Microeconomic Theory Applied to Distributed Systems

David Clarke and Brendan Tangney
Department of Computer Science,
Trinity College Dublin.

Distributed Systems Group
Department of Computer Science
University of Dublin
Trinity College, Dublin 2, Ireland.
Fax: +353-1-772204

Document Status Published
Distribution Public
Document # TCD-CS-93-30

Microeconomic Theory Applied to Distributed Systems

David Clarke and Brendan Tangney *
Department of Computer Science,
Trinity College Dublin.

February 3, 1994

Abstract

This report explores how well-known techniques from micro-economics can be used to interpret and solve aspects of resource allocation problems in distributed systems.

The principle features of the relevant economic theory are described before the report goes on to investigate how the theory can be used to solve an instance of the load balancing problem in a distributed system.

1 Introduction

The idea of modelling entities which compete for scarce resources in a distributed system as an instantiation of a free market is an intuitively appealing one as it provides a straightforward and well-understood chassis upon which to base an approach to resource allocation problems. In this report some basic theory from micro-economics is outlined. It is then shown how that economic theory can be mapped onto a distributed system. As a concrete example of the mapping in action an instance of the load balancing problem is solved.

The layout of the report is as follows. Firstly a more detailed rationale for the use of concepts from economic and market theory in distributed systems is presented. In section 3 the basic economic theory being used is explained. It is followed in section 4 by a description of how that theory can be mapped onto a resource allocation problem in a distributed system. Section 5 outlines directions in which the ideas could be extended in order to help solve more complicated problems. Finally previous applications of economics to the load balancing problem are described and assessed.

*email: brendan.tangney@dsg.cs.tcd.ie

2 Rationale for an economic approach

The principal underlying reason for adopting an economic approach to resource allocation is the conceptual simplicity of the scheme and the fact that economics provides a large body of work which can potentially be called upon to help solve problems in distributed computing. The notion of economic agents competing in a free market is inherently attractive and is empirically proven to allocate resources efficiently in an environment of imperfect information, such as that which pertains in a distributed system.

Adam Smith [Smi76], the seminal economist and free marketeer, posited the idea of an 'invisible hand' acting in the marketplace. This was his explanation of the phenomenon whereby without any explicit authoritarian intervention, and with all the agents in the marketplace selfishly attempting to maximise their own utility, market equilibrium was somehow brought about.

An analogy in a distributed systems environment is that as long as *consumers* (e.g. processors) demand *goods* (e.g. jobs), then fulfilling that demand may be regulated using a price mechanism with the action of the market ensuring efficient allocation.

The similarities between economic systems and distributed computer systems suggest that mod-

els and methods previously developed within the field of economics can serve as blueprints for engineering similar mechanisms in distributed computer systems, [KS89]. The benefits of controlling resource usage in a distributed system using microeconomic methods are many. The methods are well understood, are already tested and are proven to be effective. Simplicity, modularity, and scalability are achieved by treating processing nodes as individual agents in an interacting society. Finally, the use of a price mechanism provides a simple and decentralised framework which [KS89] notes:

> '...May reduce the complexity of designing and implementing the large distributed systems of the future.'

3 Economic theory

This section presents the micro-economic theory being used, a more detailed discussion of which can be found in [Cha86]. Section 3.1 explains the interaction of demand and supply to achieve equilibrium; section 3.3 presents the theory of an individual consumer, and from this derives his demand curve; while section 3.8 presents the theory of the economic firm.

3.1 Demand and supply

Economic goods and price Chacholiades [Cha86] defines economics as:

> 'The science of choice under scarcity.'

An *economic good* has two essential properties: scarcity and usefulness. If something is useful but not scarce, for example air or water, then consumers will be unwilling to pay for it. Conversely, something which is scarce but not useful will not attract economic activity either: most rational individuals would be unwilling to pay for a square wheel.

For those goods which are both scarce and useful their price is a reflection of both the valuation attached to the good by consumers, and also the relative levels of supply and demand. The important point here is that 'price' is not some absolute, but rather a floating mechanism which fluctuates to equalise supply and demand.

3.1.1 Demand

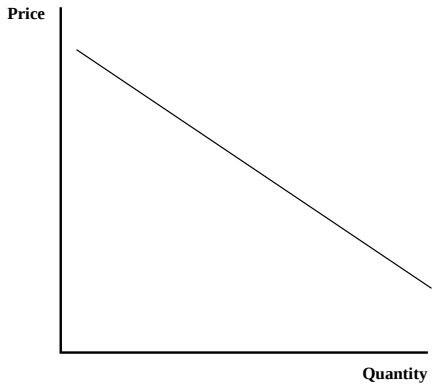

Figure 1: A demand curve

Figure 1 illustrates the demand curve of a consumer for a typical commodity. Price is marked on the Y-axis, quantity demanded on the X-axis. The curve slopes downwards, reflecting the fact that as commodities grow cheaper, people tend to buy more of them, and vice-versa. Such goods are known as *normal* goods, and the commodities considered in this paper are treated as normal.[1]

What determines demand? Demand is modelled as a multivariate function, normally with four components:

$$D = f(Px, Py, Y, Tastes)$$

The components are:

Px The price of the good.

