GRID COMPUTING: BUSINESS AND POLICY IMPLICATIONS

by

Sze Hwei Ong

B.S.E.; (Electrical Engineering) & B.S.; (Economics)
University of Michigan, Ann Arbor, 2001

Submitted to the Engineering Systems Division
in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Technology and Policy

at the

Massachusetts Institute of Technology

September 2003

Signature of Author: _____
Technology and Policy Program, Engineering Systems Division
August 15, 2003

Certified by: _____
Amar Gupta
Co-Director, Productivity from Information Technology Initiative (PROFIT)
Thesis Supervisor

Accepted by: _____
Dava J. Newman
Associate Professor of Aeronautics and Astronautics and Engineering Systems
Director, Technology & Policy Program

GRID COMPUTING: BUSINESS AND POLICY IMPLICATIONS

by

Sze Hwei Ong

Submitted to the Engineering Systems Division on August 15, 2003,
in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Technology and Policy

ABSTRACT

The Grid is a distributed computing infrastructure that facilitates the exchange of expertise and resources. It is somewhat analogous to the electric power grid in that it can potentially provide a universal source of IT resources that can have a huge impact on human capabilities and on the entire society. Currently the Grid is being deployed (in limited ways) in some research and academic institutions. As Grid computing technologies mature further, the commercial sector can also benefit. With Grid technologies enabling utility computing, enterprises will be able to access IT resources on-demand in a utility-like way.

This thesis gives a brief introduction on Grids and looks back into the history of power grids for lessons learnt. It suggests that the Grid and the power grid are both infrastructures and factors of *reliability, standardization, universal access* and *affordability* are necessary to ensure the success of any infrastructure. Once the Grid is successful, it can open up new opportunities in the field of utility computing and impact IT provision in the commercial sector. The new utility computing ecosystem would consist of five major players – the *Grid resource supplier*, the *Grid infrastructure supplier*, the *utility service provider*, the *re-seller* and the *end user*. Further industry analysis reveals that there are new roles for current players in the traditional IT provision industry and opportunities for new entrants in this new ecosystem. The thesis attempts to identify the characteristics of each of the five major players to help the IT industry better understand the requirements of these new roles. Current players in the IT provision industry would have to decide which of the above roles to play in this new utility computing ecosystem and to re-define their market strategies accordingly. New entrants to the field would likely be players in the telecommunication sector who want a share of this growing pie and whose existing relationship with bandwidth subscribers can be leveraged upon. This thesis concludes with recommendations on several policy issues: Grid standardization for inter-operability, decentralized Grid governance to encourage optimal resource sharing and mechanisms for transcending cultural/organizational barriers inhibiting the commercial adoption of Grid computing.

Thesis Supervisor: Amar Gupta
Title: Co-Director, Productivity from Information Technology Initiative (PROFIT)

ACKNOWLEGMENTS

This thesis writing process has been an extremely challenging albeit rewarding experience. However it would not have been made possible without the help and support of many professors, friends and family.

First and foremost, I would like to express my deepest gratitude to Dr. Amar Gupta for his invaluable advice and guidance. His insights and suggestions based on his broad exposure and experience have enabled me to constantly improve on my thesis. At the same time, his confidence and faith in me to produce a substantial thesis have also enabled me to rise to the challenge even when everything seemed to be impossible to accomplish. Thank you Dr. Gupta.

Next, I would like to extend my appreciation to Professor Joon Park and Professor Lee McKnight from Syracuse University for taking the time to give me feedbacks on my thesis. My thanks also go out to Mr. Dan Kusnetzky of IDC and Mr. Satwik Seshasai of IBM for providing me with valuable research materials.

I would also like to thank Sydney Miller of the Technology & Policy Program for her help and support throughout these two years.

This also goes out to my fellow classmates and friends who have been extremely supportive and understanding throughout this entire process. To Hongfei Tian and Madhav Srimadh, thanks for being my bouncing board for ideas. To Ee-Leen, Kenneth, Shung, Mark, Janice, YinThai, Elana, David, Charlie, Christine and Shaheen, thank you for your friendship and for making my stay at MIT such an enjoyable one.

And I would also to thank a special friend, Raymond Han of Cornell University. Thank you for your constant encouragement and confidence in me, without which, I may not be able to finish this thesis on time.

Last but not least, I would like to thank my family for the love and care that they have showered upon me all these years. My family has been my greatest source of strength and inspiration. I owe all my achievements to them and am really thankful to know that they will support me in anything that I do. Daddy, Mummy, Sze Ling, Wei Ling and Wei Din, thank you for always being there for me.

CONTENTS

LIST OF FIGURES AND TABLES

FIGURES

TABLES

THESIS OVERVIEW

The aim of this thesis is introduce the Grid concept and to examine the business and policy implications that would arise as a result of Grid computing. The business implication is the emergence of utility computing (made possible through Grid technologies) and its impact on current IT provision models. As a result, current players in the IT provision industry would have decide which role to play in this new utility computing ecosystem and re-define their market strategies accordingly. At the same time, new entrants are drawn to this growing pie and how they would fit into this new ecosystem would be examined. Several policy implications are also be explored. The policy challenges are to ensure Grid standardization for inter-operability, decentralized Grid governance to encourage Grid resource sharing and proper consumer education to ensure that cultural/organizational barriers would not inhibit the commercial adoption of Grid computing.

This thesis comprises of seven sections. Section 1 attempts to define the Grid and the motivations behind it. Sections 2 and 3 introduce the readers to the basics of the Grid, such as the resources, the functions and the architecture. Section 4 details the emergence of utility computing and its impact on current IT provision models. Section 5 delves into the business implications by providing an industry analysis of this new utility computing market and the opportunities that would emerge for both current IT provision players and new entrants. Section 6 highlights the policy issues of standardization, governance and cultural / organizational barriers, and recommends ways to improve on them to encourage the global adoption of the Grid. Lastly, the conclusions obtained in this thesis are presented in Section 7.

1 INTRODUCING "THE GRID"

What is commonly referred to as "The Grid" (also known as the Computational Grid) is really a distributed computing infrastructure that is to a certain degree analogous to the electric power grid. Like power Grids, Computational Grids will be able to provide a universal source of computing power and this can potentially have a huge dramatic impact on human capabilities and on the entire society.

This introductory section attempts to define the Grid infrastructure, draws parallels between the Computational Grid and the electric Grid, and provides the motivation for the Grid concept.

1.1 Definition

The concept behind Grid computing is very simple: connect many heterogeneous systems to create a virtual pool of resource, by which any user on each individual system can tap into the vast pool. The array of heterogeneous systems can include personal computers, workstations, computer peripherals, servers, etc, combined to create more computing power, storage space, bandwidth, software applications and hardware functionality than any single system alone. (For more details on Grid components, please refer to Section 2)

In simple terms, the Grid is an infrastructure that supports *large-scale, coordinated resource sharing* among *heterogeneous systems* that span *institutional/geographical boundaries*, in a *dynamic manner*. The resource sharing would be characterized by having direct access to computers, software, data, and other resources on the Grid. Just like the World Wide Web that

enables widespread information sharing, the Grid is also not bound by institutional or geographical boundaries; any user can access any resource made available on the Grid by a provider. The resource sharing is dynamic because consumers can request for Grid resources as needed and resource providers provide them as and when demanded.

However all this is only possible with adequate standardization of Grid protocols. Only common protocols can ensure interoperability, which is critical for multi-institutional sharing of heterogeneous resources. Many Grid technology research groups and forums have since emerged to develop and promote standard Grid protocols, such as the Globus Project (www.globus.org) and the Global Grid Forum (GGF) (www.ggf.org).

As mentioned above, the Grid is an infrastructure and this infrastructure has the potential to bring about a revolution in computing, very similar to the electric power Grid infrastructure that revolutionized access to electricity. The following subsections highlight the nature of infrastructures and how Computational Grids can look to history for lessons learnt.

1.2 Motivations for Grid

The backbone of scientific work today rests heavily on intensive computation, extensive data analysis and more often than not, collaboration between peers. Ten years ago, biologists were content to compute a single molecular structure; now they want to do comprehensive human genome mapping – an extremely computationally intensive task that took two years to complete in year 2000 using the computer system of Celera Genomics with a bandwidth of just under two teraFLOPS (Philipkoski, 2001). After the mapping of the human genome, scientists still need to

unravel the mystery of other biological entities such as genes and proteins, their individual functions and interactions, in order to make breakthroughs in the field of medical sciences. This would not only require faster supercomputers but also lots of information exchange and learning from one another since these are problems that are much bigger than anyone could take on alone.

Technology may progress in line with the three laws of Moore, Metcalfe and Gilder[1]; however this just is not enough to keep up with the demand. As problems become increasingly complex, there is an increasing demand for more/better computational resources by researchers. The lack of collaboration presents obstacles to progress in the field of Research and Development (R&D). It is not that researchers do not want to share resources or cooperate on a project - there is just no good way to go about doing so in an easy, cost-effective and timely fashion. The Grid infrastructure is conceived to facilitate the exchange of expertise and resources. A secure Grid infrastructure would enable experts from geographically distributed locations to come together to work simultaneously on a problem. It would also enable researchers to tap into idle computational resources, expensive software, hardware and applications that they would otherwise not be able to afford. This would bring scientific collaboration to a whole new plane.

So far, computational resources have been used intensively in research work where they could be afforded, mainly at well-funded research and educational institutes and bottom-line driven companies, especially in scientific and engineering fields. However in non-research areas such as city planning, computers are less likely to be made use of. Yet non-research areas can also

[1] Moore's Law: The performance of computers doubles every 18 months; Metcalfe's Law: The value of a network grows as the square of the number of users increases; Gilder's Law: The total network bandwidth doubles every nine months.

benefit greatly from the use of computers. If city planners had made use of computers to select new routes for new roads, a lot of the bad traffic problem could have been alleviated (Foster and Kesselman, 1999). What may be deemed as not being important can have a great impact on the everyday lives of people.

Hence the motivations behind the Grid are two-fold: (i) to provide a platform by which scientific collaborations can take place, through sharing resources and expertise and, (ii) to enable easy access to computational resources thus providing the opportunity for everyone to tap into this substantial increase in computational power, whether they are engineers, scientists, city planners or simply any Grid user.

1.3 Learning from Electrical Power Grids

The Grid, which refers to the Computational Grid, got its name from electric power Grids. The electric power Grid is one of the technological marvels of the 20th century. It is a distribution infrastructure that enables the provision of electricity to billions of devices in a relatively cheap, reliable and efficient fashion. Likewise, the Computational Grid strives to provide a universal source of computing resources – hence the name - "The Grid". From history, we know that successful infrastructures can potentially have a huge dramatic impact on human capabilities and society. In order for an infrastructure as such to be successful, it needs to have the following features – *reliability, standardization, universal access* and *affordable* (modified from Foster and Kesselman, 1999).

Reliability

After the emergence of the power Grid, users expect to receive a reliable supply of electricity whenever they need it. There must be an assurance that the service will not be unnecessarily disrupted. Unless the Computational Grid can ensure the provision of dependable services of non-trivial quality – be it computing power or network bandwidth, the Grid will find it difficult to gain acceptance.

