RESOURCE ALLOCATION ON SUPERCOMPUTERS

SHARES VERSUS CHARGING

ROBERT C. BELL

CSIRO SUPERCOMPUTING SUPPORT MANAGER
55 BARRY STREET
CARLTON VIC. 3053

ABSTRACT

There are a number of ways to apportion resources on shared facilities such as supercomputers. Some of them, such as charging, lead to well-known problems such as gross under-utilisation.

A review of some of the schemes tried by CSIRO will be given. The author will then discuss the use of the Fair Share Scheduler on the Joint Supercomputer Facility's Cray. The scheduler provides a mechanism for partitioning the machine between the parties and is used by CSIRO to share resources between Divisions according to the contribution to the development fund.

A simple model of the scheduler is used to explore its fairness.

Use of the fair share scheduler is preferable to most other resource allocation schemes.

INTRODUCTION

Obtaining funding for the purchase of supercomputers or any other central facility is a difficult task. Once the facility is in operation, problems of continuing funding and resource allocation to users seem to get intertwined in the user pays principle.

In this paper some schemes for cost recovery and resource allocation used by CSIRO in recent years will be reviewed. The current schemes in use for the Joint Supercomputer Facility (JSF) Cray Y-MP2/216 will be described.

CSIRO CENTRAL COMPUTING

When I first became aware of charging for usage of the CSIRO Division of Computing Research (DCR) computers some kind of "funny money" scheme was in operation. Divisions were given an allocation of money which could be spent only on DCR computing for which there was charging. Similar schemes were, and probably still are, in use in Universities. The arrangement requires someone to allocate money in advance to Divisions and requires end of year money shuffling between Divisions for the central facility to receive its full allocation.

At some stage this funny money scheme became a real-money scheme where the Divisions had to pay for their central computing with real money. This led to a number of problems:

1. Given the choice, research managers followed their usual empire-building mentality and chose to acquire their own machines, immune from CSIRO policy changes. This led to a fragmentation and isolationism in CSIRO computing which continues to this day.

2. Charging with real money inhibited scientists from using the central computers. A colleague suggested to me that charging for in-house computing usage is akin to charging a scientist $100 every time he sets foot inside his laboratory.

3. Both of the above problems caused the central facility to be short of funds. So each financial year the charging rates were increased in order to try to recover more money. This produced a vicious circle. As fewer and fewer people could afford to use the facility the rates had to be increased again. Of course, in the end, the lie about real money and user pays was exposed when the annual reports revealed that the facility's deficit was made up from central corporate funds.

4. All of the above led to an enormous waste of resources with expensive machines lying idle because users could not afford to use them. This was very frustrating to CSIRO scientists especially as the incremental cost of using a machine compared with leaving it idle is negligible.

5. Another problem with the charging schemes was that in contrast to the usual controls over expenditure in CSIRO, any computing user could and often did run up bills for thousands of dollars without proper authorisation. This could wreck a research program budget and cause considerable damage to a Divisional budget.

CYBER 205

The Cyber 205 service commenced with a similar real money charging scheme but with the addition of a research grants scheme (open to CSIRO and universities) similar to that used for many supercomputers. This scheme added an extra unwanted research review for CSIRO scientists. It was hoped that the deficit would be funded by paying commercial customers but there was little usage by them or the universities. An annual levy to meet the deficit was then proposed, to be based on past usage. The imposition of levies of several hundreds of thousands of dollars on large user Divisions was met with strong opposition. A discounted charging scheme was also proposed, but in the end, usage of the Cyber 205 was made free to CSIRO divisions, apart from charging for disc usage above a quota. This final regime resulted in a well utilised machine, something which had not been seen in CSIRO for some time. Unfriendly, costly front-end facilities together with user and management dissatisfaction were enough to prevent saturation occurring.

JSF CRAY Y-MP

The Joint Supercomputer Facility Cray Y-MP is shared between three parties, CSIRO, Leading Edge Technologies Pty. Ltd. and Cray Research (Aust) Ltd. The Fair Share Scheduler (FSS) of Kay and Lauder [1] is used to ensure that the parties receive their correct shares as specified in the contract.

The FSS decides which task to execute by looking at the user's recent usage and the share assigned to the user. At each scheduling tick it assigns the selected tasks to the CPU so that the desired percentages are achieved. When only a single user or group of users are using the machine, they have access to the entire machine so there is no waste. Afterwards, when other users logon or submit batch jobs, they are initially given a greater share than their allocation, to catch up on the first users. There is a decay factor applied to the usage, so that a user who does a burst of computation is not penalised for long. During the day, the released value of 1 minute for the decay half-life was retained, but outside prime time, the value was increased to fifteen minutes to provide better conformance to the desired shares when long batch job predominate.