Py The price of substitute good. If X and Y are very close substitutes, then the number of consumers who 'change over' from X to Y when the price of X increases will be large.

[1] Some goods, usually necessities or luxuries, have different shaped curves but these are not of interest here.

Y The consumer's income.

Tastes The tastes of the consumer obviously dictate the levels of consumption of a particular good.

Aggregate demand The aggregate demand for a commodity is just the sum of the quantities demanded by each consumer in the market at each price level.

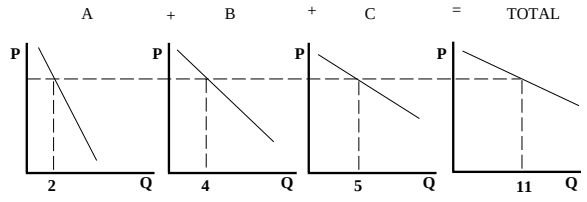

Figure 2: Aggregate demand

The situation is depicted graphically in figure 2. At a price level of 5, A demands (say) 2 widgets, B demands 4, and C 5. The aggregate demand is thus 11 widgets at a price of 5. This process is repeated for all consumers at each possible price level, and this generates the overall market demand curve.

3.1.2 Supply

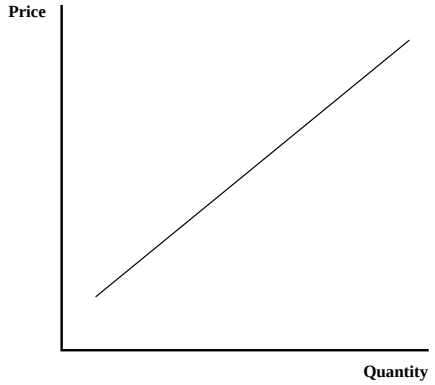

Figure 3: A supply curve

The supply curve depicted in figure 3 is a typical supply curve faced by a firm which has decided to manufacture or market a particular commodity. It may be viewed as a 'willingness schedule', which indicates how many units of output the firm will produce at each level of price. It slopes upwards, reflecting the fact that as the market price increases, the firm is willing to produce more and more of a good. This is because his raw material costs don't change, so if the market price is rising, then the supplier is making a growing profit on each unit sold. Given that the supplier could in principle operate at any point on the supply curve, what determines his actual level of output? Clearly two things are important.

- The level of demand for the good. The interaction of demand and supply to establish market price and quantity is discussed in section 3.2.
- The costs faced by the enterprise. The costs are comprised of *fixed* costs (land, electricity etc), and *variable* costs (raw materials, labour etc.), and the nature of these costs serves to constrain the acceptable levels of production. Costs are considered further in section 3.8.

3.2 Market equilibrium

Now that demand and supply have been studied in isolation, the question of how they together determine the market equilibrium must be analysed.

Consider figure 4. This overlays the industry demand and supply curves on one plane. If the initial market price is set at P1, then clearly the level of supply exceeds the level of demand, causing prices to drop, as the good is not sufficiently scarce to justify its price.

If the price now falls to P2, then demand exceeds supply, so consumers are willing to pay more money to secure the commodity. Thus the price rises. Dynamical considerations [Sam92] dictate that the price eventually settles at Peq, representing the market clearing price, at a quantity of Qeq. At this price level, demand equals supply, and the market clears. There is no dynamical incentive to deviate from this position in the absence of a shift in demand (owing to a change in tastes etc.), or supply.

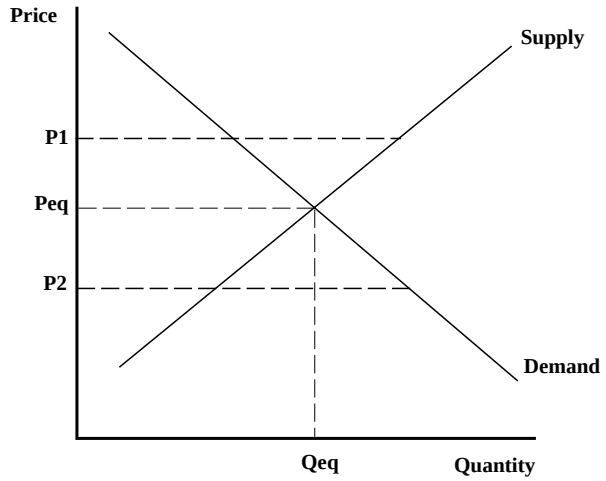

Figure 4: Market equilibrium

It is worth clarifying an earlier reference (section 6) to the 'law of one price'. This states that the price of a traded commodity tends to uniformity throughout a closely-linked market. The action of arbitrage tends to ensure this. Thus in any market with perfect consumer information (usually assumed), the prices of individual goods should not vary from place to place, but settle at the market equilibrium.

3.3 Consumer theory

The demand curve is a concise and informative graphical summary of a consumer's consumption schedule. However, it cannot simply be plucked out of thin air. Any consumer would most likely be unable, if asked, to rapidly and accurately sketch out demand curves for a number of goods, and indeed it is unreasonable to expect the consumer to do so. Instead, from information regarding *tastes* and *preferences*, it is possible to construct a demand curve.

The purpose of this section is to outline how consumer choices are modelled in economics, and how to move from statements of preference to demand curves. The first step in this process is to find some way of quantifying a consumer's preferences.