Standardization

It is just amazing how easy it is to plug in an electric device and flick a switch to get electricity to power it up. This is very much taken for granted because standardization of the consumer interface (wall socket), the device electric jacks and the standard operating parameters within which the different Grids operate, make it transparent to the users where their power comes from. Likewise, for interoperability among the heterogeneous resources on the Grid, common Grid protocols need to be developed and deployed.

Universal Access

As long as wires have been laid and a valid account with a utility company is set up, electric power is accessible. Similarly for the Grid to provide pervasive access, it must support access as long as the conditions supporting the Grid environment are in place. Restricted access to the Computational Grid to a select community will not enable the Grid to be successful. A

Computational Grid is successful when access to the Grid is ubiquitous and people use the Grid in every aspect of their lives without stopping to think how to make use of them as though it comes to them naturally. It is just like how one turns on the light switch and expects the lights to come on without thinking how the electricity comes about.

Affordable

Last but not least, access to the Grid has to be affordable. To promote usage, it must be inexpensive relative to income, so that people will be able to afford it. Just like individuals and businesses make use of the power Grid for getting electricity today on daily basis because the cost is reasonable, the Computational Grid has to offer the same economies in order to be broadly deployed.

The above factors highlight the need and characteristics of an infrastructure. These factors come into play because such an infrastructure thrives on network externalities. The paradigm of Grids is very much like the paradigm of the World Wide Web (WWW) in that it thrives on positive network externality. Positive network externality is the notion that each individual user gets more benefit out of consuming a good if there are more users consuming the same good. As more people have access to the WWW, one's ability to access the WWW becomes increasingly valuable since one will have more information resources and can reach out to more people. Similarly, if Grids provide reliable, affordable services universally operating on common standards, more people would make use of them, thus opening up more opportunities for collaboration, more sources of computing, storage and network resources, making the Grid more

ubiquitous. Ubiquity would mean broad Grid deployment, thereby ensuring the success of the Grid infrastructure. Since Grid technologies are still in nascent stages of development, Grid developers, in particular, need to pay special attention to the above factors, in order to steer development of the Grid in the right direction.

2 GRID CONCEPTS AND COMPONENTS

This section delves into the basics of the Grid by introducing the various Grid resources and functions.

2.1 Grid Resources

As mentioned in Section 1.1, the Grid is really an infrastructure that enables any user to access a virtual pool of resources. Grid resource consumers are bound by sharing policies that resource providers set. Grid resource providers will need to define clearly and carefully what they want to share, with whom they want to share and the conditions in which the sharing should take place. The various Grid resources and how they are being utilized on the Grid would be discussed further below (derived from Ferreira et al, 2002).

2.1.1 Computing Power

Computing power refers to the computing cycles provided by the machine processors that are connected to the Grid. There are three methods (with accompanying reasons) to harness idle computing power. The first method is to run an application on the Grid rather than on one's local processor. This usually takes place if one's local processor is not adequate to handle the application. The second method is break down a huge application into small parts which can then be sent to different machines on the Grid to be processed in parallel. This applies to very huge applications that can be broken down, processed separately and then recombined easily. The last method is the parallel execution of an application on different machines on the Grid because the

application needs to be executed many times, e.g., verifying some simulation results that require precision and accuracy.

2.1.2 Storage

The next most common resource on the Grid is data storage. There are two basic kinds of storage: memory attached to the processor or other more permanent storage media such as hard drives, also commonly referred to as "secondary storage". Processor memory may provide fast access but is highly volatile and best serves as temporary storage especially when running applications. On the other hand, secondary storage that provides more permanent storage can help to increase reliability of data and performance. It is the job of the Grid scheduler to select the appropriate storage device to hold the data based on usage patterns or nature of usage. If one needs certain software to run the applications of the data, the data should be stored on storage devices directly connected to the machine with the special software or as close to the machine as possible.

The increased storage capacity on multiple machines is made more useful when used in conjunction with a file unifying system. This enables the data to be stored, whether it is a file or a database, to span several storage devices and machines, and yet provides a single uniform name space for Grid storage. This enables easy data reference, yet the exact location of the file is completely transparent to the user. Special database software can even "federate" various disparate files and databases to create a larger integrated file/database. Since data on the Grid are accessible by a large group of user, data access and update can happen simultaneously and contention can be avoided with advanced synchronization software.

2.1.3 Communications

There are two kinds of communications here – communications both within and external to the Grid. Communications within the Grid are mainly used for sending jobs and the required data to the specified points in the Grid. External communication would refer to the Internet access, in addition to connectivity within the Grid machines. The emphasis here is on bandwidth. If the Grid machines do not share the same communication paths for access both within and external to the Grid, the total available bandwidth for accessing the Grid and the Internet respectively increases. Redundancy has to be built into the communication paths to better handle network failures and to better route excessive network traffic.

2.1.4 Software/Licenses

Some pieces of software are just too expensive for each person to own an individual copy. If one machine on the Grid has the software installed, other Grid users may be able to access the software for free or at a small fee. In cases where licensing fees are very significant, having only one copy of the software installed on one machine cuts down licensing costs and yet does not hamper productivity. There is no need to physically work to the machine with the installed software, as remote software access is possible on the Grid as long as one's machine is logged on the Grid consisting of the machine with the software. Another way in which software companies can control for software access as a result of a limited licensing agreement is to limit the number of installations that can be run simultaneously at any instant despite all machines having installed the software. This is extremely useful especially in a college environment where not every

student can afford the desired software and the college only has a limited budget for purchasing software licenses. It is also applicable for companies wishing to cut down on licensing costs.

2.1.5 Special Equipment

This would pertain more specifically to the research or academic community where the use of special equipment for experimentation is necessary. Like storage devices, the equipment would be connected to the machines that are on the Grid. These equipments tend to be very expensive and only a few institutions can afford them. However with the Grid, researchers in collaboration work that spans institutions can have access to the special equipment, regardless of the exact location of the device. The equipment can be shared free-of-charge or for a fee as determined by the equipment owner. Again, this can help to cut cost (for both the resource provider and user) and enable more effective collaboration among researchers across the institutes.

2.2 Basic Grid Functions

Resource sharing is one of the central themes that runs through the motivation and function of the Grid. Some basic Grid functions are necessary to facilitate this sharing process and manage the Grid resources. The basic resource sharing and management functions that are of most interest are: job scheduling, resource brokerage, resource reservation and resource scavenging. These Grid functions serve as complements to one another and can be implemented together in a single Grid. Below is a brief introduction derived from Ferreira et al, 2002.

2.2.1 Job Scheduling/Resource Brokerage

A job scheduler is the software that is responsible for finding an appropriate machine to run a job that is waiting to be executed. The scheduler can be overridden if the user has a specific location in mind to run the job. However job schedulers are important in all Grid systems especially when the scale of the Grid is extremely large and it is impossible for any one user to be aware of all the available resources that meet the requirements of the job waiting to be run.

A slightly different version of the job scheduler is the resource broker. As the term implies, a broker has the connotation of an agent that facilitates the transaction between two parties. If resources are freely available, then the role of the broker becomes unnecessary. Unfortunately in this world, nothing comes free; hence the need for a resource broker, especially when the transaction is between two independent Grid users with no pre-existing relationship of trust. Part of a resource broker's role is similar to that of the job scheduler - identifying the appropriate resource that a Grid user needs. Another role that the resource broker plays is to enable the transaction between the Grid resource owner and the Grid resource consumer by negotiating a contract in which the Grid resource consumer would pay the producer in cash or in kind (the idea of bartering) for access to the resources.

2.2.2 Resource Reservation

Job schedulers react to the current pool of resources and if none of the resources meets the requirements of the job waiting to be executed, it just keeps on waiting till the appropriate resource becomes available. It may be some time before something is available and if one is pressed for time, waiting is not always the best solution. Hence the idea of resource reservation

comes into play. When the need arises, reservation of resources in advance for a designated set of jobs can be done to guarantee quality of service or to meet deadlines. The reservation works like a hotel reservation system where one's advanced hotel room reservation guarantees one a room even if another guest shows up before the concerned person to vie for the only available room in the entire hotel. The reserved resources will be prioritized and to be given to the user who has made the reservation for it. However this convenience is usually provided for a fee to prevent abuse of the system where users reserve resources for jobs that are of low priority or even when they are not needed.

2.2.3 Resource Scavenging

As the term implies, scavenging refers to the act of sourcing for idle resources or resources that are not being put to productive use. The scavenging is usually for idle processors that can be used to perform useful computations. Resource scavenging is very useful for jobs that are computation-intensive and can be partitioned into smaller sub-jobs. The sub-jobs should be able to run independently without dependence on one another and whenever resources are available. Due to the unreliable stream of resources from scavenging, the scavenged resources are not usually the primary resource used for the jobs. Dedicated resources have to be in place to ensure more predictive behavior.

Scavenging is usually implemented in Grids where resource owners are willing to share their unused computing cycles in periods of low-peak activity. In other words, they are resource donors since there is no monetary exchange or barter of any kind involved. As current development shows, it is usually confined to causes that resource donors are willing to

participate in, such as the recent Smallpox Research Grid Project (www.grid.org/projects/smallpox/), which is an endeavor for smallpox treatment spearheaded by IBM and other vendors. Grid projects that make use of scavengers are also likely to forward a beneficial cause, such as the smallpox research project, in order to garner much public support in the form of computer resource donation. Within a Cluster Grid where the trust is not an issue and the complexity level is low (given small number and homogeneity of machines), scavenging can also be harnessed to enable productive use of idle computational resources.

Implementation-wise, scavenging is usually done in an unobtrusive manner to the resource owner. Generally resource donors download a piece of software onto the computer that they want to use. The software would identify itself as a member of the Grid. Whenever it is idle, its status would be reported to the Grid management node and the node would assign the idle machine to the next available job.

3 UNDESTANDING GRIDS

Section 1.1 has attempted to define the Grid in a logical sense – laying down the minimum conditions that a Grid should be able to achieve. This section gives a brief structural overview of the different Grid implementations. Furthermore, it would elaborate how these different Grid architectures can be configured – either as a Grid for resource sharing or a Grid to be used as an organizational tool. Resource sharing Grids would include Computational Grids, Data Grids and Utility Grids. Grids as organizational tools are Enterprise Grids, Partner Grids and Service Grids.

3.1 Grid Architecture

The focus in this subsection is on distinguishing the different levels of Grid implementations. Scale will be the comparison factor as this subsection explores three different Grid systems, of increasing scale and complexity. The three-tier architecture would include (in order of increasing scale and complexity): Cluster Grid, Intra-Grid and Inter-Grid (also known as the World Wide Grid). Other names that have been coined for this are Cluster Grid, Campus and Enterprise Grid and Global Grid respectively by Wolfgang Gentzsch, Sun Microsystems Director of Grid Computing (Gentzsch, October 2001). The table below details some of the characteristics of the Grid architecture.

Table 1: Characteristics of the 3-Tier Grid Architecture

Architecture	Grid Systems	Owner(s)	Administration
Cluster Grid	Homogeneous	Single	Centralized
Intra-Grid	Heterogeneous	Multiple	Distributed
Inter-Grid	Heterogeneous	Multiple	Distributed

3.1.1 Cluster Grid

The simplest Grid form is a connection of computers over a high-speed local area network. This is referred to as a "Cluster Grid". A cluster would consist mainly of homogeneous end systems that are machines that share the same hardware architecture and the same operating systems. With the homogeneity of the end systems, integration of the machines on the Grid becomes a much easier task. This system would most commonly be found in a single division of an organization where there exists one single administrative control over all the machines. The use of the Grid would not necessarily require special policies or security concerns since there is centralized control monitoring all the processes. Given that the use of the Cluster Grid is primarily contained within the division, only one Grid site is needs to be maintained. Although some people would refer to this as simply a "cluster" (due to the centralized control), the author has chosen to classify this as the simplest Grid form and renamed it as the Cluster Grid.

