Computer usage control

M. M. Lehman*

Department of Computing and Control, Imperial College, London SW7 2BT

CASCOM is a system of administrative procedures and accounting programs that together comprise an integrated multiple-computer, multiple operating systems, pricing and usage control mechanism. The original version described in this paper, was implemented in APL/360 with full interactive capabilities. It has been followed by a non-interactive, PL/I, version. The latter has been in full, closed-loop, operation at the IBM Research Division Laboratory, Yorktown Heights, since February, 1972.

The paper discusses the philosophy underlying CASCOM, outlines its functional specification and briefly discusses the implementations. Operational experience is however not included.

(Received February 1972)

This paper was originally written as a justification and architectural specification for the institution and control of a computer usage accounting system at the IBM Research Laboratories, Yorktown Heights, New York. An initial APL\360 interactive version (The Design System) went into operation in November 1970, some three months after discussions on a control and charging scheme were first initiated. A refined APL\360 version went into operation in February 1971 and a PL/1, non-interactive, system has been in closed-loop operation since February 1972. An interactive facility is now being added.

The present paper describes the concepts, architecture and design as they evolved during development and early operation of the two APL systems. Operational experience cannot be reported here as the author is no longer with IBM.

1. Introduction

CASCOM is a system of administrative procedures and accounting programs that comprise a multiple computer, multiple operating systems, pricing and usage-control mechanism. Used in conjunction with flexible allocation and pricing policies it was conceived to help achieve increasingly effective usage of the computer installation at the IBM Research Division Laboratory at Yorktown Heights.

Overall control is dependent on the creation of an *integrated* usage and accounting record for *all* users of *any* of the installed computer systems. Time-dependent usage rates determine the total computation expenditure of each group and/or each individual user in CU's (Compute Units). Periodic budget allocations limit usage and are used to control distribution of the Yorktown computer resources as detailed below. Various functions are provided for the maintenance, control and interrogation of the data operation. APL\360 was selected as the most effective language for the initial design and implementation of a complex, interactive, system that was to provide wide on-line access to the database (accounts) that it generates and controls. It was recognised from the start that after stabilisation of CASCOM, a final version would probably have to be implemented in some other language, to overcome the performance, workspace capacity, I/O and computing resource limitations of APL\360. As discussed in the previous section this has now been done with the implementation of the PL/1 version. Because of the limited implementation resources available this latter system was however not initially given interactive interrogation capabilities.

2. Objectives

The prime objective of CASCOM is increased installation *utilisation* and *effectiveness*. A secondary objective is the *fair* distribution of the available resources amongst the user population based on the usage *priorities* set by management. This latter sub-objective requires the measurement or, at least, estimation of the value of different computing activities. Such value judgements are however a management responsibility and should not, indeed cannot, be made by computer installation management.

Efficient operation and utilisation of a computer installation is always important but especially so at a time of economic stress (Neilsen, 1970). The overall demand for computer services at Yorktown is continuing to grow at a rapid pace as a consequence of both the expansion of existing usage and the development of new applications. The supply of raw computer power on the other hand, after an extended period of rapid growth, is likely to remain essentially constant for some time.

To meet the general objective, it is essential to spread the total compute load evenly and equitably over the entire day and week. That is, the pricing and control scheme should operate as a load-levelling mechanism. It should reduce demand, and hence improve response and turnaround times during computer prime time, by shifting a portion of the load to the night hours and weekend periods. At these latter times the installation is, at present, often only lightly loaded.

A significant increase in available resources may also be obtained by inculcating good usage practice in users. Good habits can help to minimise and balance *sub-system* usage. For example terminals should be connected to a time sharing system *only* when the system is being used and not as a means of ensuring future access. Equally, core-storage requests during execution under multi-programmed batch processing should be minimised, and data should be retained (even on archival disc storage) only if it is likely to be used again at a later date. Such storage in turn should be organised and compacted to reduce the amount of disc space required.

