FORECASTING DEMAND FOR CLOUD COMPUTING RESOURCES

An Agent-based Simulation of a Two Tiered Approach

Owen Rogers and Dave Cliff
Department of Computer Science, University of Bristol, Merchant Venturers Building, Bristol, U.K.

Keywords: Utility computing, Market-orientated computing, Resource reservations, Cloud computing, Options markets.

Abstract: As cloud computing grows in popularity and usage, providers of cloud services are facing challenges of scale and complexity; how can they ensure they are most efficiently using their existing infrastructure, and when should they invest in new infrastructure to meet demand? We propose a two-period model which utilises a third party called the Coordinator, who interacts with a population of resource-buyers. The Coordinator uses two mechanisms to aid the provider in future capacity planning. Firstly, the Coordinator extracts probabilities from the buyers through an options market to determine their likely usage in the next period, which can subsequently be used to schedule workloads. Secondly, the Coordinator uses previous market demand to predict if cost can be reduced by investing in a reservation over a longer period. This upfront investment contributes to the provider's capital expenditure in new capability and implies that Coordinator intends to further utilise such an investment. We implement the model in an agent-based simulation using actual UK market data where a pool of users submit different probabilities based on previous market demand. We show that the Coordinator can make a profit when faced with different market conditions, and that profit can be maximised by considering the utilisation of previously purchased reservations.

1 INTRODUCTION

Grid, cluster and, most recently, cloud computing have all promised to transform computing resources into a commodity, that can be delivered in a manner similar to that of existing utilities, such as electricity, gas, water and telephone services (Buyya et al. 2009). Cloud computing in particular is primed to deliver a new level of freedom to the consumer, allowing different levels of service and quality to be delivered on an as-needed basis without the need for capital investment

This *utility model of provisioning* gives users the ability to purchase computing resources as if they were any other commodity such as coal or steel. By providing a suitable mechanism for buying and selling, *market oriented computing* opens up a wide range of trading possibilities - CPU cycles, storage capacity, and memory allocations could be bought and sold, for current or future use. This is already happening to some extent in the marketplace, and a wide range of economic and resource sharing models for grids, clusters and clouds are public (Yeo and Buyya 2006; Hilley 2009). To fully realise this goal, however, providers must be able to interoperate so that consumers can move between providers easily and so that providers can utilise each other's capability when demand is high. This *federated cloud* is the ultimate aim of cloud computing (Buyya, Ranjan, and Calheiros 2010).

Currently, users purchase capability from the utility-computing provider directly. Problems of interoperability and lock-in are preventing consumers from being able to easily change supplier. Should standardisation be achieved, such a federated cloud would enable the use of centralised compute-resource "exchanges" and intermediary aggregators and brokers. This is not yet widespread but nevertheless seems likely to grow in significance over coming years.

These centralised mechanisms would enable a true Service Orientated Architecture where customer needs are matched to the most suitable computing

Rogers O. and Cliff D..
FORECASTING DEMAND FOR CLOUD COMPUTING RESOURCES - An Agent-based Simulation of a Two Tiered Approach.
DOI: 10.5220/0003717201060112
In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence* (ICAART-2012), pages 106-112
ISBN: 978-989-8425-96-6

resources using brokers or Coordinators. This would be controlled through Service Level Agreements (SLA) which would define the targets, for various metrics, (e.g. uptime, latency) that must be achieved; and would also define the compensation due to the customer if the targets are not achieved.

To meet these SLA's, the provider must ensure they have enough resources to meet demand; otherwise the provider will need to pay compensation to those customers whose performance criteria have not been met. Such a prediction will ensure adequate investment in new technology, and optimal utilisation of existing capacity.

The provider could obtain information on likely future requirements by letting users reserve resources through a derivatives market involving futures and/or options. A futures contract is a contractual agreement to buy or sell an asset for a certain price at a certain time in the future. An options contract gives the contract holder the right to buy, or sell, an asset by a certain date for a certain price, without a binding obligation to do so (Hull 2005).

It has been proposed that swing options, originally developed for trading electrical power, can be used to price a future reservation of computing resources (Clearwater and Bernando Huberman 2005). As with electricity, computing resources are non-storable and have volatile usage patterns, so such a model would provide customers with flexibility in terms of amount and duration of resource requirement, and enable resource providers to estimate demand.