3.1.2 Intra-Grid

The next level of Grid system is named as such because it refers to the fact that this Grid comprises machines that span the intranet connection of an organization. This is also sometimes referred to as the "Campus Grid" or "Enterprise Grid" since the bounds of the Grid are kept within a college campus or an enterprise. Relative to Cluster Grids, Intra-Grids introduce more complexity as it increases in scale of implementation.

Since the end systems in an Intra-Grid are owned by multiple divisions, it is likely that the types of resources would be more diverse and consist of heterogeneous systems. With different divisional ownership of end systems, separate administration of individual systems would be

likely to emerge, contributing to the heterogeneity. Under separate administration, the emergence of different Grid sites is highly possible (i.e. divisions form their own Cluster Grids within the Intra-Grids). The presence of physical and administrative heterogeneity and the increased number of end systems would lead to difficulty in creating an accurate and updated global knowledge of resources available on the Grid. All these add up to form the various complexities that characterize the implementation of an Intra-Grid. Even with the autonomy given to the respective divisions, it is reasonable to assume that the exertion of a minimal level of centralized administrative control from the parent organization in order to assume the role of a watchdog. On the other extreme, the parent organization can force decisions using a top-down approach, to enforce the uniformity in Grid management across division or collapse the management into a centralized administration that exists in Cluster Grids, reducing Grid complexity.

Unlike Cluster Grids that are mainly set up to support the coordinated use of multiple resources, Intra-Grids provide a channel for data sharing (such as a database) and access to specialized services (such as a special software or equipment on the Grid). With the sharing of resources across divisions becoming possible in Intra-Grids, an interface to facilitate the process has to be created. Some policies and more sophisticated security measures have to be in place to prevent the leakage of private and sensitive data of one division to others. This is not necessary in Cluster Grids, where the Grid complexity is lower. Due to the different functions that Intra-Grids play relative to Cluster Grids, other services that are not usually needed in Cluster Grids may have to be implemented such as resource sharing and resource brokerage services.

Sometimes the Intra-Grid can cross geographical boundaries if an organization has facilities in different cities/countries. Dedicated communications connections (such as Virtual Private Network (VPN)) may be used to connect the Grids in different parts of the organization. Sharing policy and security considerations become even more critical in such cases.

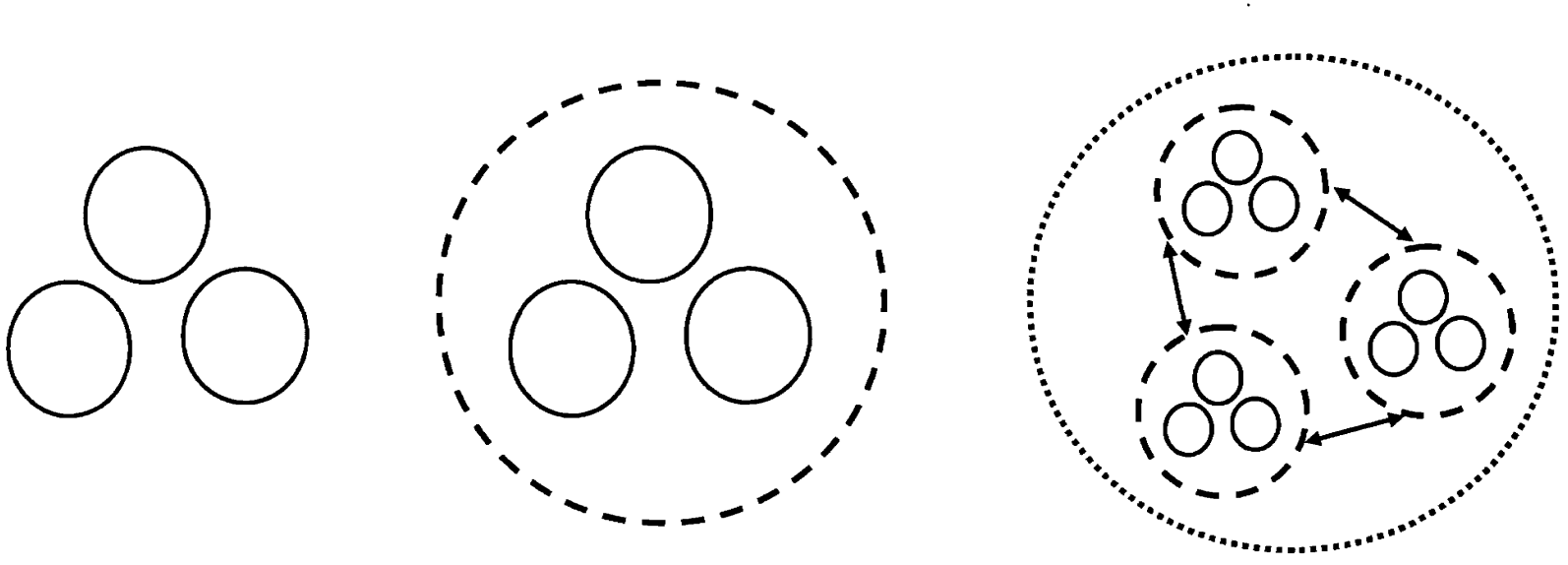
3.1.3 Inter-Grid

This is the most complex Grid system that is being considered because it spans not one but multiple organizations and is possibly implemented on a global scale, connecting inter-networked systems regardless of boundaries. Like Intra-Grids, the end systems on an Inter-Grid are heterogeneous and numerous. In addition to the complexity of Intra-Grids (heterogeneity, separate administrations and lack of global knowledge), Inter-Grids introduce more complicating factors.

Just like the World Wide Web (WWW), which sits on the Internet, it is impossible to exercise centralized control over something that is so diverse. With a different administrative control exerted over individual Cluster Grid or Intra-Grid site, it is impossible for a central authority to guarantee operability or the quality of service. The geographical distribution of the various end systems would also impact network performance. With increasing distance, latency and bandwidth issues can be exacerbated. With geographical diversity, it would not be surprising that Inter-Grids would someday operate on a global level like the WWW, with not just organizations utilizing the Grids but independent individuals tapping into the Grid capability as well. Hence sometimes the Inter-Grid is also known as the "World Wide Grid" or the "Global Grid". International issues such as legislation, intellectual property rights become very different when

they involve international borders. In order for global interaction and transaction to take place, such issues need to be resolved and as history has shown, it is not going to be easy to get many countries to agree on such issues easily. Hopefully, Grid developers and regulators can draw on the experience from the development of the WWW and not repeat the mistakes of the past. Another issue worth mentioning is security. With Intra-Grids, it is possible to assume a certain level of trust between the end systems as they ultimately work for one parent organization. However in Inter-Grids, there are potentially millions of independent Grid resource owners with conflicting interest and there is a definite lack of trust. In order for Inter-Grids to function smoothly, it is almost imperative that there be ways to easily authenticate the user and control the access to one's shared resources. Security measures need to be much tighter and sophisticated than the above two Grid systems. Sharing policies become more important in controlling one's shared resource and need to be clearly articulated.

With Global Grids, one would be able to reach out to a wider audience and tap into a much bigger pool of resources. However in dealing with so many independent entities, current state of Grid development leaves much to be desired. Trust among the end systems is the key to the issue. Cluster Grids can operate on a bare bone Grid without any advanced features because the scale is contained within entities that know and have a pre-existing relationship of trust. As one progress to Intra-Grids, with divisional politics at play and the sheer amount of resources made available, a minimum level of software features needs to exist to facilitate the transactions. As for Global Grids, more sophisticated software services such as resource management, security, authentication, distributed data management, etc, would need to be developed and successfully implemented before the Global Grid can realize its full potential.



Cluster Grid
Local clusters deployed on a departmental/divisional basis.

Intra-Grid
Merging Cluster Grids into an Intra-Grid that is often within a college campus or an enterprise.

Inter-Grid
Merging Intra-Grids into an Inter-Grid that spans across organizations, regardless of location.

Figure 1: A Simplified Depiction of the 3-Tier Grid Architecture

(This figure is modified from Gentzsch, August 2001)

3.2 Grids as Resource Sharing Tools

One of the main drivers of Grid computing is the ability to pool computing resources, enabling them to be used more efficiently. This sharing can enhance collaboration among research institutions, education institutions and even the commercial sector. This subsection elaborates on how the different Grid architectures can be configured specifically for different uses. These various Grids have been given different names to highlight their functionalities and they are summarized in the table below.

Table 2: Grid as Resource Sharing Tools

Grid Type	Main Provides
Computational Grid	Computational Power
Data Grid	Data Access and Storage
Utility Grid	On-demand Access to All Kinds of Grid Resources

This table does not cover all the different resource sharing Grids. In fact, new Grid technologies are being developed as this is being typed and more resources can be shared through specific Grids. What this is intended is to give a summary of the more common resource sharing Grids so that readers can get a better idea of the potentials of Grids.

3.2.1 Computational Grid

Computational Grids are primarily concerned with the sharing of computational resources. These Grids aggregate the computational power of distributed systems and provide secure access to the huge pool of processing power. Applications that make use of Computational Grids usually require intensive computations. The two general applications of Computational Grids are computation intensive computing and high throughput computing.[2] Due to the nature of these applications, some dedicated Grid machines may be necessary to solely handle work on the Grid. These dedicated machines would not be preempted by outside, thereby enhancing the performance of the Computational Grid.

Other than processing computational intensive applications, another reason for setting up the Computational Grid (especially within an organization with various computational resources), is to enable better utilization of these resources. By harnessing the power of these idle resources to process applications, there is little need to buy supercomputers to do the job or even constantly upgrade the systems to better handle the increasingly complex workload.

3.2.2 Data Grid

Unlike Computational Grids, Data Grids are focused on providing access and storage resources for data. Data Grids would primarily provide shared and secure access to distributed, heterogeneous databases and file systems that could contain tera- or even peta-bytes of data. Through a new concept of "*federated databases*", a Data Grid is able to bring together a pool of

[2] High throughput computing aims to increase productive use of processing cycles. (Foster and Kesselman, 1999)

databases and make them appear to the end user as a single virtual database. The federated database would provide a single query point and ensures data consistency. (Ferreira et al, 2002)

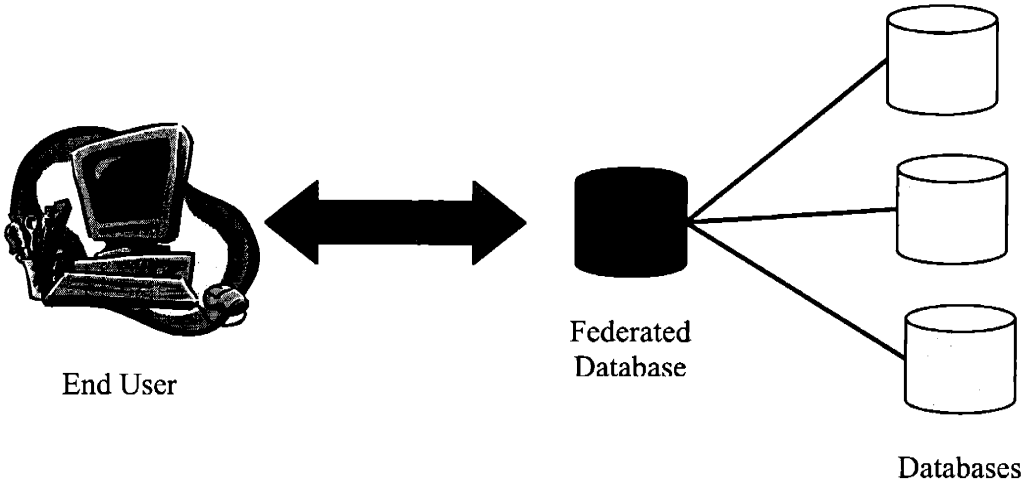


Figure 2: The End User View of a Federated Database

Data security is a big issue as the data involved can be sensitive and confidential. Special measures to ensure authorized access, simultaneous data updates by various parties on the Data Grid have to be implemented. Since data can be presented in various contexts, getting the semantics right would be a challenge to current technology.