3.4 Utility

When an economic good was defined earlier (section 3.1) as having the properties of *scarcity* and *usefulness*, no attempt was made to quantify 'how useful' a particular good was deemed to be, or how the relative usefulness of goods was to be judged. The economic measure of usefulness or satisfaction is known as *utility*.

The notion of **ordinal** utility requires that the consumer be able to make *relative* judgements between 'bundles' or baskets of commodities, e.g. 'I prefer three apples to two oranges'. While ordinal utility is accepted as being the most useful version of utility in modern theories an earlier notion of **cardinal** utility, where consumers are required to assign an absolute utility values to goods, may actually be more appropriate for a computer system.

3.5 Axioms of rational consumer behaviour

Before presenting an analysis of consumer equilibrium under conditions of ordinal utility, it is necessary to present the assumptions governing rational consumer behaviour. It is worth noting that although human consumers rarely fulfil these assumptions, the behaviour of a computer executing a set of instructions in software and hardware will be systematic, and fully consistent. Its preferences are timeless, and not subject to any of the normal vagaries of individual choice. It is thus perfectly valid to attribute *rational* behaviour to computers.

With humans elaborate psychometric testing is unnecessary to ascertain rationality. The economist, however, makes no such psychiatric value judgements, rather, a consumer is deemed 'rational' if his behaviour is consistent with the following axioms:

- **Rationality**
  The consumer aims at maximisation of his utility, given his income and market prices. It is assumed that he has full knowledge of all relevant information.

- **Ordinal utility**
  It is taken as axiomatically true that the consumer can *rank* his preferences according to the relative satisfaction afforded by consuming the various commodities.

- The total utility of the consumer depends on the quantities of the commodities consumed:

$$U = f(q1, q2, ..., qn)$$

- **Consistency and transitivity of choice**
  Consistency of choice implies that if the consumer prefers good A to good B at any time, then he must always prefer good A to good B. Transitivity implies that if he prefers A to B, and B to C, then he must also prefer A to C. Strong ordering is also assumed; any two bundles of goods are comparable.

- **Nonsatiation**
  Nonsatiation states that the consumer must always prefer more to less.

From these assumptions, a graphical model of utility and choice may be derived, which is, as has been noted, particularly appropriate to an environment where the computers are the consumers.

3.6 Indifference curves and maps

For simplicity, when analysing consumer behaviour, it is normal to consider choices between bundles of only two types of goods[2]. Graphically, this situation may be represented as in figure 5.

Points A and B on the diagram correspond to two *choice points* in the overall choice space. They each represent a possible basket or combination of goods X and Y. Given the earlier assumptions of strong ordering of consumer preferences, there are three possibilities, that the consumer:

1. prefers bundle A to bundle B.
2. prefers bundle B to bundle A.

[2]It is easy to generalise this to all goods by letting good Y be 'all goods except good X'.

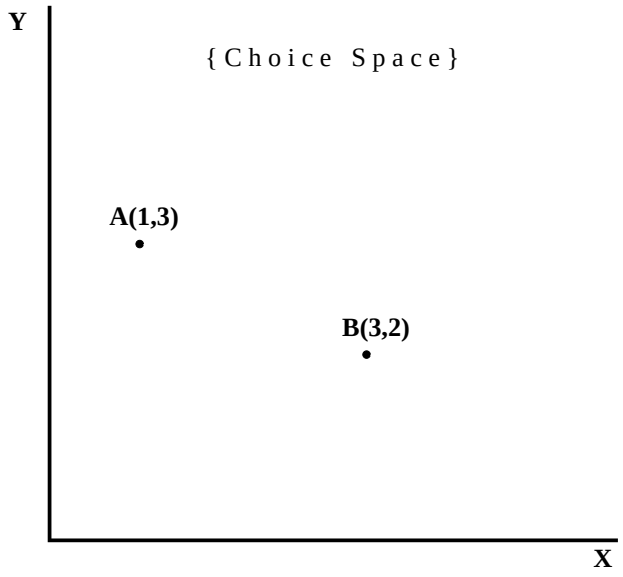

Figure 5: Choice points

3. is *indifferent* between the two bundles - values them both equally.

By systematically interrogating the consumer about his (ordinal) preferences, one can devise sets of bundles, the elements of which he is indifferent between. He will, however, prefer some sets to others - in particular, he prefers sets containing *more* items to those containing less. The locus of a particular set of bundles between which the consumer is indifferent, i.e, which yield him the same level of utility, is known as an *indifference curve*.