Use of such derivatives presents two problems. Firstly, how can users accurately predict their future resource requirement? Determining and hedging their future demand for a resource is not an easy task; the variable nature of IT usage means that pricing the service so that competitiveness and profitability are balanced has an element of risk (Khajeh-Hosseini, Sommerville, and Sriram 2010)

Secondly, how can the user be trusted to submit a true representation of their likely resource requirements?

The first issue can be solved using a forecasting tool, such as that proposed in Clearwater (Clearwater and Bernando Huberman 2005) or by analysing historical market data such as that proposed by Sandholm et al. (Sandholm and Lai 2007). For the second issue, Wu et al. proposed a reservation model which was shown to lead to a truthful reservation on the user's part (Wu, Zhang, and BA Huberman 2008).

Wu et al.'s model involves a number of users who require the resource, plus a central authority ('the Coordinator') responsible for receiving and resolving resource requests. The Coordinator and users take part in a two-period game.

In this paper, we extend the model so that the Coordinator uses two mechanisms to predict future usage, while remaining profitable.

We create a practical implementation of the model, where market demand varies to typically observed dynamics using data obtained from the UK Government and where users have a degree of intelligence when submitting future resource requirements. Our objective is to determine if the model can be developed into a commercial offering, and be profitable in different market conditions.

2 BACKGROUND

2.1 Wu et al. Two Period Model

Wu et al. proposed a two-period model for resource reservation in which in the first period the user knows her probability of using the resource in the second period, and purchases a reservation whose price depends on that probability.

Consider N users who live for two discrete periods. Each user can purchase a unit of resource from a service provider to use in the second period, either at a discounted rate of 1 in Period 1, or at higher price C, where C > 1, in Period 2. In Period 1, each user only knows the probability that they will need the resource in Period 2 - it is not known for certain until the next period.

A third agent, the Coordinator, is introduced who makes a profit by aggregating the users' probabilities and absorbing risk through a two period game described below:

1. Period 1: Each user i submits to the Coordinator a probability, q_i, which does not have to be the real probability, p_i, that they will require a unit of resource in Period 2.
2. Period 1: The Coordinator reserves $q_i n_i$ units of resource from the resource provider at the discount price for use in Period 2, where n_i is the number of units of resource required by each user. For simplicity in this simulation, n_i =1 for all users.
3. Period 2: The Coordinator delivers the reserved resources to users who claim them. If the amount reserved by the Coordinator is not enough to cover the demand, the Coordinator purchases more from the resource provider at

the higher unit price C.
4. Period 2: User *i* pays:

$f(q_i)$ if resource is required
$g(q_i)$ if resource is not required

The contract can be regarded as an option if $g(q_i)$ is paid in Period 1 (i.e. as a premium), and $f(q_i)$ - $g(q_i)$ is paid in Period 2 (i.e. as a price) should the resource be required. In Period 1, the resource is reserved, but the user is not under any obligation to purchase.

Wu et al. showed that if the following conditions could be met, the Coordinator would make a profit:
• Condition A: The Coordinator can make a profit by providing the service.
• Condition B: Each user prefers to use the service provided by the Coordinator, rather than to deal with the resource provider.

The following truth-telling conditions are not completely necessary, but are useful, for conditions A and B to hold:
• Condition T1 (truth-telling): Each user submits his true probability in Period 1 so that he expects to pay the lowest amount later.
• Condition T2 (truth-telling): When a user does not need a resource in Period 2, it is reported to the Coordinator in the same period.

The following specific case was proved to meet these conditions, where k, a constant chosen to alter the price paid by the customer, is set to 1.5 and C is set to 2:

$$g(q_i) = \frac{kp_i^2}{2}$$

$$f(q_i) = 1 + \frac{k}{2} - kp_i + \frac{kp_i^2}{2}$$

2.2 Previous Simulations

In an earlier paper (Rogers and Cliff 2010a) we simulated the reservation model proposed by Wu et al., in a multiple-user, heterogeneous, variable-demand market. Through a simulation model, we showed that honesty benefits both the user and a Coordinator when the market varies uniformly, and that the user-base evolves to be more honest over time.