Initially the released algorithm of the FSS was retained, which meant that only CPU time was taken into account when the FSS calculated a user's usage. Later on memory usage was also taken into account but had to be withdrawn when it was found that suspended jobs were still considered to be using a share, in some cases over 20%, because of memory size. The FSS uses the task memory size rather than the actual use of memory which seems to be a deficiency.

The Fair Share Scheduler has a hierarchy of up to four levels. Within each of the major parties, groups and users can be assigned different shares.

CSIRO FUNDING

When the arrangements for the JSF were completed CSIRO was faced with a significant cost above the corporate budget for supercomputing. It appeared inevitable that charging would be instituted with all the attendant problems.

However the CSIRO Institutes were persuaded to fund the deficit, and in exchange, Divisions were asked to finance a Development Fund to provide for hardware and software enhancements. Divisions are far happier to contribute to a fund for enhancements rather than to meet costs.

Rather than charging Divisions for the use of the JSF Cray I proposed that Divisions be asked to contribute to the Development Fund and be rewarded for their contribution by being given a share of the machine proportional to their contribution. A target for the Fund was set, with Divisions contributing monthly according to their likely usage in the coming month.

The concept is new to CSIRO but has been well received by users and Divisions when it has been explained that contributors are receiving not a fixed amount of time but are purchasing a share of the capacity when there is contention for resources on the machine. Users can get more than their share without significant penalty when the machine is lightly loaded. The scheme has been in operation since mid-August.

One of the problems not yet resolved fully is how to make sure the Divisions contribute at a reasonable level so that the target amount for the Development Fund is reached. In theory, Divisions could collude in their bidding for shares and ruin the scheme. Fortunately for this year a major using Division has put in a standing order with a large contribution. This puts a floor price into the system and means that Divisions which want to use the Cray have to contribute significant amounts of money.

Usage of the Cray remains very cheap to Divisions. If the Development Fund target is reached and CSIRO uses precisely its share, the cost of Cray usage is about $45 per hour.

The share scheme is quite simple to operate and avoids most of the problems of retrospective charging.

FAIRNESS OF THE FAIR SHARE SCHEDULER

There are some areas where the FSS may not be effective in delivering the required shares.

Firstly, management has to decide how usage of the various resources such as CPU, memory, i/o and system calls are to be weighed in an overall measure of usage. At present only CPU time is considered (partly for reasons given earlier) and this is not a good measure with a mix of jobs which are i/o and memory bound as well as CPU bound.

Secondly, there is no interaction between the FSS and the initiation of batch jobs. This means, for example, that a long job belonging to a user with a very low share or a lot of recent usage could be initiated ahead of a far more worthy customer. This job would take a long time to complete, getting only a very small proportion of CPU time and tying up other resources such as memory, swap space and job slots. An enhancement to the FSS, or manual scheduling for long time or large memory jobs would be desirable.

Thirdly, because of the hierarchical nature of the FSS, users who belong to a group with few users are at a disadvantage compared with users in a big group. This is because a single users in a group with many users can get the whole of the group's share.

Consider a simple example of three users, U1, U2 and U3, each with a share of 1/3. Suppose U1 and U2 belong to one group, with a share of 2/3, and U3 belongs to another group, obviously with a share of 1/3. Although the nominal share for each user is 1/3, and that will be the share received when all are requiring resources, whenever one of U1 and U2 is absent and the other is present, the one present will receive a share of 2/3.

Let p be the probability that any of the three users requires the CPU at any time. The following table shows the shares received for all the possible cases.

User Demand			Probability	Share Received		
U1	U2	U3		U1	U2	U3
n	n	n	$(1-p)^3$	0	0	0
n	n	y	$p(1-p)^2$	0	0	1
n	y	n	$p(1-p)^2$	0	1	0
n	y	y	$p^2(1-p)$	0	2/3	1/3
y	n	n	$p(1-p)^2$	1	0	0
y	n	y	$p^2(1-p)$	2/3	0	1/3
y	y	n	$p^2(1-p)$	1/2	1/2	0
y	y	y	p^3	1/3	1/3	1/3

Then the expected share for each user can be calculated by multiplying the probability by the share received for each case and summing. Let S1 and S3 be the expected share for U1 and U3 (U2 receives the same as U1, and will not be considered further). Then:

$$S1 = p - 5/6\ p^2 + 1/6\ p^3$$
$$S3 = p - 4/3\ p^2 + 2/3\ p^3$$

When $p = 0$, $S1 = S3 = 0$ and when $p = 1$, $S1 = S3 = 1/3$, as expected. However, in between, $S1 > S3$. Figure 1 shows S1 and S3, and the ratio $R = S1/S3$ as a function of p. The maximum value of R is 1.27 when $p = 0.63$.

From this simple example it can be seen that belonging to a group with multiple users is a distinct advantage and becomes more of an advantage as the number in the group increases, provided the share per users is the same. However, further analysis with unequal shares shows that when the share of U3 is small, that user's share can be considerably more than intended because of the possibility of getting a complete use of the machine. Some modification of the share scheme for CSIRO users should be attempted to remove the bias favouring the large users Divisions.