3.2.3 Utility Grid

Although the two resources commonly utilized on Grids are computation power and storage, other Grid resources can be equally important as well. Therefore the concept of Utility Grid is born. Utility Grids are Grids that provide ubiquitous access not only to computation power and storage, but also software, special equipment and any other resources that Grid users want to share. Utility Grids may possess the functions of Computational and Data Grids but its primary concern is to ensure on-demand access to all kinds of resources and not limited to either the

computation power or storage. The concept of the Utility Grid is tied to utility computing whereby users can request for resources whenever needed (on-demand) and only be charged for the amount being used. The Utility Grid supports this by providing the infrastructure for utility computing to take place. Companies and even individual users need not purchase expensive storage or software for a specific project but can instead choose to "rent" or share (especially among trusted parties with sharing agreements in place) the required resources for a limited time and this may be cheaper than buying the actual resource. Such Grids are especially useful for on-demand applications or collaborative applications. The concept of utility computing is further discussed in Section 4.

3.3 Grids as Organizational Tools

The above subsection categorizes Grids according to their specific functions. Another view of Grid categorization originated from the IT (Information Technology) industry. This view assumes Grids as organizational tools, with the scale and scope of the Grid as the differentiating factor. The terms "Enterprise Grids", "Partner Grids" and "Service Grids" are born during the hype of utility computing (a concept which is further discussed in Section 4) to showcase the immense potential that Grid technologies hold to enable utility computing in the commercial sector.

The main difference in the three Grid models of utility computing lies in the scope of resource sharing. In the Enterprise Grid, resources are located within an organization whereas for the Partner Grids, it is simply linking up the various Enterprise Grids among business partners (e.g.: between buyers and suppliers) for resource sharing and collaboration among the companies. The

Service Grid (which is still a vision) is only possible when the sharing of computing resources becomes ubiquitous. All computing resources and services required by businesses will be provided by dedicated external service providers (ESPs) and delivered over the Grid platform as a service. For a discussion of the emerging opportunities for providing IT resources through the utility model, please refer to Section 5.

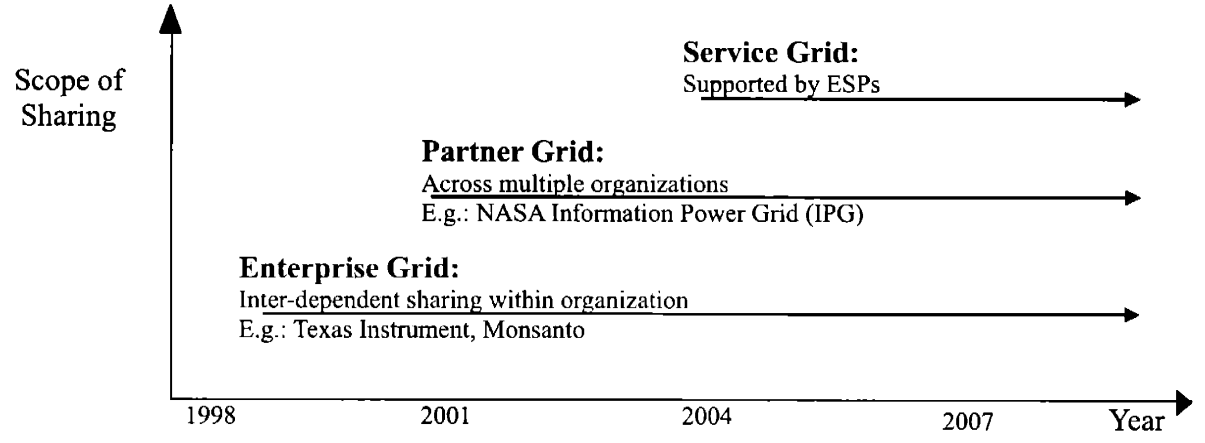


Figure 3: Utility Computing Timeline For the Industry

(This figure is modified from the Platform Computing Paper (Platform, November 2002))

3.3.1 Enterprise Grid

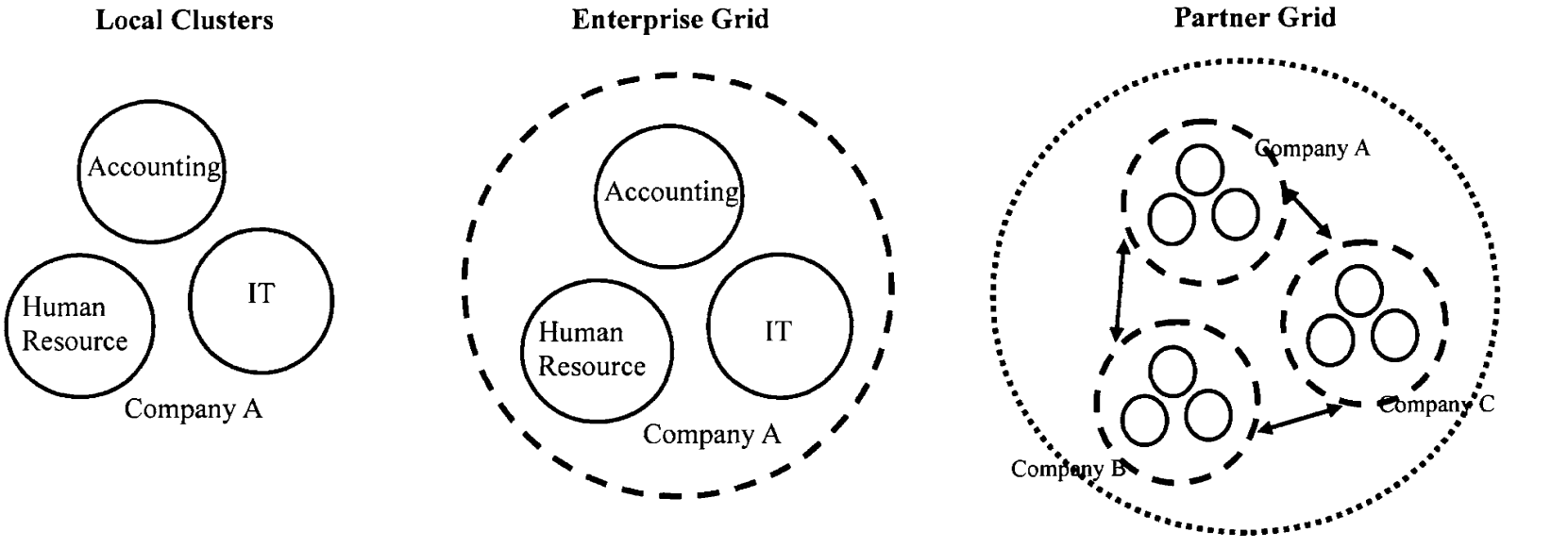
As the name suggests, this Grid is usually within an enterprise. This is the name of a Grid that is built especially for a business organization. Just as most companies have their own Intranet that is protected by firewalls and only available within the confines of the companies, many companies would build their own Enterprise Grids that are only accessible by company staffs. In the education sector, such a Grid that spans the whole institution campus would be termed the "Campus Grid".

The Enterprise Grid takes the form of an Intra-Grid and shares many of its characteristics. Just as an Intra-Grid is comprised of many Cluster Grids, an Enterprise Grid (an Enterprise-Wide Grid to be exact) is the formation of many Departmental Grids (which is really a Cluster Grid). To better illustrate this point, a figure of the Enterprise Grid is shown in Figure 4.

As depicted in the Figure 4, each enterprise has many departments such as Accounting, Human Resources, IT, etc, which would have their Departmental Grids. By linking these Departmental Grids together, an Enterprise Grid is formed.

3.3.2 Partner Grid

A Partner Grid is formed when many Cluster Grids are linked up. Partner Grids are primarily set up between partner organizations, hence the name "Partner Grid". The aim of Partner Grids is to facilitate collaboration by providing access to shared computing resources between partner organizations, such as between sister organizations or between buyers and suppliers. In the field of research whereby research institutions are constantly sharing data and results, a Partner Grid is the ideal medium to enable such collaboration. One such example is the NASA Information Power Grid (IPG) - "a high-performance computation and data grid that integrates geographically distributed computers, databases, and instruments" (www.ipg.nasa.gov). The IPG is essentially a Computational Grid that is made possible through a collaborative effort between several research agencies such as the NASA Ames Research Center (ARC) and the NASA Glenn Research Center (GRC). In fact, Partner Grids are beneficial in any market that involves extensive business collaboration and sharing.



Clusters Within Company A
Local clusters deployed on a departmental/divisional basis within a company.

Company A Enterprise Grid
Merging Cluster Grids into a Partner Grid that is within a company.

Partner Grid between Companies A, B, C
Merging Enterprise Grids of many companies into a Partner Grid that spans across organizations.

Figure 4: A Simplified Depiction of the Enterprise Grid and the Partner Grid

(This figure is modified from Gentzsch, August 2001)

3.3.3 *Service Grid*

The vision of a Service Grid is "a reliable, ubiquitous, and scalable infrastructure for generic service delivery" (Weissman and Lee, pp. 1). The Service Grid follows the Utility Grid model, in that instead of just delivering resources over the Grid, Service Grids deliver resources in the form of services in a ubiquitous manner. Resources such as computational power, storage and software will be delivered in the form of business solutions by dedicated external service providers (ESPs). In the tradition of Utility Grids, Service Grids also follow a pay-as-you-use billing model. The Service Grid would also further the Global Grid (or Inter-Grid) concept in that its outreach and scope would be immense, being able to reach out and connect to many individuals and organizations.

4 EMERGENCE OF UTILITY COMPUTING

Utility computing can happen on a commercial scale or in a shared environment context. In this thesis, the focus would be on the business implication of utility computing in the commercial sector. Firstly, there is a brief introduction of the relationship between Grid computing and utility computing. Next, current IT resources provision models are explained to enable a better understanding of the marketplace that utility computing will impact. Finally, the impact of Grid computing on IT provision is examined.