Figure 6 displays some indifference curves. With reference to this diagram, the consumer is indifferent between a bundle containing 6 of Y and 1 of X, and a bundle containing 2 of Y and 3 of X - they yield him equal satisfaction, and he cannot choose between them on the basis of utility alone. Note, however, that he would prefer a bundle containing 3 Y and 4 X to either of the other bundles - or indeed any other combination on indifference curve A. All points on B yield higher satisfaction than any on A, and similarly for all points on curve C relative to curve B[3]. Thus the consumer prefers

[3]And therefore, indifference curves may not intersect -

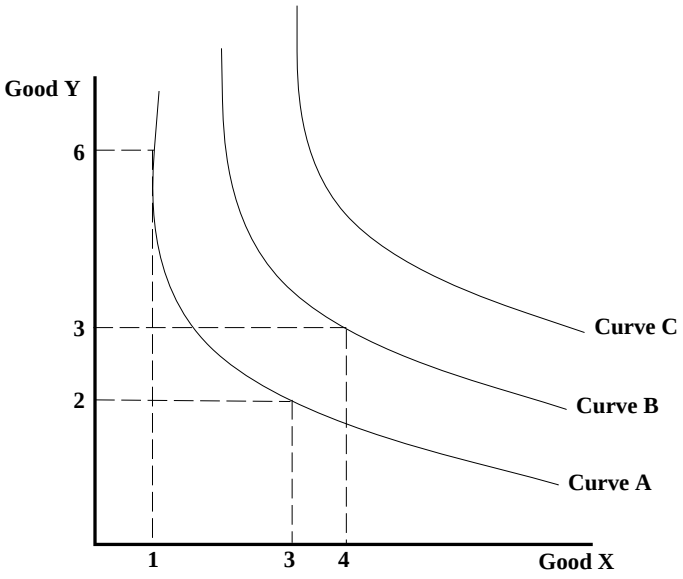

Figure 6: Indifference curves

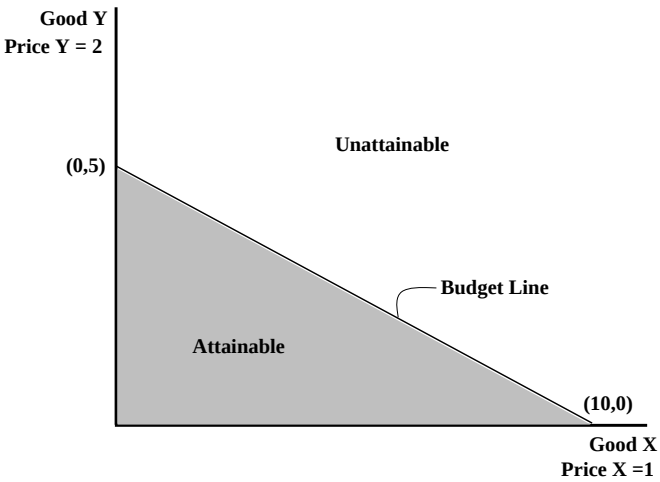

Figure 7: the budget line (Income = 10)

being on a *higher* indifference curve.

The whole set of indifference curves in the choice space is known as the *indifference map* for the consumer.

The budget constraint Now that the consumer's indifference map is known, the question arises as to how his level of demand is to be ascertained: on which particular indifference curve will he be at equilibrium, and which 'bundle' on this curve will he choose?.

While nonsatiatable consumers might prefer 50 books and 50 videos to 6 books and 6 videos most of them also lack the income to support such extravagant consumption. Two options therefore face the impoverished consumer: to acquire the extra 44 books and videos in an illegal fashion (highly inadvisable), or to settle at the equilibrium point between desires and means (much more socially acceptable). The situation is depicted graphically in figure 7. The budget line for a consumer partitions his choice space into 2 segments: attainable bundles (that which he can afford), and unattainable ones (that which he can't).

In figure 7, the budget space for a consumer with an income of 10 units is shown. Commodity Y costs 2 units, commodity X costs 1. The limiting bundles are therefore 5 Y and 0 X, and 10 X and 0 Y, both costing the full 10 units of income. Joining these extreme points defines the budget space.

All that is required to locate the consumer equilibrium is to overlay the budget line on the indifference map, as in figure 8, and identify the point at which the budget line intersects the highest possible indifference curve (recalling that consumers prefer higher indifference curves).[4] This corresponds to the 'optimal' level of consumption which the consumer can afford, and he is at equilibrium at this point.

3.7 From indifference to demand curves

So where does the demand curve come from? Considering again figure 8, it is clear that if the price of good X were to drop, with income and the price of good Y remaining constant, then the consumer's budget line would pivot *outwards*, intersecting with a higher indifference curve, and corresponding to a higher level of consumption of good X. As the price

to have two levels of utility or satisfaction simultaneously is meaningless.

[4] The intersection will in fact be tangential. [Cha86]