In a second paper (Rogers and Cliff 2010b) we extended our simulation, so that market demand did not vary uniformly, but instead underwent a period of high or low resource availability. It was found that the Coordinator benefits more when resources are in abundance, and less when resources are scarce. However, it was also found that when resources are abundant, the Coordinator does not always benefit financially as the honesty of the user-base increases. There is an optimum honesty that occurs when there is no surplus or deficit of resource purchased by the Coordinator.

3 RESELLING RESERVATIONS

Wu et al.'s model was found to be profitable amongst a group of heterogeneous users, and was found to promote honesty in the user-base.

However, the provider is only made aware of future demand one period in advance which may not be of any use for planning larger investments. If this information is used to plan additional capacity in the next period, the provider may have to make an investment in technology without having any guarantees regarding its longer-term utilisation.

To offset some of this risk taken by the Coordinator we propose a new model. The Coordinator now has the option of purchasing resources from the service provider using one of the following schemes:

- In Period 1, the Coordinator can purchase a *reserved instance*. A reserved instance gives the Coordinator access to a resource for a fixed term (36 months). The reserved instance costs a fixed sum at the beginning of the term, but gives the Coordinator access to the resource at a lower cost per unit time
- In Period 2, the Coordinator can purchase an *on-demand instance*. An on-demand instance is charged at a higher cost per unit time than a reserved instance, but there is no one-off cost.

This primary benefit of this approach to the provider is that they have a longer term view of future demand through the purchase of reserved instances by the Coordinator. In the short-term, information on likely utilisation in the next period could be used to efficiently schedule workloads on servers in an off-line fashion so that servers are fully utilised (Stage and Setzer 2009). Upfront payments received for reserved instances demonstrate to the provider that the Coordinator believes a resource will be utilised in the future. In the longer term, the provider can reinvest this upfront payment towards new infrastructure with at least some evidence that it will be paid back. Both sources of information could be used to calculate spot market prices.

The Coordinator is now a wholesale reseller of resource - the purchased reserved instances can be provided to whichever users need to use the resource in that period and wastage is reduced.

For the user, their expenditure is reduced as they can reserve a resource without having to pay full price should they not need to use the resource later. However, in our implementation of the scheme, the user must anticipate that she will take full advantage of the resource available to them during the month.

3.1 Methodology

To investigate the performance of the model, a computer simulation was constructed. The nature of the new model allows its performance to be evaluated using actual commercial cloud offerings and actual market conditions.

Period 1

1. Each user i submits to the Coordinator a probability, q_i, which does not have to be the real probability, p_i , that they will require a unit of resource in Period 2.
2. The Coordinator must reserve $\sum q_i n_i$ units of resource to be executed in the next month. For simplicity in this simulation, $n_i=1$ for all users.
   a. If the Coordinator has previously purchased enough reserved instances for the predicted demand, no further instances are purchased.
   b. If the Coordinator does not have enough resources available to meet the anticipated demand, it may need to purchase additional reserved instances. It will consider the performance of additional reserved instances over the past 36 months:

A = array [Last 36 months monthly resource demand]
B = array [Current resource capacity for next 36 months]
U = array [**A** – **B**]
Marginal Resource Utilisation (MRU) = (number of items in **U** > 0) / 36 months

The *MRU* is the fraction of the life of an additional reserved instance that will be utilised over the next three years based on past performance.

The *Threshold* is a ratio determined by the Coordinator to maximise profit.

   c. If *MRU* > *Threshold*, the Coordinator will buy a new reserved instance for 36 months at cost R as it is likely it will be used enough to make a return on the original investment
   d. If *MRU* < *Threshold*, it will be probably be more profitable for the Coordinator to buy an on-demand instance at cost D_h in Period 2.

Period 2

3. The Coordinator delivers the reserved resources to users who claim them. If the amount reserved by the Coordinator is not enough to cover the demand, the Coordinator purchases more from the resource provider at the on-demand instance cost D_h. For the reserved instances, the reduced cost of R_h is paid.
4. User i pays
   $f(q_i)$ if resource is required
   $g(q_i)$ if resource is not required
   where $f,g : [0,1] \rightarrow R^+$

3.2 Agent-based Simulation

A computer simulation was programmed in Python and for each of the market segments shown in Table 1, a simulation was implemented with 1000 users. Each simulation was run 100 times with a different threshold, between 0 and 1, in 0.01 increments.