Equally in an operating environment including a number of machines and several operating systems, users should be encouraged to make discriminating use of the alternative systems offered. That is, they should receive encouragement to migrate between systems when it is in the interest of the installation or in their own interest to do so. Thus for example, in the Yorktown installation it may be justifiable to develop a new program interactively on TSS/360 or CP/CMS but to compile and/or execute in batch mode under OS-MVT which runs on a larger, more efficient machine.

Similarly it may be desirable to *rewrite* a program first developed in APL\360 so that regular and repeated production

*Formerly at IBM Thomas J. Watson Research Centre, Yorktown Heights, New York.

runs can be executed from compiled mode rather than in interpretive mode.

In general, charging schemes should be designed to maximise the total *value obtained* from a computer installation. Simultaneously they should *improve*, not degrade, the computing services offered, benefiting the individual user, not hindering him. At best this can be achieved only for most of the users, most of the time.

3. Implications

Certain implications derive directly from the above objectives.

No idling in presence of demand

In designing the accounting and control system and selecting a *control strategy*, the aim could be to ensure maximum value to the body of users. The mere statement of this objective, however, does not help to determine just what constitutes the best value in computing service. Alternatively, the installation objective might be to achieve maximum return from the installation or high system utilisation on worthwhile activity. Determination of an appropriate, quantifiable, measurable and attainable objective is, in itself, a non-trivial task, but one element of any objective can be stated with certainty. In any event the control mechanism must not lead to resource *idling* as a consequence of a lack of 'funds' amongst just those users who wish to use the machines at a given time. *Resources that idle are lost forever.*

Sub-system charges

Usage measures, and rates on which charges are based, must be applied to all appropriate *sub-systems* of all available systems so that the usage of scarce resources (in particular) can be measured, charged for and hence controlled. **Table 1** lists the resources initially included in the CASCOM scheme.

Integrated charges

The controlling system must *integrate* charges over the entire installation so as to be able to encourage desirable inter-system migration, to prevent saturation of any one system or resource by adjusting relative rates.

Adjustable rates

Usage rates can *initially* be cost-related, but must be adjustable up or down according to the scarcity of a resource and the availability of cheaper alternatives. Off-line disc storage, for example, should be cheaper than on-line storage because it does not occupy a disc drive. Similarly evening or weekend charges for an interactive service should be cheaper than mid-morning weekdays. In general the rate should reflect the cost of a sub-system, demand for its services, availability of resources to meet that demand, availability of alternatives and the impact of incremental demand on system balance and performance.

Differential shift rates

Differential shift rates must be instituted to encourage the flow of usage from presently prime to presently slack usage periods. For conversational usage the shift rate applied should be determined by the actual time of use. For batch submissions the user should be able to choose between alternate classes of service for a given job, thereby specifying the maximum rate he is willing to pay and hence an acceptable turnaround time.

In the Yorktown implementation, batch users are able to specify a first and second choice of shift or User Period (UP) for execution of their work. This facility is, of course, not only reflected in CASCOM but has also led to a need for (minor) modifications to the various operating systems. The rate actually charged is the lower of the rates for the UP requested and for the UP during which execution actually occurred. Where the second choice is honoured, jobs are put at the head of the queue for that shift and surcharges are incurred for the priority gained. Discounts are given where neither of the UP requests could be met.

Global allocation

In a well balanced installation and configuration, user groups and users should receive a single allocation to cover any desired combination of system and resource (sub-system) usage. That is, the controlled usage of scarce resources and encouragement for the use of surplus resources should be obtained by relative rate adjustments. *Ultimately only the user can know his true needs.* Management that allocates potential usage can only make a *global* judgement of different usages' relative value.

