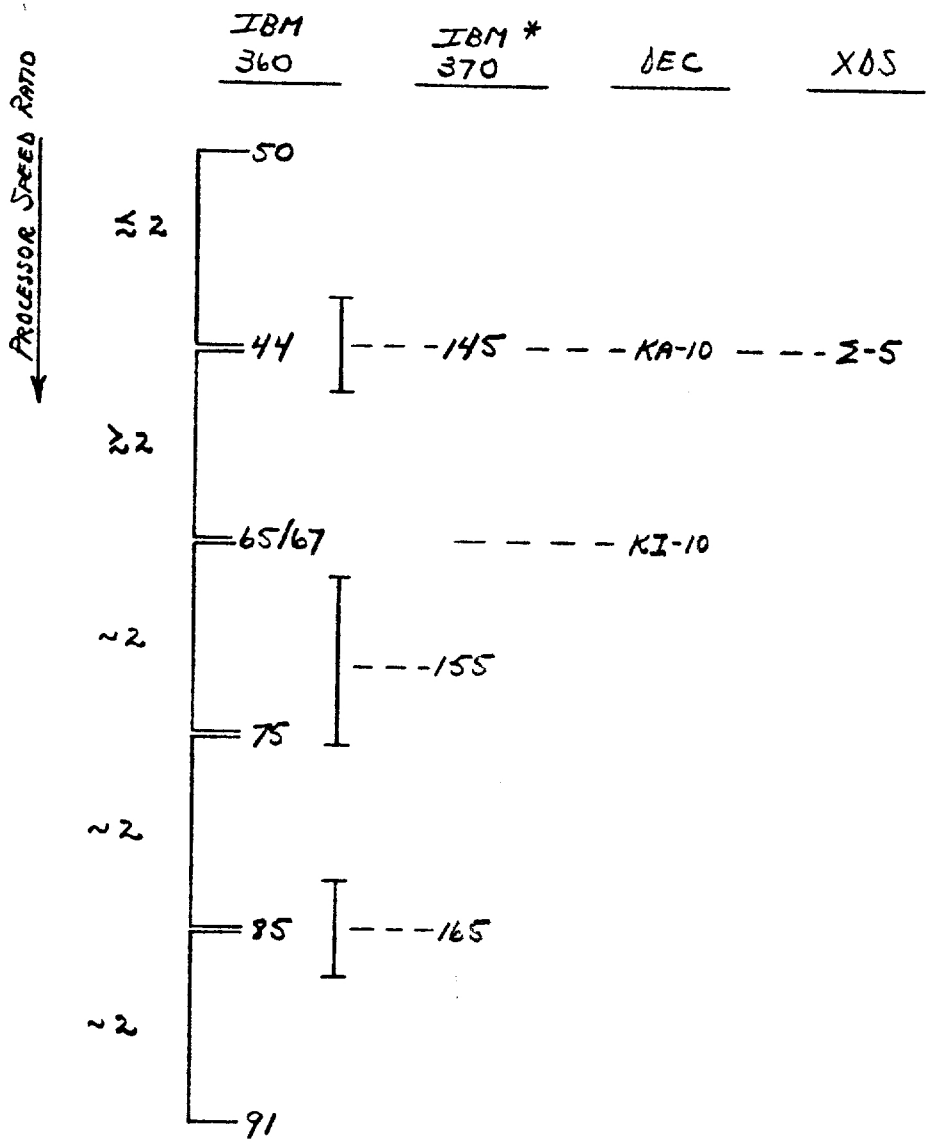
TCR 2/17/72



PROCESSING TIME INTERDEPENDENCE

FIGURE 2

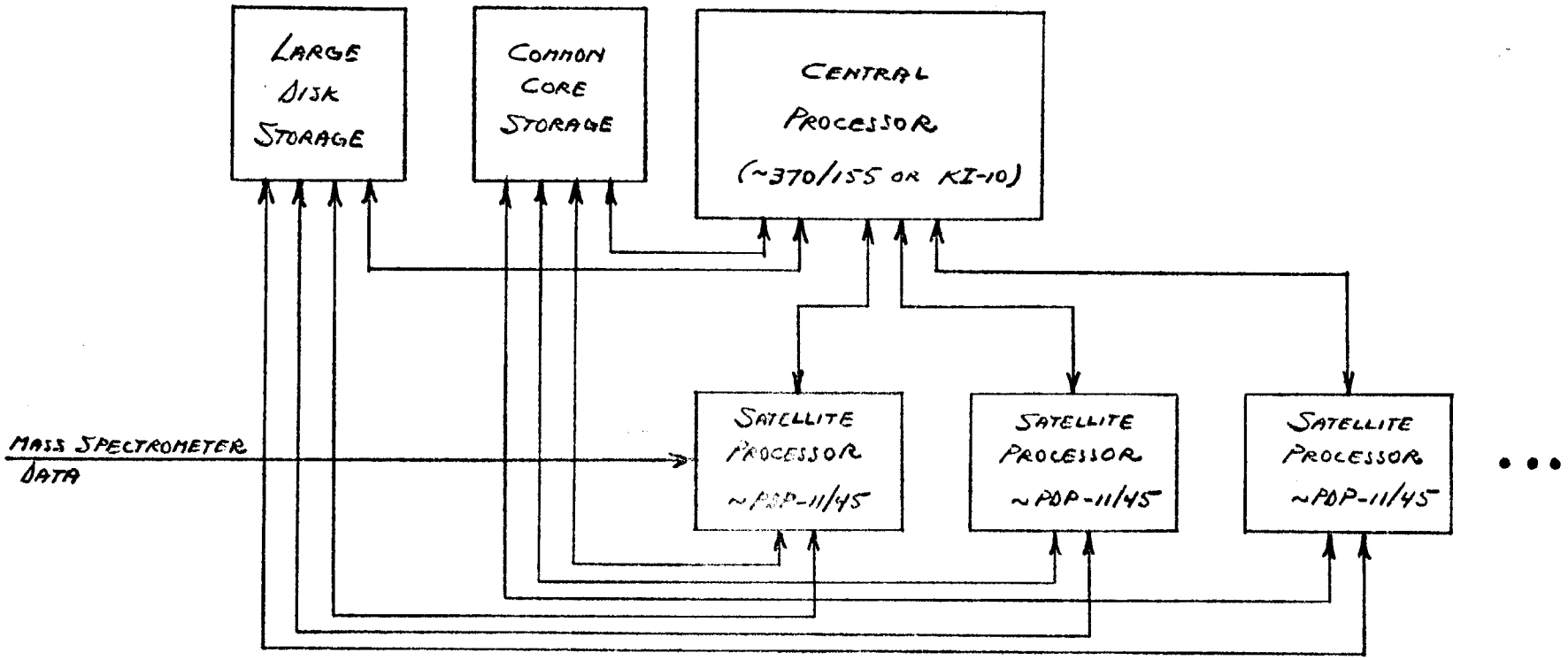
TCP 2/17/72



* The performance of System 370 machines depends upon a fast cache memory whose utility may be degraded for "random" addressing code such as is required in LISP-like programs

GROSS CPU SPEED COMPARISON

FIGURE 3



SATELLITE COMPUTER CONFIGURATION

FIGURE 4