The simulation was prepared with the following characteristics:

3.2.1 Market Demand Data

Datasets were obtained from the UK National Statistics Office on the Non-Seasonally Adjusted Index of Sales at Current Prices from 1988 (earliest available) to 2011 for four different market segments, as shown in table 1. These segments were chosen as they have a strong relationship to IT usage and they vary differently over the period, therefore allowing the model to be simulated across a wide range of market conditions. These were normalised between 0, where none of the N users submit a resource request, and 1, where all N users submit a resource request. The period of these statistics represents a typical period of modern times where demand has changed frequently, with both periods of recession and growth. As such, it is a suitable model of market variance.

3.2.2 User Agents

In the first period, the user will submit a probability based on the market demand in the same month from the previous year. The probability is chosen at random from a uniform distribution between the previous year's market demand and 1. This approach means that when a high market demand was experienced during the same month in the previous year, more users will submit a high probability to the Coordinator, than when market demand was low.

3.2.3 Service Provider Agent

The resource being purchased is an Amazon Web Services EC2 Standard Small Instance (US East). At the date of simulation (July 2011), these were being advertised at a cost of D_h = \$0.085/hour for an on-demand instance, and R = \$350 plus R_h = \$0.03/hour for a 36 month reserved instance.

3.2.4 Pricing Structure

Users are charged a price based on the values of $f(q_i)$ and $g(q_i)$ suggested by Wu et al. However, as the standard monthly on-demand cost charged by the service provider is around \$60, the Coordinator can charge the user anything up to this value such that condition B is met. To achieve peak profit while ensuring the Condition B is met, the Coordinator increases $f(q_i)$ and $g(q_i)$ by a factor of 60.

4 RESULTS

Plots of annual profit for each of the four segments over time with no optimisation and maximum optimisation are shown in figures 1 to 4 in the Appendix. Plots of customer demand and capacity reserved by the Coordinator for each of the four segments over time with no optimisation and maximum optimisation are shown in figures 5 to 8 in the Appendix.

Table 1: Profit increases for market segments.

	Profit £M - Period			
	No Opt	Max Opt	+/-	Opt Thre
Non-store retail: All businesses	3.62	4.63	28%	78%
Retail: IT Equipment	5.02	5.92	18%	93%
Non-store retail: Small businesses	4.26	5.15	21%	82%
Non-store retail: Large businesses	4.10	5.05	23%	81%

Table 1 shows the profits achieved by the Coordinator over the period when there is no optimisation (threshold=0) and when there is maximum optimisation (when threshold is set at that which produced the maximum profit).

Table 1 shows that the Coordinator makes a profit even when not optimising across the four market profiles, which implies that the Coordinator is likely to survive and prosper in a variety of conditions. The total profit is related to the demand of the market and the accuracy with which the Coordinator predicts future usage.

Figure 8 most clearly shows the Coordinator tracking changes in market demand, but a similar pattern can be seen in figures 5 to 7.

When the Coordinator's threshold is set to 0, annual profit generally varies with market demand as shown most clearly in figures 2 and 6, but cycles every 3 years due to the need to buy additional reserved instances whenever a deficit is anticipated.

However, the Coordinator regularly reserves more resources than are required. This is due to users submitting probabilities based on previous performance as a way of guaranteeing access to a resource in the event of high demand – this is shown as the difference between the resources demanded and the capacity available in figures 5 to 8.

From table 1, it can be seen that a significant increase in profits can be made by considering past performance before deciding to invest in a reserved instance.

It is common sense that the Coordinator will profit most when there is a large demand for resources which has been fully anticipated by the Coordinator. This means that all resources are delivered to the users using the cheaper reserved instance rate, and no new resources must be purchased at the higher on-demand rate. It also means that advance purchases of resources are being wasted. The profit is therefore maximised when the Coordinator is able to predict future demand most accurately.

When the threshold is set to the optimum threshold achieved during simulation, we see that the profit stabilises and no longer cycles as in figures 1 to 4. The Coordinator now only buys reserved instances when it believes it will be used enough times to payback, and thereby reduces expenditure and maximises profit.

5 CONCLUSIONS

This paper has shown how modification of a truth telling reservation model for computing resources described by Wu et al. can provide the basis for a real-world implementation of an options market in a federated cloud which is price-competitive for the user, profitable for the coordinator and beneficial to the service provider.