4.1 Grid Computing & Utility Computing

As Grid computing technologies mature, it becomes more of a reality to be able to provide enterprises with IT resources on-demand as like how one gets water through turning the tap or the electricity through the flick of a switch. In fact, this has been termed "utility computing" in reference to comparison between IT resources and utilities. The term "utility computing" is coined to describe the utility-like way in which enterprises can get its computing needs. Grid computing and utility computing are working towards a similar goal, i.e. to tap into the huge amount of underutilized "dark" computational resources worldwide and make them available in a safe and secure manner, through a commercial or voluntary setting. The difference between the two is that the Grid is really the enabling infrastructure that will help to further the business concept of utility computing.

4.2 The IT Provision Models

The delivery of IT resources to enterprises is known as IT provision. IT provision has changed as technology improves and provides different platforms for resource delivery. On the one extreme, we have the vertically integrated model whereby an enterprise has an internal IT department to deliver the IT resources; on the other hand, some companies are outsourcing their IT resources to focus on their core competencies. With Grid computing in the picture, the face of IT provision is changing once again. The IT utility model enters the picture as the Grid makes it possible for IT resources to be provided and delivered just as how one gets one's electricity and water. The figure below shows how the various IT provision models differ.

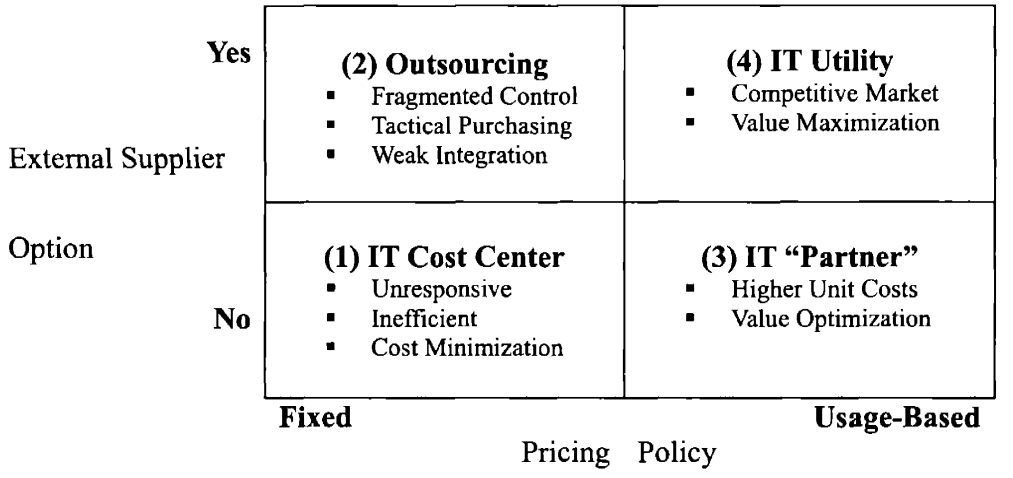


Figure 5: IT Provision Models

(This figure is modified from Gerrard, 2001)

The horizontal axis is a measure of how enterprises are being charged for IT resource usage and the vertical axis is a measure of the competition level in the IT resource supplier market. The various IT provision models are discussed below.

4.2.1 IT Cost Center

The business units in an enterprise often have no choice but to rely on the internal IT department to provide the IT resources they need because there are no other external services providers (ESPs). The IT budget is set aside by the senior management and the business units contribute to this budget according to what the Chief Financial Officer (CFO) presumes to be a fair amount. Since the senior management perceives IT budget as a fixed cost, they are always trying to find ways to control this cost. At the same time, business units who have no control over their IT contribution, often support this strategy of trying to lower the IT budget. Being perceived to be a cost center and "squeezed" by so many parties, the internal IT department is often left with little capacity to invest in its operations, thus making it unresponsive and uncompetitive. This IT provision model is suitable only if IT does not make a strategic contribution to the enterprise.

4.2.2 Outsourcing

IT outsourcing refers to the provision of IT resources through ESPs. Whole outsourcing would refer to the fact that all IT needs within the enterprise are met by ESPs. There is also a hybrid of outsourcing and the internal IT department provision, which is termed selective outsourcing. IT outsourcing became 'popular' in the early e-business investment era whereby business units rush to outsource their IT needs since they felt that the internal IT departments are not adequately meeting their needs. While a lot of money was spent on ESPs, internal IT budget was constrained. This rush of outsourcing made its mark by leaving an IT environment that is fragmented, increasingly complicated and disintegrated, causing enterprises to lose overall control (Gerrard 2001).

4.2.3 IT "Partner"

The IT "Partner" model is very similar to the IT Cost Center model because in both cases, the IT departments are operating within the enterprise. However the big difference is that in the former model, the IT unit operates like a business unit that is concerned with the bottom line and charges business units using economics. Since the business units do not see IT cost as a fixed cost but rather, a variable cost that follows their usage pattern, they will be more inclined to work closely with the IT department (hence the term 'partner') and collaborate when necessary. However given the natural choice of the business units to deal only with the IT-as-a-partner department without competition from other ESPs, there will be less incentive in this case to maximize price performance, leaving to higher internal cost.

4.2.4 IT Utility

The IT utility model is generally defined as a form of outsourcing model as well except that instead of dealing with dedicated IT resources belonging to the enterprise (as is the case with the outsourcing model explained in Section 4.2.2), the IT utility provider owns the common, shared resources that all their customers can leverage on. Simply put, IT cost will be transformed from fixed costs to variable costs as companies get charged for how much they use, instead of paying for the IT resources that they own. A true IT utility model would function like the water or electric utility that we know today. When the IT utility market becomes mature enough, there will be many players and this can become a very competitive market since IT resources will reach near commodity status. The differentiating factor in cases as such would be the trust in the brand name of the providers and the provision of services, such as consulting services, that will best meet the needs of a specific enterprise.

4.3 Impact of Grid Computing on IT Provision

As Grid computing technologies mature, new business opportunities for IT provision will emerge as ESPs take advantage of the Grid platform to deliver IT resources. The hype around Grid computing has created immense interest in IT provisioning through the IT Utility model described in Section 4.2.4.

However, neither enterprises nor ESPs are going to make a switch immediately from the traditional Outsourcing or IT Center model commonly used today to the IT Utility model. The changes are too drastic for any company to handle since the changes would entail a total re-structuring of an internal IT department (which would be met with resistance from IT staff who are used to the old ways of doing things) or a total re-positioning of an ESP's business direction (i.e.: treading on untried risky territory of trying out new business models). As the popular saying goes: "If it isn't broken, don't fix it". Traditional methods of IT provisioning has worked well so many industry observers are adopting that attitude. Besides technology has not reached a point whereby a true IT utility model can exist. Hence various forms of utility computing have since emerged to introduce this new concept via methods that are still within the boundaries of the traditional IT provisioning models.

4.3.1 Different Forms of Utility Computing

According to IDC (Tapper, 2003), utility computing can come in three different ways.

1. **Insourced/In-House Private Utility.** Basically, the private utility model is achieved through building an Enterprise Grid and is managed by the internal IT staff. This self-management retains a little flavor of the "IT Cost Center" and the "IT Partner" model. The utility is dedicated to a single enterprise but the IT assets and infrastructure can be under the ownership of either the enterprise itself or the ESPs. As a first step towards utility computing, it is logical for enterprises to build an enterprise Grid to manage their private utility as a means of achieving efficiency in resource utilization.

2. **Managed Private Utility.** This is similar to the Insourced Private Utility model in that it is still a private utility dedicated to a single enterprise but it differs in that it is being managed by a third party (ESPs) instead of internally. Again the ownership of the IT assets and infrastructure can vary depending on contractual agreements.

3. **Public Utility.** This is a true "IT Utility" model. Different enterprises share a common infrastructure and the ESPs are responsible for the ownership and management of the IT assets and infrastructure.

Table 3: Characteristics of the Various Utility Models

Utility Model	IT Assets & Infrastructure		
	Management	Ownership	Dedicated/Common
Insourced/In-House Private	Internal	Internal/External	Dedicated
Managed Private	External	Internal/External	Dedicated
Public	External	External	Common

With the initial introduction of utility computing, the skepticism of enterprises and the limitations of technology would restrict utility computing to the first two models – the insourced private utility or the managed private utility. In fact, from a recent demand-side survey from IDC, asking 34 enterprises which utility they prefer, the public utility received the most negative response (see Figure 6).

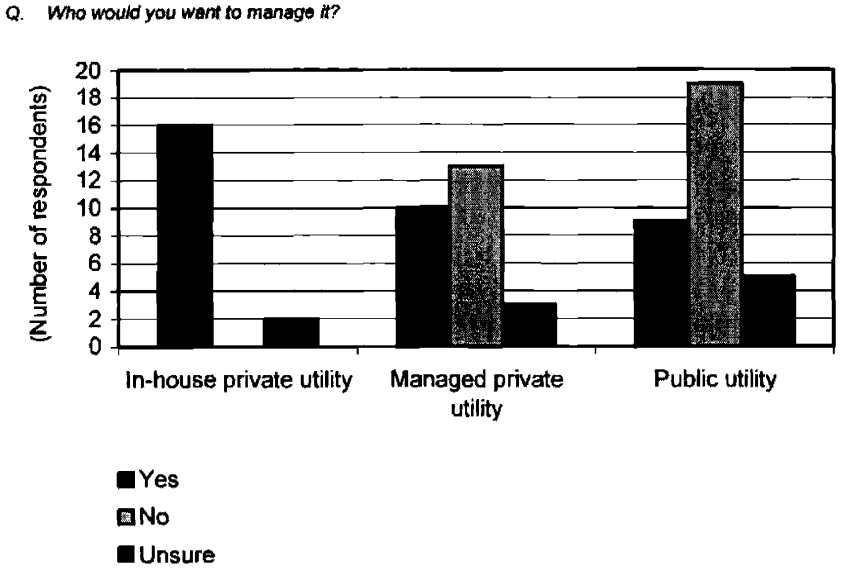


Figure 6: Private vs. Public Computing Utility Solution

(This figure is derived from Tapper and Goepfert, 2003)

4.3.2 Related Utility Computing Concepts

Variants of the utility computing concept have also since emerged. Two other similar concepts of service-centric computing and on-demand computing are discussed below. The fundamental issue with many of these computing approaches is the *management* of infrastructure on behalf of

end users (Villacastin) and the various names have been coined to emphasize what the respective IT providers perceived is the selling point of their services.

On-demand Computing

On-demand computing is built on the utility computing concept but underscores the provision of IT resources on an *as-demanded* basis, whereas utility computing is more focused on the pay-as-you-use paradigm. Both computing concepts are essentially not very different since both terms have been used inter-changeably in the media. Although on-demand computing can be provided utility-style, the marketing hype surrounding on-demand computing is the fact that companies do not have to buy excess IT resources but can purchase them whenever needed. On-demand computing is being championed by IBM who is focusing their resources in this area. This form of computing is particularly appealing to companies who face uncertainly in IT usage (due to cyclical/seasonal demand) and yet do not wish to pump in the high capital investment on IT resources and find themselves stuck with resources that are severely under-utilized. The value proposition of on-demand computing is about the optimum use of resources.

Service-Centric Computing

Service-centric computing takes utility computing idea one-step further. Instead of focusing on the hardware and software to deliver as embodied by the utility computing concept, service-centric computing would concentrate on delivering the required IT services thereby abstracting the complexity out of day-to-day IT maintenance and allowing enterprises more time to focus on

their core competencies. In accordance with the utility concept, the services will be paid for based on usage. In other words, service-centric computing is the modern-day answer to the traditional IT solutions provider who not only provides IT resources but also IT services that help to solve the IT problems that companies faces.