SUMMARY

The Fair Share Scheme as implemented for CSIRO Divisions has the following advantages:

- users are not inhibited from using the machine
- waste or idle time is rare
- no extra grant proposals and review committees are needed
- there is no unauthorised expenditure
- the scheme is simple to operate.

The Scheme has the following disadvantages:

- the new concept is hard to communicate
- there is the danger of collusion
- it is difficult to ensure that the target is reached
- it requires Divisions to contribute real money
- there are some biases in the hierarchical scheme.

I believe that the advantages far outweigh the disadvantages and that the scheme is an advance on previous funding and resource allocation arrangements.

REFERENCE

1. Kay J and Lander P, A fair Share Scheduler, *Communications of ACM*, 31, 1988, pp 44-45.

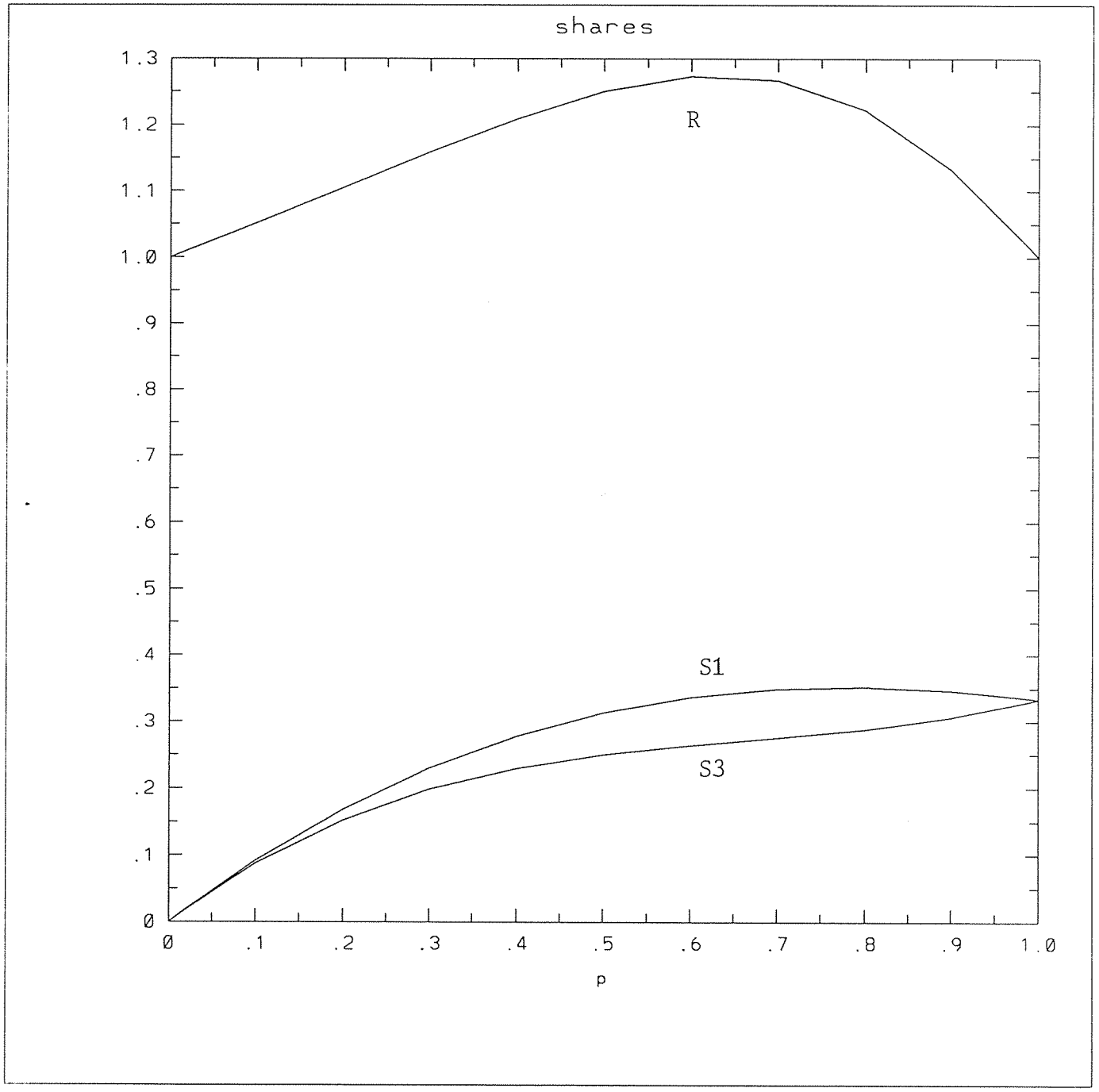

Fig. 1

Shares for Users 1 and 2 as a Function of p, the probability of requiring CPU Time.