Hierarchically distributed allocation procedure

Allocations for resource usage should be made by those able to judge the relative value (to the organisation and according to its priorities) of the various competing activities. Intrinsically, decisions as to allocations must be spread hierarchically with only the top, departmental, allocation being decided by top management. Thereafter allocations should be made at each level by those most intimately aware of the immediate relative needs, priorities, claims and values.

Levels of sub-allocation

Budget allocations in computer units (CU's) should be made

Table 1 Chargeable Resources and Units

	RESOURCE NUMBER					
SYSTEM	1 6 DIGITS	2 7 DIGITS	3 5 DIGITS	4 5 DIGITS	5 7 DIGITS	6 7 DIGITS
APL/360 (Code A)	CPU Time Seconds	Connect Time Minutes	Storage Workspace-days			
CP/67 (Code C)	CPU Time Seconds	Connect Time Minutes	Spindle-Residence Cylinder-hours		Supervisor State CPU-seconds	Virtual Storage Kilobyte-min.
OS-MVT (91) (Code O)	CPU Time Seconds	RJE or TSS 'SHIP' Usage	External Storage Cylinder-days	Printed Output Pages	EXCP's	Core-Time Product Kilobyte-min.
TSS/360 (Code T)	CPU Time Seconds	Connect Time Minutes	On-line Storage Page-days	Off-line Storage Page-days	TSLC-Time Product Seconds	Virtual Storage Page-minutes

down to the project and individual user level. Thus they can directly encourage responsible individual usage.

For CASCOM, at Yorktown, the Director of Research or his representative allocates CU's to departments. Department directors are *encouraged*, but not obligated, to sub-allocate to projects, project managers to groups or individuals and so on. Certain individuals even need to have more than one account. The CASCOM system records and administers the allocations, and enforces control down to whatever level sub-allocations are made.

Initial allocation

In allocating the totality of available resources, initial distribution should relate to the actual recorded usage in some preceding time period rather than to some theoretical ceiling of resources available. In the absence of other authorised demand, the CASCOM scheme permits users who have exhausted their allocation to continue usage. Also allocations are easily changed. Hence the primary allocation strategy should under-allocate so as to hold units in reserve for the additional demand and changing usage patterns that will inevitably develop.

Changeability of allocation

Budget allocations must be easily changeable. This will ensure minimum inconvenience to users, smooth convergence to an optimum distribution of the totality of resources, and an ability to adjust allocations to changing demand in a dynamic application and usage environment.

Exhaustion of allocation

Any user (or project or department) having used one hundred per cent of his allocation in a given control period should, after appropriate warning, be relegated to lowest priority access for the remainder of the control period, or until more CU's can be allocated.

Creation of a 'Low Priority' classification is preferable to totally denying such users access to the systems for the remainder of the control period. It is less disruptive of ongoing projects, and permits utilisation of resources which might otherwise idle and be wasted.

In CASCOM users who have used up their allocation are given two working days notice before being relegated to low priority status. If, after receipt of such notice, a user reaches one hundred and ten per cent of his allocation, he is placed in the 'Lowest Priority' class without further warning.

Users in the lowest priority class submitting batch jobs, have them executed only if no other jobs with a higher status are ready for execution. On the time-sharing systems, users in the low priority class are permitted to sign-on when the load is below some critical level. If the system reaches the critical level while they are signed-on, they may be requested to sign off or they may be disconnected from the system without additional warning.

All usage in the low priority category is charged to the appropriate account, though at a rate reflecting the low priority treatment. Batch users may also *submit* jobs in the first place, with a request for low-priority service, so as to benefit from this low rate.

Remanent allocations

To avoid the dangers of over allocation and inflation, and since idle computer resources are lost and cannot be recovered, remanent allocations at the end of any control period should not be retained or added to the next allocation.

Annual allocations

A yearly departmental CU allocation underlies the monthly allocation. This should be made together with all other divisional resource allocations to ensure overall compatibility.