An extension of Wu et al.'s model was implemented in an agent based simulation using actual data on consumer demand over a typical

period in modern history, using costs of an Amazon Web Service cloud instance, and where users submit probabilities based on previous demand. It was found that the coordinator profits in such a situation in a number of market segments, thereby demonstrating that a stable commercial implementation is feasible.

It was also found that the Coordinator is better off considering past performance when decided to invest in another reserved instance, and this can increase profits by up to 28%.

Wu et al.'s model provides a suitable theoretical model for an options-market in computing resource. However, the service provider would have to provide specific pricing to support the Coordinator, and this might not always be profitable for the service provider. Our extension to this model does not require new pricing to be agreed, but contract restrictions on reselling may be a barrier to commercial implementation.

Our work shows that a probability-based options market in computing capability is a viable commercial proposition, and that all parties can potentially benefit as a result of such a system. The advantage of this approach is that a forecast of future usage requirements is obtained, which can subsequently used to plan future capacity requirements and so that targets on performance as detailed in a Service Level Agreement can be met. These are currently issues for widespread federated cloud adoption.

The simulation has shown that the reservation model may be suitable for real-world application. The model provides a platform for further risk assessment work to be undertaken and, as discussed, the simulator can be further extended to simulate a variety of market conditions, or specific user demands.

The optimum threshold is the value at which market demand is fully anticipated by the Coordinator, and which is fully provisioned through reserved instances. Determining this threshold mathematically is likely to be challenging due to difficultly in determining market dynamics over a very long period. However, an empirical simulation using actual market data could produce such a threshold for commercial implementation.

By taking the results from this paper and extending them with future research into the performance of the model under different conditions and inherent honesties, in different segments, a commercial offering that is profitable to the coordinator, beneficial to the user, and with a calculated level of risk looks likely to be achievable.

ACKNOWLEDGEMENTS

We thank the UK EPSRC for funding the Large-Scale Complex IT Systems Initiative (www.lscits.org) as well as HP Labs Adaptive Infrastructure Lab for providing additional financial support.

REFERENCES

Buyya, R., Ranjan, R., & Calheiros, R. N. (2010). InterCloud : Utility-Oriented Federation of Cloud Computing Environments for Scaling of. Network, 13-31.

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems, 25(6)

Clearwater, S. H., & Huberman, Bernando. (2005). Swing Options : A Mechanism for Pricing IT Peak Demand. Proceedings of 11th International Conference on Computing in Economics.

Hilley, D. (2009). Cloud Computing : A Taxonomy of Platform and Infrastructure-level Offerings Cloud Computing : A Taxonomy of Platform and Infrastructure-level Offerings. Technology, (April).

Hull, J. C. (2005). Fundamentals of Futures and Options Markets.

Khajeh-Hosseini, A., Sommerville, I., & Sriram, I. (2010). Research Challenges for Enterprise Cloud Computing. Arxiv preprint

Rogers, O. and Cliff, D. (2010a), The Effects of Truthfulness on a Computing Resource Options Market, Proceedings of the Conference on Advances in Distributed and Parallel Computing

Rogers, O. and Cliff, D. (2010b), The Effects of Market Demand on Truthfulness in a Computing Resource Options Market, Proceedings of the 3rd International Conference on Agents and Artificial Intelligence.

Sandholm, T., & Lai, K (2007), A statistical approach to risk mitigation in computational markets. Proceedings of Conference on High Performance Parallel and Distributed Computing

Stage, A., & Setzer, T. (2009). Network-aware migration control and scheduling of differentiated virtual machine workloads. 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing, 9-14.

Wu, F., Zhang, L., & Huberman, Ba. (2008). Truth-telling reservations. Proceedings of 11th International Conference on Computing in Economics.

Yeo, C. S., & Buyya, R. (2006). A taxonomy of market-based resource management systems for utility-driven cluster. Cluster Computing, (June), 1381-1419. doi: 10.1002

APPENDIX

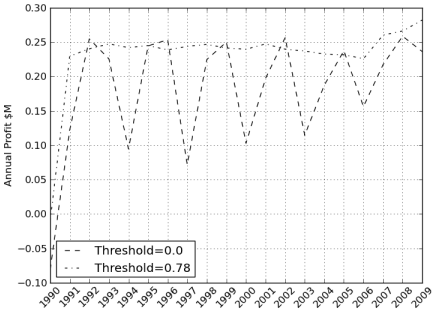

Figure 1: Non-store retail profit.