Hence the utility computing provider is to the IT resources provider as the service-centric computing provider is to the IT solutions provider. Another way to look at this is to realize that enterprises will no longer be interested in knowing how their IT problems are solved but rather they expect the solution to yield the result they desired – whether be it faster customer response time or the ability to meet peak consumer demand in a timely fashion without the server crashing.

5 BUSINESS IMPLICATIONS

As Grid technology matures, providing IT services using the utility model becomes easier. Combine that with the underlying trends driving the growth of outsourcing, IT utility models will eventually replace traditional outsourcing models as the primary mode of IT provision. There is definitely a strong interest in IT utility services as shown in the Gartner research brief (Caldwell, December 2001) whereby a survey of 91 end users expressed as strong a interest in IT utility services as traditional outsourcing services. In fact, Gartner forecasts that "the North American IT utility (ITU) market will grow to $8.6 billion in 2003, on its way to more than $25 billion in 2006." (Claunch and McCoy, 2003) Similarly, IDC has also noted a developing trend toward "provisioning IT as a set of utility computing services" (Tapper, 2003).

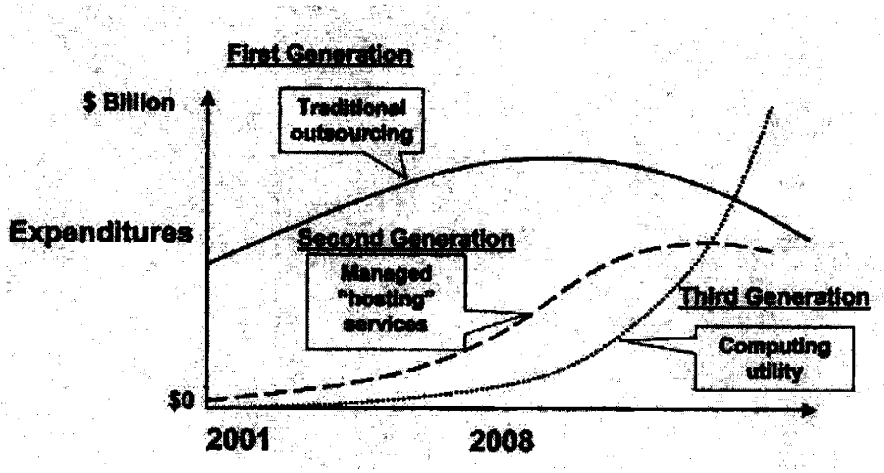


Figure 7: The Trends of the Various IT Provisioning Models

(This figure is derived from Tapper, 2003)

Judging from the figure above, the IT utility industry can provide plenty of opportunities for both existing players in traditional outsourcing and emerging IT utility providers to take advantage of.

If industry players in traditional outsourcing fields do not play catch-up while the industry is still in its nascent stage, they find themselves displaced by emerging IT utility providers. Hence it is critical to segment the utility computing market, understand the roles of the various industry players and find a fit for current or new players.

5.1 Segmenting The Utility Computing Market

Generally, the entire utility computing market can be segmented into the following categories as shown in Figure 8.

1. **Grid Resource Suppliers.** They are the ones who supply Grid resources (as mentioned in Section 2.1) in this whole utility computing ecosystem. They can supply either hardware (to provide computing power, storage and bandwidth) or software (such as software licenses or applications) to the utility service providers.

2. **Grid Infrastructure Suppliers.** As the name implies, they are the ones who make the Grid tick by supplying the appropriate hardware and middleware[3]. Their role in this value chain is to facilitate the connectivity of the Grid resources to the end users. They are also likely to deal directly with the utility service providers, rather than the end users.

3. **Utility Service Providers (USPs).** This would be the new name that IT outsourcers adopt in the new utility computing market. Utility Service Providers (USPs) still perform the

[3] Middleware is the "sets of tools and data [i.e.: the software layer] that help applications use networked resources and services". (Klingenstein, 1999)

role of an outsourcer but their responsibilities in this new market entail more than just providing the appropriate IT resources. They would also be responsible for upholding the terms of the Service Level Agreements (SLAs), monitoring the IT usage and billing the clients appropriately. The USPs and the re-sellers are the only ones in the value chain who would probably have direct contact with the end users.

4. **Re-sellers.** Another variant of the utility service providers, these providers are likely to operate on a smaller scale. The re-sellers would re-package the services provided by the bigger utility service providers and sell them (e.g. adding on enhanced services on top of the bare-bone services provided by the utility service providers). They act as the middleman between the end users and the utility service providers.

5. **End Users.** There are two distinct groups of end users – the business user and the leisure user. The business users would refer to enterprises that require an extensive array of IT resources on a daily basis whereas the leisure user has a sporadic and relatively low demand as compared to their business counterparts. The leisure users would refer to users that use IT resources for personal purposes such as a college student running a simulation or a potential homeowner calculating his/her mortgage rate.

The above is just a brief introduction to the different parties in this ecosystem (the term coined by IDC (Tapper, 2003)). The following subsection attempts do a detailed analysis on each of the five members of the utility computing ecosystem.

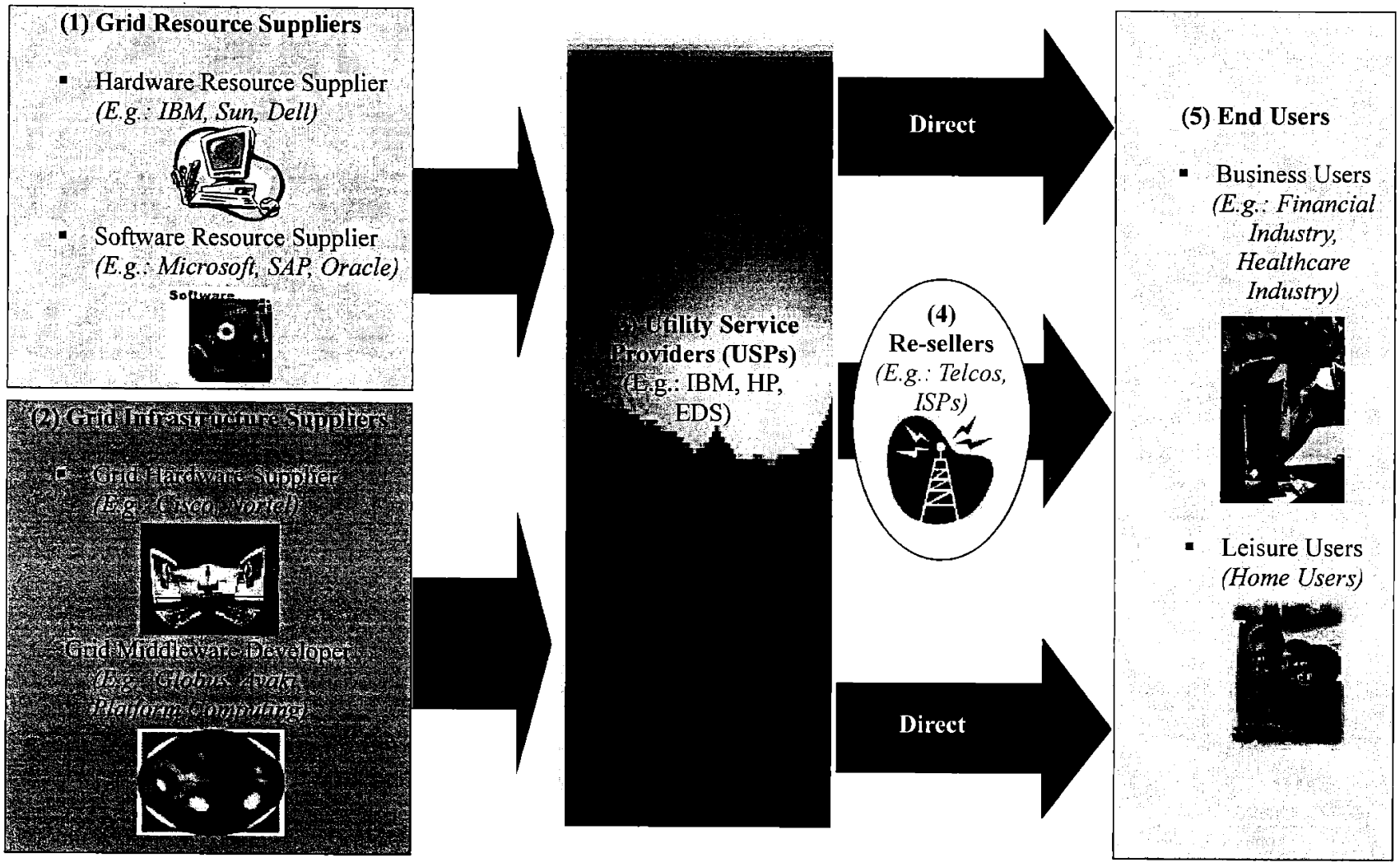


Figure 8: A General View of the Utility Computing Ecosystem

(This figure is modified from Tapper, 2003)

5.2 Players In The Utility Computing Market

5.2.1 Grid Resource Suppliers

Recall that in Section 2.1, the various resources that can be "shared" via the Grid were introduced. The suppliers of these resources are known in this context as the *Grid Resource Suppliers*. The Grid resource suppliers do not offer end-to-end solutions for IT problems that businesses face but rather, they concentrate on developing hardware and software products that can be used by the USPs to deliver IT solutions to solve problems that end users face.

Hardware Resource Suppliers

A hardware resource supplier would supply hardware such as computing power, storage and networking capabilities to the Grid. Traditional hardware resource suppliers would include companies such as IBM, Sun Microsystems, Dell and Hewlett-Packard (HP) that currently cater to all the IT needs of the enterprise from servers to storage and networking products.

Before the advent of utility computing, these traditional hardware suppliers provide quality IT products and services which subsequently enabled them to build a reputation for themselves, establishing a brand name that is synonymous with good hardware provision. Since enterprises usually go with the same company for most of their needs to avoid the complexity of integrating incompatible resources and non-interoperability, these suppliers are able to lock their clients in. However with the hardware delivery mode undergoing a major change due to utility computing, traditional hardware providers are slowly losing the factors that enable them to be successful. Firstly, they are losing the "stickiness" factor for clients to stick to their services since the clients

are likely to interact solely with the USP instead. With the USP handling the service end for the clients, traditional hardware providers become easily replaceable. Without the value-added services, the hardware they produce (even if they were of high quality) becomes a commodity that can be readily copied by new upstarts, especially by the low cost hardware manufacturers emerging in China and India.

However traditional hardware resource suppliers can continue to grow in the Grid resource supplier role as long as they invest in Research and Development (R&D) to constantly improve on their resources to stay ahead of their low cost competitors. Couple that with their established reputations which is not something that a new low-cost entrant can easily imitate and replicate, traditional hardware supplier should be able to stay ahead of the game as a Grid hardware supplier.