Table 2 The main IBM Research Division (Yorktown Heights) computing systems

HARDWARE	OPERATING SYSTEM	SUBSIDIARY SYSTEM
360/91	LASP-OS/360-MVT	APL 360
360/67	TSS/360	—
360/67	CP/67	CP/CMS CP-Virtual Machines

In CASCOM the monthly sub-allocation in month m reflects one part in $(12 - m + 1)$ of the remaining allocation *unless* and *until* responsible managers initiate allocation changes.

Account accessibility

There must be simple and up-to-date accessibility to the current state of each account so that users can be forewarned of an approach to the budget ceiling. Equally, in the event of overspending, any individual authorised to reallocate CU's at the departmental or group levels must be able to determine under-usage at that level so as to be able to draw on under-utilised allocations.

Visibility of usage detail

Users should also be able to ascertain the makeup of the CU usage in terms of both systems and sub-systems. Such detailed visibility of usage and accounting data is necessary to detect inefficient usage patterns, to control the use of individual resources and to encourage use of the most appropriate system and time for each application and circumstance.

CASCOM was originally conceived as an interactive system and provided with the control and interrogation functions summarised in Section 4.3 and more completely specified elsewhere (Lehman, 1971). These provide managers and users with the requisite degree of access to up-to-date data. In this APL implementation, accounts were updated once in 24 hours, since there existed no on-line communication between the APL and the other Research systems. The PL/I operational version has, however, been implemented without the interactive interrogation capability and account visibility is obtained by the daily and weekly circulation of appropriately detailed printout to appropriate personnel together with usage statements on each batch job printout and upon sign-off after each interactive session.

Control period

The control period should be relatively *short* (CASCOM—one month). In this way allocations can be more dynamically maintained in step with changing needs. In addition, temporary inconvenience of access will not be crippling to a project or an individual.

Staggered starts

Staggering the start of the control period for different groups produces a smoothing of demand and hence peak load reduction. That is it causes the user population to be comprised, at all times, of individuals at the beginning, in the middle and at the end of their control period. This is a condition that is most likely to prevent large variations in demand.

In CASCOM the control period for department XYZ ends on day $(3Y + 1)$ of the calendar month.

Interchangeability of computer units

The desirability of making computer units interchangeable with other budget funds represents a problem to which no clear answer exists. It will be very much a function of local conditions and policy questions. At Yorktown, CU's cannot be interchanged.

Enforcement
The rules relating to users exceeding their allocation must be *rigorously enforced* if the control system is to achieve its stated objectives. Since users will be given adequate warning of impending excess usage and since allocation changes will be simple, *there should be no reason to make exceptions at any time.*

Changeability of system
The behaviour of economic systems controlled by pricing mechanisms and self regulation are notoriously difficult to plan or predict. Thus any system implementation and the database it generates should be capable of extension, modification and growth.

The above implications of the original objectives underly the development of the basic CASCOM philosophy. They guided the implementation of the interactive system, CASCOM/APL. This served as an interim control system, and the practical experience gained with it formed the design base for the current PL/I system. There are only minor differences in detail between the concepts outlined here and the implementation now running, except that as stated above, the present implementation has, for practical reasons, been implemented without the interactive facility. But the PL/I system does not share the main disadvantage of APL: a severe limit of available high speed storage, no on-line communication with the other Research systems and interpretive execution which leads to excessive computer resource usage by the control system itself.

CASCOM generates a detailed record of usage patterns, which has materially assisted in the detection of bad or inefficient usage. Its mere existence has also generated an awareness of the value and scarcity of computer resources and of the communal and individual advantage of thoughtful and knowledgeable usage of computer resources. Equally CASCOM has forced management to think about methods of evaluating the *value* of various computing activities relative to organisational and its own objectives. Finally, the system and its accompanying protocol has proved itself as a tool for learning how to use a charging scheme as a self-regulating control mechanism for computing resources.