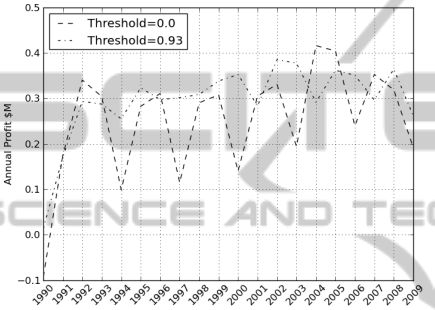

Figure 2: Computer equipment profit.

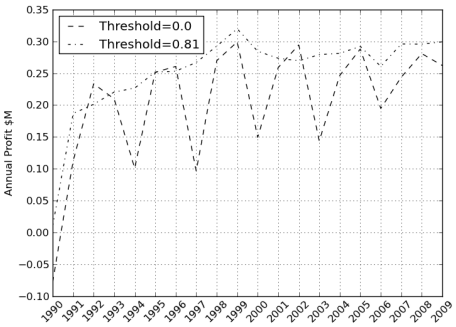

Figure 3: Non store, small business profit.

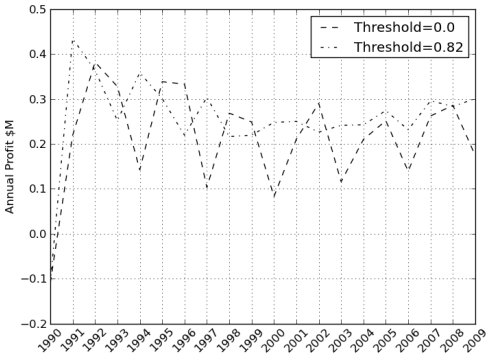

Figure 4: Non-store, large business profit.

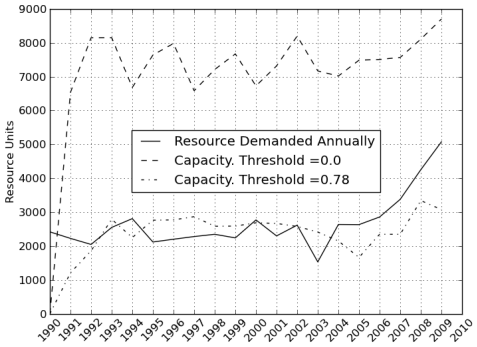

Figure 5: Non-store retail units.

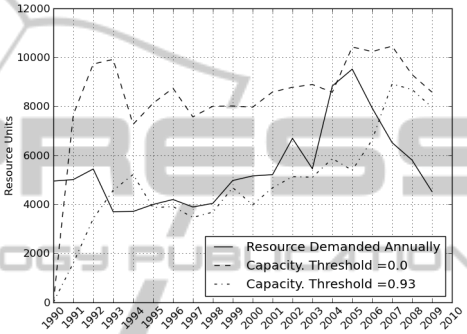

Figure 6: Computer equipment units.

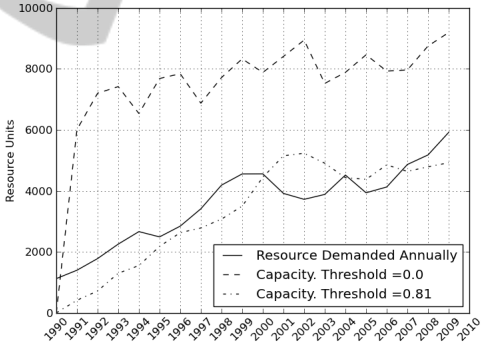

Figure 7: Non-store, small business units.

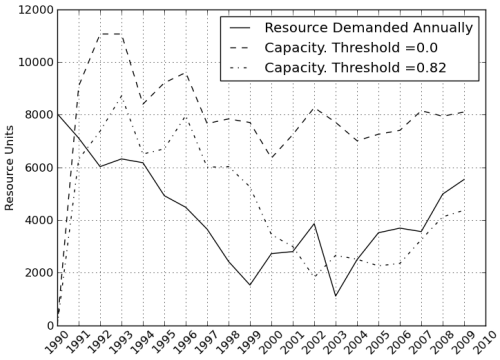

Figure 8: Non-store, large business units.