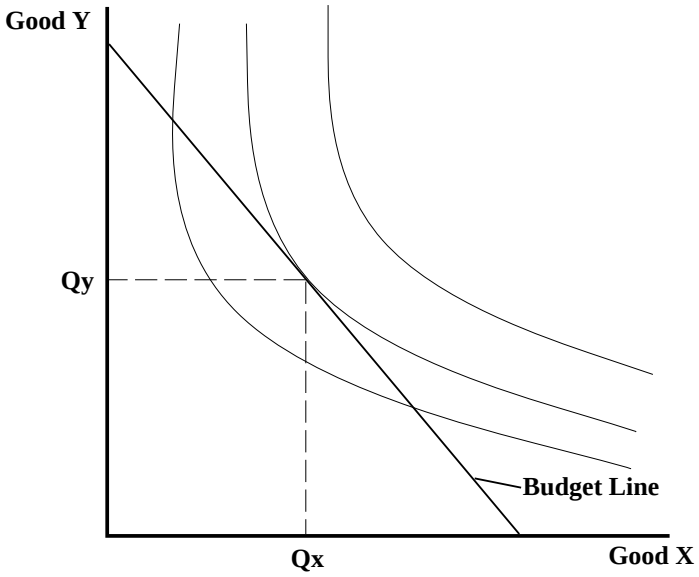

Figure 8: Consumer equilibrium

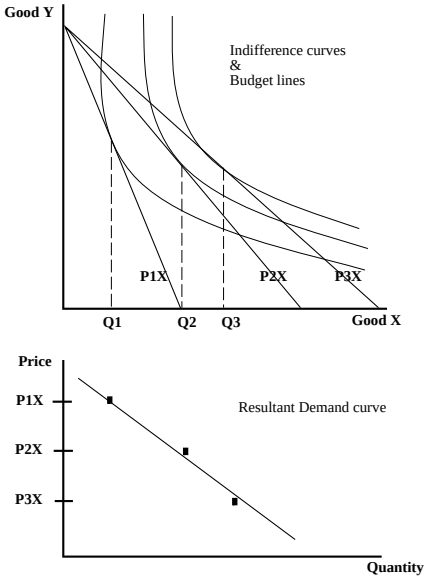

Figure 9: Deriving the demand curve

falls further, the effect is repeated. By plotting quantity against price, as in figure 9, the demand curve results.

This analysis provides a solid behavioural underpinning for the shape of the demand curve which was postulated in section 3.1.1.

3.8 Market structures

Now that the theory of consumer and market demand has been presented, it is necessary to outline some characteristics of the supplier model.

The model appropriate to the problem tackled in the next section is monopoly and the situation faced by monopolists is straightforward. Since the firm *is* the industry, the demand curve faced by the firm is the aggregate market demand for the product. Given a demand curve, the monopolist may choose to set *either* the price, *or* the quantity. The other is yielded by the demand curve. The decision of whether to set price or quantity, and at what level, may be made in a variety of ways: the price or quantity decision may be strategic; a regulatory body may force the monopolist to sell at a certain price, or he may attempt to produce at the point which maximises his profit. Any of these are viable.

3.9 Summary

This section has presented the economic theory underlying the model used. Section 3.1 described demand, supply, and their interactions in bringing about market equilibrium. Section 3.3 developed the idea of utility and the rational consumer, leading to a derivation of the demand curve from an indifference curve perspective. Finally, section 3.8 introduced the economics of a monopolistic firm. This theory is applied in the in the following section to propose an economic model to solve a problem in load balancing.

4 Applying the theory

This section maps the economic theory described in the previous section onto a model for the load balancing of parallel jobs of the form "do these N independent pieces of work and report the results back". Many problems in parallel computing are of this form and the particular example used

here is the parallelisation of the UNIX *Make* utility. The motivation for this is not so much to parallelize Make, which has been done already, but to see how well the economic model handles a typical distributed systems resource allocation problem.

Section 4.1 provides an interpretation for price and quantity in a distributed system, and also describes the mapping of the distributed system onto the monopolistic market structure explained in section 3.8. Section 4.2 extends the indifference curve analysis presented in section 3.6. A model of processor income allocation is sketched in section 4.3, and related to local and global system load. The ensuing sections derive demand curves for each 'consumer' in the market, aggregate this demand with respect to the monopolistic market model, and assess the quantity demanded by each consumer at this price.

4.1 What are 'price' and 'quantity'?

Any market consists of *buyers* and *sellers*. In the case of a parallel Make operation, there is *one* seller (the machine on which the Make was originated) and this machine wishes to sell a commodity (files to be compiled) to the buyers - the other nodes on the network. Thus the 'quantity' of the good demanded by each of the nodes in the network is the number or size of files that it is willing to 'buy' - and hence compile - at the price determined by the market. The quantity demanded by each machine is determined with reference to its demand curve at the market price. The demand curve is itself determined, as has been shown, by income levels and indifference curves. Thus if an income allocation strategy is adopted which reflects system **load** - i.e., if lightly loaded nodes have more money than heavily loaded ones - and **processing power** - i.e. powerful processors have more income than weaker ones - then these nodes will purchase more of the available files at the market price (recalling that the higher the income, the greater the quantity demanded at each price). **Thus the majority of the work will,** *inter alia***, migrate to the less busy nodes, effecting load sharing through the medium of the market.**

Note that price is merely an abstract mechanism (determined with reference to aggregate demand and quantity supplied), which determines levels of individual consumption by defining what can and cannot be afforded, while quantity is a measure of the number or size of the files to be compiled.

The bulk of this sections now proceeds to illustrate how income, price, and quantity are to be determined.

4.2 The shape of a processor's indifference curve

As described in section 3.6, the indifference curve of a consumer is a measurement of his attitude towards various baskets of two goods, and may be generalised to the n-good case. One possible approach to adopt in a distributed system is to consider the 2-good case:*local* and *remote* jobs. Thus the indifference curve will reflect the node's preference for either local or remote jobs.

So what are these preferences? Clearly, from the processor's viewpoint, there are none. It makes no difference whether it processes jobs that originated locally or remotely, **provided that it is always kept busy if there is work available.** It is necessary at this point to add the following to the axioms of rational consumer behaviour enumerated in section 3.5:

> **The laziness hypothesis:**
> 'Computers would rather process any job than be idle.'