Other companies such as the Storage Service Providers (SSPs) and networking companies (e.g.: Cisco) can also play a role in providing storage and providing bandwidth respectively when the Grid develops as a delivery platform for hardware. The failure of the SSPs to realize the full potential of on-demand storage will become a thing of the past with the development of the Grid. (For a short history of SSP evolution, please refer to Couture, 2003.) Companies that have failed in the past would be able to re-enter the industry since utility computing would open up a whole new market demand for storage utility. There will also be a new demand for bandwidth, as utility computing will increase bandwidth consumption. According to Insight's Research's study (2003), "Grid Computing: A Vertical Market Perspective 2003-2008", "the [Grid] technology will likely form the foundation of a fourth wave in IT, and present new opportunities for the

long-suffering telecom industry" (UtilityComputing.com, May 2003). Networking companies and telecommunications companies with plenty of "dark" fiber can play the role of hardware resource suppliers by providing bandwidth on demand.

Software Resource Suppliers

Currently software is often provided on disks, complete with licenses for usage. Software licenses can be very costly and this has deterred enterprises and individuals from purchasing them. All that is set to change when Grid computing enters the picture. With the Grid as a medium of software delivery, software resources can be provided as a service on a pay-as-you-use basis. That was the promise of the ASP (Application Service Provider) model that never quite took off with enterprises due to the limitations of the technology then. In fact, Marc Benioff, the chief executive of startup salesforce.com is the "cheerleader for the idea of delivering software as a subscription-based utility, like water or electricity – an idea that many firms have tried, mostly without success" (The Economist, June 5, 2002). With Grid technologies in place, the software-as-service model is beginning to take shape.

Software-as-service model will appeal to enterprises that are experiencing chaos in the computer room after years of purchasing the latest and greatest software. The difficult to use or flat out buggy software costs the U.S. economy $59.5 billion annually (Kerstetter, 2003). As salesforce.com is making its foray into this new domain, other software developers and independent software vendors (ISVs) cannot afford to ignore this trend that can potentially change the way they do business. Companies that are likely to be affected by this change are software developers such as Microsoft and the ISVs such as Oracle, SAP and PeopleSoft. In fact

this is deemed a serious enough development to warrant the software maker community to sit down together to discuss issues such as agreeing on the standards to adopt by focusing on a single Web-programming language that allows their software to tie together more easily (Kerstetter, 2003).

When software can be easily delivered over the Grid platform, commoditization starts to set in very quickly. With the pay-as-you-use model and the easy delivery of software, enterprises are no longer tied to their software providers and easily switch to another provider due to cost or quality reasons. Similar to hardware resource suppliers, software resource suppliers have to innovate fast enough to stay at the forefront of creativity and technology, or they may find themselves losing market share and subsequently overtaken by low cost upstarts or forced out of business by the more established companies.

5.2.2 Grid Infrastructure Suppliers

As the name implies, the Grid Infrastructure Suppliers are the providers of hardware and middleware that makes the Grid infrastructure tick. This group in the ecosystem consists mainly of Grid hardware suppliers and Grid middleware developers.

Grid Hardware Suppliers

Grid hardware suppliers are the providers of hardware for the Grid. Since the Grid platform rests almost entirely on the network, traditional hardware networking companies such as Nortel and Cisco can still play this role. As bandwidth becomes an important issue for Grid access and

managing Grid traffic, the telecommunication companies (Telcos) and the Internet Service Providers (ISPs) may have a role to play here.

Grid Middleware Developers

With today's Grid technologies still at the nascent stage of development, Grid middleware developers face a daunting task of setting standards for Grid technologies. The forerunner of the Grid middleware developer role is The Globus Project (www.globus.org). Other active middleware developers are Avaki and Platform Computing. The role of Grid middleware developers currently centers on helping end users build Intra-Grids.

5.2.3 Utility Service Providers (USPs)

The Utility Service Providers (USPs) are companies that provide the connectivity to link the end users to the resources supplied by the hardware/software resource suppliers (also known as the Grid resource suppliers). To be more explicit, they deliver the solutions to IT problems that end users face, using the hardware/software that the Grid resource suppliers provide. The end users should only see the delivered results and the interactions between the Grid resource suppliers and the USPs should remain transparent. USPs generally need not possess any hardware/software ownership. However there are currently some hybrid USPs that do posses their own data center. To differentiate between the two kinds of USPs, the former shall be known as a *pure-play USP* and the latter, a *hybrid USP*.

Pure-Play Utility Service Providers

Companies playing the pure-play service provider role in traditional outsourcing are the systems integrators, IT consulting firms and IT solutions providers. These companies generally do not own any of the hardware/software supply but they have the expertise to come up with solutions to specific problems that end users have, by using the resources provided by suppliers.

As Grid technologies mature, these companies should incorporate the use of the Grids in their solutions in order to be able achieve true IT utility and evolve to become pure-play USPs. Other than providing the connection services, pure-play USPs must come up with appropriate metering services and pricing methods in this new business model to cater to the needs of end users.

One of the first few known pure-play USP in the market is Electronic Data Systems Corporation (EDS). EDS is an outsourcing services company and by nature of its pure-play services model, has appeared to assume the role of a pure-play USP (Tapper, 2003). This is an area that systems integrators, IT consulting firms and IT solutions providers should be looking into in order to re-define their market strategies appropriately to keep up with the pace of technology changes.

Hybrid Utility Service Providers

The first few companies to jump onto the utility service provider bandwagon are IBM and HP. Both of these companies started out as hardware resource supplier companies. Having the foresight to envision IT resource commoditization, they ventured into service provision to raise the profit margin. In order to enter the service provider market, IBM created the IBM Global

Services arms and HP developed the Data Utility Center to strengthen their IT expertise in the service-provider area. These companies possess both hardware/software ownership and provide the end-to-end solutions that end users require. In order to be both a resource supplier and utility provider, the company needs to own excessive amounts of hardware/software that can be a huge capital drain on the finances and small pure-play service providers may not be able to afford the upfront investment.

It will be more difficult for pure-play USPs to become hybrid USPs because of the huge capital investment involved in acquiring the hardware/software. Besides if the pure-play USP is late in the resource/infrastructure supplier industry, it may have to play constant catch-up and not yield anything profitable in the long run.

On the other hand, it is extremely attractive for pure hardware/software providers to assume a hybrid USP role. As mentioned in Section 5.2.1, Grid resource suppliers face the threat that they can no longer lock-in their clients and face increasing commoditization of their products. Hence to look into expanding their hold on the market, they may follow in the footsteps of IBM and HP.

A major issue that may arise as a result of hybrid USP is that clients may suspect the objectivity of the hybrid USP's decision in using their best-in-breed technologies. Hybird USPs may also be internally pressured to use their own resources to which may not always be the most suitable or cost-effective for the client's needs. A way around the issue is "to establish a separate business and operating unit, unencumbered by the technology division, with the technology division competing for business, alongside other similar competitors" as suggested by Tapper from IDC

(Tapper, 2003). However he further argues that this hybrid model may not be as effective as a pure-play USP approach.

The fact that the strongest USP players in the utility computing ecosystem are IBM and HP is not without its reason. The author believes that Grid technologies are still in the transitional phases to full maturity and IBM and HP may be more poised to deliver managed or private utility computing solutions that are more readily accepted by the enterprises today.

5.2.4 Re-sellers

The re-seller channel usually refers to smaller pure service-centric provider. They usually do not posses the clout of big service providers but probably cater to a niche market that is exclusive to them. The re-sellers are like the CLECs (Competitive Local Exchange Carriers) who have to "purchase" the rights to access the networks of ILECs (Incumbent Local Exchange Carriers).

One such niche market is the leisure user market. The Telcos and the Internet Service Providers (ISPs) are in many ways more appropriate to meet the needs of the leisure users of IT services. By nature of their relation to the leisure user (the interaction via the monthly Internet bill), the Telcos and ISPs are ideal to be the service provider of utility computing to the leisure users. Having a huge subscription base of users, Telcos and ISPs can easily use Grid technologies to aggregate the idle computing resources of their subscribers and provide a service for them by creating a business environment whereby they can have access to computing resources in an easy, safe and secure manner. Subscribers would benefit from the convenience of having an aggregated Internet/Grid bill and not having to go through the hassles of subscribing to a new

service with a new provider. They also need not worry about not being a lucrative enough market for utility providers of business users to cater to them and having to deal with uncertified third-party utility service providers who may otherwise be unqualified to provide utility services.

Other enterprising new entrants can also capture the leisure user market by setting up the necessary billing and monitoring infrastructure and ride on the networks of the Telcos and ISPs to reach the users. That way, Telcos/ISPs do not have to invest unnecessary resources into what is not considered their core competencies and "outsource" that to the dedicated third parties to handle. The users still go through the Telcos/ISPs and need not worry about the unnecessary hassle of dealing with another set of bills. If the new entrants can develop through positive network effects (e.g.: eBay) and gain enough clout to establish this on their own, a new business model would emerge.

Another area that a re-seller may be necessary is when a generic USP cannot cater to the specific needs of the industry. Some industry needs can be very complex and require specialization in a specific industry area to better service them. Such a niche market may be too specialized for a generic USP and hence the specialized re-seller can have a role to play. These specialized re-sellers can currently be IT consulting firms for specific industries. The industry would place more trust in their expertise than they would a generic USP.

5.2.5 End Users

All end users have differing preferences. In this context, end user preferences can clearly segment the user population into two distinct groups. The first group of end users is has simple

needs and takes budget into serious consideration whereas the second group has other expectations of the purchase other than cost (e.g.: quality of service, speed of service, etc). End users in the former group is likely a leisure user who shops around for the best prices and is willing to tolerate a little delay or service dissatisfaction; the latter group is likely to consist of business users who are willing to pay more for timely and quality service from the Grid.

Price Sensitivity

Given that a business end user may be concerned with the kind of resources he/she is getting and whether they meet his/her expectations, he/she is likely to be paying more to ensure that his/her expectations are met. On the other hand, the average leisure end user is usually willing to tolerate some service delays to get a better price. He/She would be more price sensitive and shop around to get the best deal.

Demand Preferences

Since it is expected that a business end user has bigger and more complicated problems for Grid resources to process, his/her demand for Grid resources is likely to be larger and more complex. The business end user may require bundles of different Grid resources and other differentiated services such as a service guarantee to ensure that the processing can be finished on time. An example would be the demand for special software to simulate huge data sets and requiring storage for the results with the expectation that this can be finished in less than two hours. A leisure end user is likely to be dealing with much simpler work that would only need one or two

kind of Grid resources. A typical straightforward demand would be to request computer cycles to process a small-scale simulation for a school project.

Differences \ Type of End Users	Business	Leisure
Price Sensitivity	Low	High
Demand Preferences	Differential and complex array of services required	Simple and straightforward

Table 4: A Comparison Between Business End Users and Leisure End Users

Implications

It can be inferred from the above that there is a need for service discrimination. Service providers and re-sellers should make available a wide range of Grid resources and complementary services, so that users with differing preferences can all be satisfied. Even though Grid resources are near commodity items, the differing end user preferences can clearly segment the target market for service providers and enable them to price differentiate.

5.3 Wireless Grids & Wireless Utility Computing

Wireless technology has found its way into the lives of people from all walks of life. Whether it is for work, for school or for leisure, wireless has become ubiquitous as people communicate using their cellular phones, wirelessly check their emails on their PDAs (Personal Digital Assistants) or surf the web on their laptops in their backyards. Businessmen, students and

housewives alike, are using wireless technology to a large extent. Wireless technology is certainly becoming more indispensable as people come to rely on it for their everyday needs.