4. CASCOM

4.1. *General description*

CASCOM monitors and controls the computer usage of all users of the four major systems at Yorktown Heights, outlined in **Table 2.** It is also intended to include various smaller dedicated systems, 1130's, 1800's and a 360/44 in the control scheme.

Each department is given a total yearly allocation of Computer Units (CU's) subdivided into monthly allowances. The latter provide the basic control at departmental level. Where, as expected, a director sub-allocates his CU budget to projects and/or individual users, such sub-allocations become the primary controlling element. Control periods (months) are staggered by department. A user may maintain separate records for different applications within a single project by signing on with minor-variation (say initials) of his name ID. They will however not be encouraged to do so, to avoid proliferation of user ID's.

Each of the operating systems has its own individual accounting system but these are not compatible with one another. The outputs of the individual accounting systems are integrated by means of a nightly transfer to CASCOM. Each data transfer corresponding to a batch job or an interactive user-session includes a date, serial number, system and shift ID, department, project and user ID and record of resource utilisation. Details of the resources being accounted, the units in which they are measured and the transfer format are given elsewhere (Lehman, 1971).

In addition to providing integrated accounting for all Research Division computing activity, CASCOM also serves as the vehicle for control of and communication with the user population. Thus its functions provide for the registration or deregistration of users to one or more of the systems, for partial or complete changes in ID (Department, Project, Name) or allocation. Such changes are input to CASCOM and passed on to the relevant system(s) together with listings of users who have exceeded their allocation and are therefore to be restricted to low priority usage. A year-to-year record is also maintained for the division, by department.

CASCOM also generates accounting statements for non-Divisional, dollar-accountable users. Billable charges for such users are computed from a separate set of dollar-based cost-recovery rates. These may differ from the CU rates which are fixed to encourage usage habits that improve system effectiveness.

A common question for all charging schemes is whether to make CU's interchangeable with real money. In the current system this was not done for legal reasons. An important economic agreement *against* convertibility in organisations where the purpose of a scheme is efficient utilisation and fair distribution of limited resources, rather than cost recovery (with profits), is the fact that charging rates reflect the scarcity of a resource and the policies of the installation management rather than the actual fiscal value of resources used.

The interactive system included a number of user accessible APL workspaces each of which held usage, charge and/or summary records, for a department or group of departments. Each of the workspaces contained APL functions, called by appropriate mnemonics, that permitted interrogation of the records.

The update procedure includes a listing of the ID's of users having reached some pre-specified level of usage of their allocation. It also provides calendar month statistics of usage by system, resource and department. Other functions permit the interrogation of the database to extract detailed or summary records of usage, charge and budget-status data on individuals, projects or departments. Finally service functions permit the accounting system administrator to make adjustments to allocations, accounts, rates and system resources being charged.

4.2. *Security*

Ease of access and visibility is considered absolutely essential to a successful control scheme. At the same time people should be permitted read-access only to such data that concerns them. Similarly only authorised personnel should be permitted write-access to the database. Any system must, therefore, include a security scheme. That developed for CASCOM was a five level scheme as follows:

Level	*Access*	*Individual*
0	*Read* and *Write* to entire database	CASCOM Administrator
1	*Read* to entire database	Director of Research
2	*Read* to departmental data and below *Write* to project and individual *allocations* subject to project and department allocation ceiling	Department Directors
3	*Read* to project data and below *Write* to individual allocations subject to project allocation ceiling	Project Managers
4	*Read* to individual (own) data	Each user

Each individual should be able to authorise others to his own level of authorisation by passing on his security key. Interrogation of the system should be possible by signing on to the system under the user's identifying code or under a general CASCOM enquiry number and identifying oneself with the correct security key to obtain access to the database at the desired level.

4.3. *Functions*

4.3.1. *Summary*

Interactive CASCOM included three groups of functions. Any function in any group could be used only after the user's authority to do so had been recognised.