To determine the resultant shape of the indifference curves, given these assumptions, a special case of indifference curves - for closely substitutable goods - is considered.

4.2.1 Substitute goods

Substitute goods have already been referred to (section 3.1.1). If goods are highly substitutable, for instance two brands of floppy disk, then they have a high positive cross-elasticity of demand. If

they are perfectly substitutable, then the cross-elasticity of demand becomes infinite.

Ordinary indifference curves (e.g. figure 6) have a curved shape, because the slope of the tangent at any point on the indifference curve represents the *marginal rate of substitution* (MRS) of X for Y[5]. This obviously varies along the indifference curve, with the slope *increasing* as consumption of good X increases. This is because, as consumption of X increases, the consumer becomes increasingly willing to trade more items of X for one of Y, as he is approaching 'satiation' with good X, and would like some variety.

If two commodities are close or perfect substitutes, this should not happen. Since the goods are virtually indistinguishable, it shouldn't matter to the consumer which he chooses. Thus MRS(X,Y) remains constant throughout the choice space, so the slope of the tangent to the indifference curve must also remain constant. This necessarily implies that the indifference curve between substitute goods is a straight line with negative slope, of the form Y = -X + C, as in figure 10.

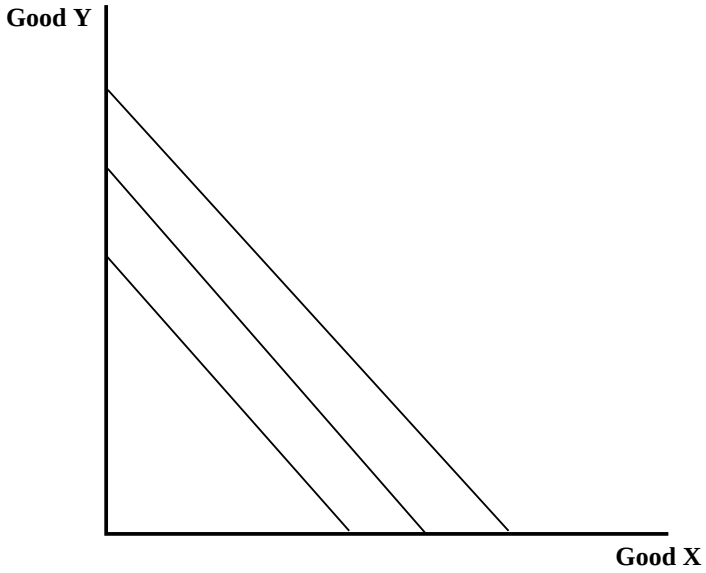

Figure 10: Indifference curves for perfectly substitutable goods

As has been said, processors don't care whether they process local or remote jobs; thus they may be viewed as freely substitutable commodities. Equally, all nodes have these same characteristics (by virtue of the laziness hypothesis), and so the indifference curves all have the form of figure 10. A linear indifference curve can lead to a 'corner solution', or 'mono-mania', in which a consumer spends all of his income on one commodity. Such behaviour is extremely uncommon amongst (sane) humans, but unremarkable in a machine context - they follow their behavioural patterns unflinchingly.

4.3 Income and system load

Load conditions on nodes around a distributed system vary considerably owing to erratic job inter-arrival times. Nodes which are heavily loaded at a given time should not have enough income to 'buy' much extra work, and vice-versa. Thus the income allocation policy should reflect system load by making a monetary endowment to each node which is proportional to its load relative to other nodes. For example, if there are 100 units of income to be allocated between three nodes with loads of .5, .4, .1 respectively, then clearly the incomes should be 50, 40, and 10 units respectively. It is also possible to weight the income strategy so that more powerful processors in a heterogeneous environment get more money, to reflect their greater processing ability. Equivalently, load statistics could be interpreted as a percentage of capacity, rather than an absolute value.

How is the total amount of money which is to be distributed as income determined? Figure 11 illustrates a *closed economy*, in which all income derives from consumers working for one firm, producing one type of good which they all purchase and then consume, with the money returning to the firm, and the cycle re-commencing. In a situation like this, there is just enough money in circulation to buy up all goods produced (management at the plant may be paid more than workers, and hence

[5] MRS(X,Y) is interpreted as 'How may units of X am I willing to forego, in order to have an additional unit of Y?'

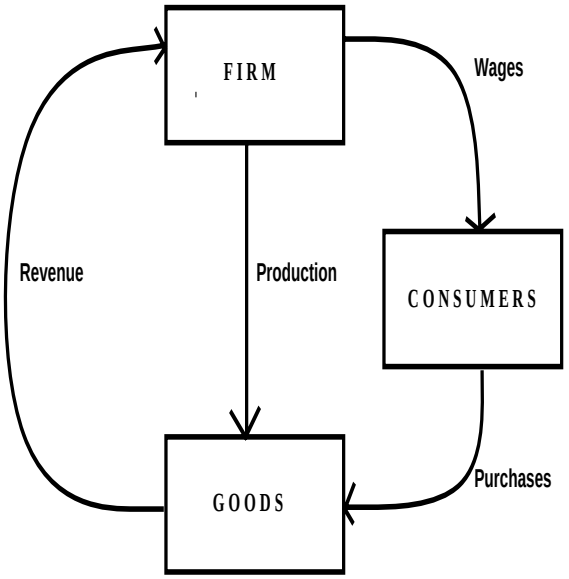

Figure 11: A closed economy

buy more of the product).

The way in which to simulate this closed economy model for the type of problems under consideration here is to stipulate some price for each unit (file/byte), and ensure that the total amount of money (processor incomes) is just sufficient to allow all the output (files/bytes) to be consumed (compiled). For example, if there are 5000 bytes of source code to be compiled, the price level may be arbitrarily initially set at 1 unit=1 byte, and thus 5000 units of income may be distributed across the network in proportion to the various loads, as described earlier.

In practise it is not necessary to divide the total amount of money out in this fashion. Rather, it suffices to give *every* node enough money to buy *all* the bytes at a price of 1 unit=1 byte. The amount of money actually allocated varies linearly with load. The effect of there being more money in circulation than is strictly necessary has the effect of causing *inflation* - a rise in the price level. Thus the price per unit is higher than it would have been if the total amount of money (income) in the system was just enough to buy all the bytes at the price of 1 unit=1 byte. Recalling that the price level is arbitrarily determined, and its absolute numerical value has little or no significance (the government could declare tomorrow that one pound consisted of 10,000 pence. Beans would immediately 'shoot up' to 3,000 pence, but their inherent value hasn't changed - just the price tag), this is perfectly acceptable.

In the spirit of ordinal utility, the *relative* differences between the purchasing powers of the nodes remain unaffected by changes in the *absolute* monetary allocation policy, and it is the relative ordering that captures the notion of lightly-loaded processors being richer than the others. Thus the income policy reflects system load conditions accurately.

4.4 Demand curves revisited

Knowing the processor income (and hence the shape of the budget line) and also the shape of the indifference curves, it is now possible to use the theory outlined in section 3.7 to derive a processor's demand curve.

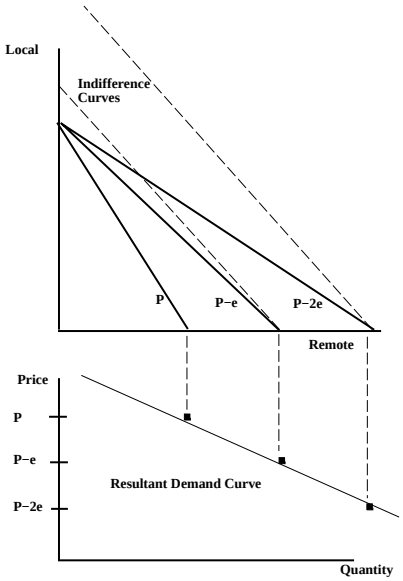

Figure 12: Deriving processor demand curves

The initial situation is as depicted in figure 12. The choice space is everywhere dense with indifference curves of the form Y = -X + C. The ini-

tial budget line, corresponding to Price(local) = Price(remote) : 1 unit=1 byte, is also of this form. The price of the remote bytes is then varied, with the local price and the income remaining the same, The levels of demand are sampled at a number of levels: P, P-e, P-2e, ..., P-ne, corresponding to various different prices of remote bytes (of source code to be compiled). The price of local jobs remains constant, because irrespective of their price, they must be executed locally as they are not eligible for export to other nodes. Thus the total consumption of the processor will be the amount of local jobs, plus the amount of its expressed remote demand which is satisfied by the market.

As described before, the consumer will seek the highest level of satisfaction (ergo, the highest indifference curve) that he can afford to be on. The point of equilibrium occurs at the point of tangency between the budget line, and the highest indifference curve. By inspecting figure 12, this point occurs *on the X-axis* - corresponding to mono-mania in favour of remote goods, as described earlier. This 'mono-mania' is consistent across all processors, so its effect is balanced. The demand curve is synthesised for each processor in a similar fashion, and aggregated by adding together the demand of each processor at each price.

4.5 Market price

The total 'quantity' is predefined, being the total number of source code bytes for compilation. The market price is determined with reference to the aggregate market demand curve.

The quantities demanded by each individual node may be ascertained with reference to their demand curves:

Y = mX + c
so P = mQ + c
so Q = (P - c) / m
where P=market price, and c and m are the parameters of the demand curve.

Each processor having demanded a particular quantity (the sum of the quantities = total quantity available), the available files are then shared out amongst them in a proportionate fashion. As mentioned earlier, quantities demanded will most likely not correspond to integral source file sizes, so the best possible fit must be sought. This is described in the implementation section.

4.6 Summary

This section has developed the model of a load balanced concurrent Make system anticipated by the economic theory of the previous section. The meaning and interpretation of price, quantity, income, demand and supply in the distributed system were presented, and applied to simulate a market-driven load balancing strategy.

The objective was to endow processors in a distributed system with resources and behaviour similar to that enjoyed by actual consumers, and then allow an economic model to allocate resources between them in the manner of a normal market. It would have been possible to have identified a general equilibrium solution to the problem by centrally solving non-linear equations, and then imposing this solution on the marketplace (network), but this complexity would have undermined the simplicity which was a major benefit of this scheme. The procedure was intended to *simulate* the behaviour of a marketplace, and not to seek exact analytical solutions.