Systems—Administrative functions available only to the system administrator in the master workspace.
User—Control functions available to Department Directors, other responsible managers and those authorised by them.
Interrogation functions available to any user of any system.
We list here a brief description of some of these functions. A more complete specification is given elsewhere (Lehman, 1971).

4.3.2. *System—Administration functions*

These are executed in locked master workspaces and include:

UPDATE	Daily usage-record update. The update includes an ID printout of all users who have reached some percentage p (say 90) of their allocation and who must be warned of their imminent approach to the usage limit, a listing of those who have exceeded their allocation and who are liable for fourth shift status, and a listing of those who have received adequate notice and no re-allocation and are being transferred to fourth shift status. A list should also be printed of those whose fourth shift status has been revoked as the result of the re-allocation of CU's. Similar action is also taken for all users of any project or department that has exhausted its allocation. UPDATE also identifies and takes the appropriate actions for any department that has reached the end of its control period(s). Finally UPDATE accumulates and prints out usage statistics for each system and resource by department for the current calendar month. The statistics include a record, by department of multiple system usage
SEPARATE	Permits the splitting of the structured database into two or more separate structured entities
SELECT	Deletes all ID's, allocations and usage data *except* those of the indicated departments, projects or users
DELETE	All ID's, allocation and usage data of one or more departments, projects or users
CHANGE RATE	Change chargeable rate for some system resource(s), for one or more UP's
CHARGE DEPT. ALLOC.	Change number of CU's allocated to some specified department
ADJ. CHARGE	Adjusts the usage record/charges incurred by some user
CHANGE FIXED CHARGE	Provides for the inclusion and any changes in charges for whole systems allocated to specific groups
ADD. SYS ADD. RES EXC. SYS EXC. RES	These utility functions add new systems or additional resources or exclude existing systems or resources from CASCOM
AUTHORISE	Gives the identified user the authority to access the named class of functions and to interrogate the system at the appropriate level
YTD	A collection of functions that permit the maintenance of a year-to-date record.

4.3.3. *User control functions*

USAGE L	Lists the allocation, actual CU usage and percentage of allocation used for all users exceeding a level of usage L CU's
USAGE L M	As USAGE L except that it lists data for users whose usage lies in the interval L to M CU's
PUSAGE L PUSAGE L M	As the two previous functions with selection based on percentage of allocation usage rather than absolute value
DEPTORD PROJORD	Lists users and data for designated groups in ascending or descending order of actual usage
DEPTORDA PROJORDA	As previous but listing ordered by allocation
STATUS CHANGE	Permits a director or his representative to register new users, deregister existing users from one or more systems, or record any changes
CHG. ALLOC.	Permits an authorised person to change allocations to projects within his department or to individual users within a project
LOCK	Permits the creation or change of a user-selected lock for those authorised to access the systems at other than the individual user level.

4.3.4. *Interrogation functions*

DEPTSUM PROJSUM USERSUM	Presents a summary of usage data for one or more departments, projects or users, respectively
DEPT. ALLOC. PROJ. ALLOC. USER. ALLOC.	Functions that provide details of the current allocation of the indicated groups and its constituents
RATES	Gives usage rates by system, resource and shift (UP).

Acknowledgements

Thanks are due to all those who participated in the evolution of the concepts and system characteristics described in this paper, in particular to R. Evans, R. Kelisky and D. Streeter, all of the IBM Research Division, Yorktown Heights, New York who participated in day-to-day discussion on various aspects of the systems and the protocol to be followed. Thanks are also due to IBM for permission to publish this paper.

References

NEILSEN, N. R. (1970). The Allocation of Computer Resources—is Pricing the Answer?, *CACM*, Vol. 13, No. 8, pp. 467-476 (see this reference for further bibliography).

LEHMAN, M. M. (1971). *CASCOM—Computer Usage Control System*, IBM Research Report RC. 3447, Yorktown Heights, New York 10598, USA, July 30th, 1971.