5 Future Work

The previous section has shown how micro-economic theory can be mapped onto a distributed system and used to solve a problem in the area of load balancing. In the process a number of issues arose which must be addressed if the model is to be used to solve more general problems in load balancing or distributed resource allocation in general. The most obvious areas requiring future work are as follows.

- The impact of so called "local jobs" needs to be further investigated. There are a number of aspects that need to be considered.

While a processor may be indifferent to whether it processes a local or remote job, the response time to the job will vary as may overall system throughput. There are also cases when jobs should be kept local as the overhead in executing them remotely is not worth the effort.

- In the model just described it was assumed that only one Make was being initiated in the system at any one time. This needs to be generalised to allow multiple Makes not only to overlap in execution but to be actually initiated at the same time.

  This then opens the door to general load balancing.

- The concept of cardinal utility may be worth investigating. The theory of cardinal utility states that every commodity consumed had a definite, measurable utility which was measured in abstract units known as **utils**. The Cardinalists claimed that, with increasing consumption, *marginal* utility declined.

  Thus the first apple eaten on a particular day might afford 10 utils of satisfaction, the second 7 units, the third 5 units, and so on. In theory, marginal utility declines to such a point that it becomes negative - the consumer will eventually be willing to pay to avoid eating more apples.

  While the theory has been rejected as being unsuitable for modeling human consumption it obviously has strong parallels with the way in which computers operate.

6 Related Work

This report concludes with a brief look at some previous applications of economics to distributed systems. These have been mainly in two areas: load balancing and resource allocation. Even within these areas, the quantity of research is minimal, making economic algorithms quite a novel area of application in distributed systems. Three economics-based systems will be briefly described.

6.1 Auctions

[FN88] proposes a system of load balancing based on processors holding auctions for jobs which wish to be processed. Each job entering the system is allocated some amount of money and it 'spends' this money buying itself processing time on local or remote machines, and also buying bandwidth on the communications channels between them, the use of which incurs a monetary cost. The idea of the paper is that when processors become heavily loaded, the cost of buying time on them will increase, and thus work will tend to migrate elsewhere on the net, bringing about load equilibrium. The work was simulated on the NEST test-bed, and revealed promising behaviour. Notwithstanding this, some aspects of the model may be criticised:

- Firstly, and most significantly, the system violates the 'law of one price'. The same commodity - i.e., computer processing time - may be offered at radically different prices on adjacent nodes. This is unrealistic in practice, as identical goods being sold to informed customers should command identical prices. As the system stands, variations in *demand* serve to bring the *price* into equilibrium, rather than vice-versa. Although this anomaly probably does not affect the performance of the system, it certainly does undermine the claimed validity of the economic model.

- Secondly, the overhead involved in administering a bidding system is substantial. Bids must be solicited, received, and then accepted or rejected. Tendering an unsuccessful bid means that a job must bid again elsewhere, and there is an appreciable possibility of 'thrashing' or instability. The system has the additional overhead of maintaining price 'bulletin boards' at neighbouring nodes, which as well as being costly again seems to violate model assumptions: commodities at auctions don't have price tags.

6.2 Spawn

The SPAWN system [Eco89] was developed at Xerox Palo Alto in the mid to late 1980s and while details of the system are sketchy, it used auctioning to control all aspects of resource allocation and it was found that it was necessary to introduce spurious inefficiencies into the market in order to optimise the performance of the distributed system, a clearly undesirable situation.

6.3 Non-auction resource allocation

[KS89] represents a different angle of economic attack. The authors have produced several papers on constrained optimisation and economic methods, with that cited being the culmination of their previous work.

The problem which they consider is the File Allocation Problem (FAP) - how best to strew the file system across the distributed system in order to optimise performance. They distinguish between *price directed* and *resource directed* microeconomic approaches, and adopt the latter, which yields satisfactory and rapidly convergent solutions to the FAP. However, the resource-directed approach concentrates on the concept of *marginal* utility [6] , not on total utility or value. Thus the system does not have a global price structure or market system and may be characterised as more co-operative than competitive. Although being therefore different from the competitive approach proposed in this paper, it nonetheless represents a novel and attractive approach to a difficult problem in distributed systems theory.

7 Summary

This report has outlined a model from microeconomic theory and should how it could be used to solve at least one common problem in load balancing a particular type of parallel job. Related work was sketched to put the model in context and some avenues for future exploration have been identified.

[6] Which considers the *added* value accruing through each decision made in the marketplace.

References

[Cha86] M. Chacholiades. *Microeconomics.* Collier Macmillan, 1986.

[Eco89] The Economist. Auctioning computers. *The Economist*, 1989.

[FN88] Yemini Ferguson and Nikolau. Microeconomic algorithms for load balancing in distributed computer systems. *Proceedings of 10th IEEE conference on distributed systems*, 1988.

[KS89] Kurose and Simha. A microeconomic approach to optimal resource allocation in distributed computer systems. *IEEE transactions on computers*, 38(5):705–716, May 1989.

[Sam92] P. Samuelson. *Microeconomics (14th ed.).* McGraw-Hill, 1992.

[Smi76] Adam Smith. *An enquiry into the nature and causes of the wealth of nations.* Clarendon press, 1976. Originally published in 1784.