With the ubiquity and indispensability of wireless technology, it will not be long before wireless makes its way into Grids. Firstly, there are millions of powerful mobile end devices sold in the market annually. Technology has progressed such that powerful microprocessors are made small enough to be embedded into mobile handheld devices. There are currently many PDAs and super lightweight laptops that are using 802.11b wireless technology (or Wi-Fi). It is a natural transition for the Grid to tap into this huge potential market of resources and optimize their usage. Secondly, a wireless Grid does offer the many conveniences over its wired cousin. With businesses talking about productivity and connectivity, the wireless Grids would enable the business person on the go, to maximize his/her time on a flight or at the airport or even on his/her drive to the office. It also offers a way to connect the business person outside the office compound, a secure way to access the company's database and get the most updated company information to the client. Couple all this with the fact that developments in wireless technologies such as 802.11, GPRS and 3G has grown significantly in spite of the downturn in telecommunications investments, they are becoming more affordable, reliable, easily accessible and ubiquitous to the man on the street.

Of course, there are challenges associated with the wireless Grid implementation that are unique only to it. Firstly, the computing power of the mobile end devices are very much reduced compared to what one is getting from a workstation on a wired Grid. Then there is the small secondary storage issue that may not even prove to be very useful. Thirdly, there is the battery

consumption issue that is always a concern with mobile end devices. As it is, device users are complaining of the short battery lives of mobile devices. By utilizing them on the wireless Grid, the battery consumption sensitivity is heightened. Lastly, there is the low bandwidth, unreliable communication problem. Wireless connection tends to be spotty and this can be a serious disruption to processes on the Grid. Despite all these inherent limitations of mobile end devices, Phan et al (2002) has argued that the limitations can be overcome and hence the transition from a wired to a wireless Grid is not at all inconceivable.

With wireless Grids, utility computing would be able to take on a wireless dimension. As Section 4.1 has already defined 'utility computing' as the utility-like way in which enterprises can get its computing needs, wireless utility computing would refer to the utility-like way in which enterprises can get its computing needs wirelessly.

5.4 Further Business Implications

The utility computing ecosystem depicted in Figure 8 is only intended for the wired Grid. However it is not completely invalid for the wireless Grid. As with the mobile Internet, the backbone of the wireless Grid still rests very much on the wired Grid.

Looking at the mobile Internet model of today, mobile subscribers usually turn to their Wireless Service Providers (WSPs) for access to the mobile Internet. However the WSPs only provide the access but not the content. Similarly, the wireless service providers (WSPs) provide the points of contact with end users in the wireless utility computing ecosystem. By drawing the parallel here, it is natural for WSPs to assume a utility service provider role. Being a service provider is not the

core competency of a WSP, hence they are likely to act as a re-seller. Just like the mobile Internet, WSPs by nature of their relationship with their subscribers, is ideal to be a re-seller. The WSPs have an existing relationship with their subscribers that is built on trust and there is the convenience of receiving one bill and not divulging more private information to another party.

Regardless of the end user (be it the business user or the leisure user), the WSPs can cater to both groups. WSPs are already differentiating between the business users and the leisure users through differentiated pricing plans. Similarly, the level of utility computing service required can be differentiated through pricing.

6 POLICY ISSUES

6.1 Standardization

In many ways, the Grid is trading on a path that is very similar to the Internet. If the Grid lives up to its hype, it will revolutionize computing in the 21st Century the way the Internet did in the 1990s. Indeed there are physical similarities that warrant this comparison. As Ian Foster (senior scientist of the Argonne National Laboratory and Head of the Globus Project) pointed out, the Grid is "a set of additional protocols and services" that sits on existing Internet protocols to support "the creation and use of computation- and data-enriched environments" (Foster et al, 2001). Hence it is appropriate to look back into the evolution of the Internet and learn some lessons from history. Learning from the history of the Internet, one can infer that standards did indeed play an important role in making the Internet what it is today. However standards alone are not sufficient to achieve the feat that the Internet did. The standards also have to be *open*. The following subsections discuss the importance of open Grid standards to the development of the Grid.

6.1.1 Why Are Standards Important?

In Section 1.3, it is mentioned that a successful infrastructure has the following four features - *reliability, standardization, universal access* and *affordability*. Indeed, standards do play a very important role in the success of any industry. Standards form the basis for product design and comparison and most importantly in this case, facilitating interoperability.

Returning to the power Grid example, standards govern the design for the wall sockets such that any electrical device can be plugged in and powered up. The reliability of the electrical power source lies in the standard operating parameters that dictate how different Grids operate. These standard operating parameters enable power sources to be routed to where it is needed (from one Grid to another if necessary) such that even when a power plant fails, it would not severely affect the neighborhood that are primarily powered by that plant.

Similarly, Grid standards have to be established in order that Grids can interoperate. This is especially true because with Grids, we are talking about heterogeneous resources from multiple domains across organizational borders and it will be impossible to expect everyone to adhere to the same products. Without standard protocols, different proprietary vendor solutions would not be able to communicate with one other, resulting in disjoint islands of Grids. When different Grids adopt their own standards and they cannot interoperate, the overall value of the Grid as an infrastructure to bring about collaboration through coordinated resource sharing would be reduced. This would severely delay the realization of the Grid's full potential. Furthermore, the reliability of the Grid is likely to be compromised by the lack of interoperability. With disjoint islands of Grids, there will be fewer resources to be tapped into and this would result in difficulty in meeting demands for resources.

Therefore, standards are indeed important in that they play an important role in facilitating interoperability and reliability that are critical for the Grid. Without standards, Grids would not work based on proprietary and incompatible protocols, interfaces and application environments. In fact, the importance of standards has not gone undetected by Grid developers. Many Grid

technology research groups and forums have since emerged to develop and promote standard Grid protocols, such as the Globus Project (Globus) and the Global Grid Forum (GGF). Both Globus and GGF are non-profit groups that aim to develop open Grid standards that will benefit the Grid community as a whole. The concept of free standards is further discussed in the following subsections and the function of these groups is detailed in Section 6.2.

6.1.2 Open Standards

Grid standards are important for interoperability reasons. However it is even more important that these standards are open. Standards that are open would imply that they are freely available to anyone who wanted to use them and there is no need to get anybody's permission to utilize them. By making standards open and free for all to use has the following consequences.

Firstly, Grid developers would be more willing to use open standards to develop Grids since they do not cost a cent and there is no possibility of being locked-in by proprietary solutions that would cost even more to maintain in the future and not to mention, the additional cost incurred in trying to enable the Grids to interoperate. (Which is why Web services are so hyped because they solve precisely this interoperability issue that has hindered inter-enterprise cooperation for decades.) Open standards would further the affordability feature of a successful infrastructure and encourage widespread adoption, which leads to the second point.

Looking back at the development of the Internet, TCP/IP was developed by the U.S. government without the formation of an internationally recognized standards authority. However the government's decision to make the protocol a free standard "led to much more universal

adoption of the common Internet protocols than any of the other more centrally mandated alternatives" (Gillett and Kapor, 1997, pp. 9). Taking a leaf out of history, the Grid community has realized the importance of free standards and formed non-profit research and forum groups such as Globus and GGF to standardize Grid APIs (Application Program Interface)[4] and protocols, in order to encourage their universal adoption – an important feature of a successful infrastructure.

As ironic as it seems, since the standards developed would be free for all, more interested parties would be willing to contribute toward a collective goal in which they perceive their participation to further their individual interest. This is most obviously seen in the growing Linux movement in the 1990s. Countless developers, programmers and individuals all over the world put in their time and expertise voluntarily to assist Linus Torvalds in the creation of Linux – the free Unix-type operating system. On a similar note, Globus and GGF are attracting expertise from all domains (academic, government and corporate – even competitors) to partake in the development efforts of free Grid standards. Besides, such forums and working groups would be the perfect place to ensure that one's competitors do not advance their self-interest at the expense of one's own. Free of financial interest, the free standards that would emerge would be a quality product that everyone would be willing to adopt.

Lastly, the generality and the affordability of the free Grid standards also encourage the development of appropriate applications. The Grid developers are limited by their imagination as

[4] An API defines a standard interface (e.g.: a set of subroutine calls or objects and method of invocations in the case of an object-oriented API) for invoking a specified set of functionality. (Foster et al., 2001)

to what people would use the Grid for. Even in the Internet era, developers had not anticipated the use of electronic mail and were astonished at the popularity of such an application. Letting application developers have free, easy access to Grid standards would allow them to exercise their creativity and imagination which would ultimately benefit end users who get access to useful applications.

6.1.3 Standards Organizations

After dwelling on the importance of free standards, it would be appropriate to examine the role of standards organizations. As mentioned above, the initial developments of the Internet came about without a standards organization. However as the commercial interest in the Internet started growing, it became necessary to establish neutral organizations (such as the IETF – Internet Engineering Task Force[5]) to ensure an open and fair standards process. With 20/20 hindsight, Grid developers have started GGF and Globus before Grid developments explode beyond control.

Global Grid Forum (GGF)

GGF is a community-initiated forum, formed in November 2000, through the merger of Grid Forum, (a primarily U.S.-based grassroots organization of individuals developing, deploying, or using Grid technologies), eGrid, the European Grid Forum, and Grid leaders from Asia-Pacific. Casting aside territorial and cultural boundaries, the GGF was formed to prevent duplication of

[5] The Internet Engineering Task Force (IETF) is a large open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet.

work and the ability to realize a vision of interoperability across Grids. Recognizing a convergence in their efforts, GGF subsequently merged with the New Productivity Initiative (NPi) and the Peer-to-Peer Working Group (P2PWG) in year 2002 and formed an alliance with the Distributed Management Task Force, Inc. (DMTF) in year 2003. Both NPi and DMTF are industry organizations. Currently, there are more than 5000 "individual researchers and practitioners working on distributed computing, or "Grid" technologies. GGF's primary objective is to *promote and support the development, deployment, and implementation of Grid technologies and applications via the creation and documentation of "best practices" - technical specifications, user experiences, and implementation guidelines."* (www.ggf.org) GGF receives financial support through membership dues from GGF sponsor members.

The Globus Project

The Globus Project "conducts research and development to create fundamental technologies behind the "Grid," produc[ing] open-source software that is central to science and engineering activities and is the substrate for significant Grid products offered by leading IT companies". (www.globus.org) Globus is primarily funded by government agencies and a few corporations (namely IBM, Microsoft and Cisco).

Potential Issues & Problems

While it is great that we are learning from the Internet experience and setting up standards organizations to oversee the entire standards process, there are some potential issues and problems that the author foresees.

1. **Lack of end user representation**. The membership profiles of GGF consist of many industry vendors/service providers and whereas those of Globus are confined to the government/research agencies and selected corporate sponsors. The people who are potential commercial users of the Grid technology are starkly missing. Of the over 700 participants in the Seventh Global Grid Forum (GGF7) hosted by GGF in Tokyo (March 2003), less than five percent were commercial end users as observed by MacKinnon (2003). While vendors/service providers have a vested interest to be part of the loop (to influence standards in their favor) and government/research institutes are represented by nature of the Grids' roots in research, it is also critical to get corporate end users excited and involved. After all, they are the ones who would be able to bring about the growth of the Grid through its widespread deployment. If end users are not aboard on this in the early stages, many of the standards and technology would be oriented towards the interest of the vendors, service providers, government and research institutes. Hence to have a direct voice in shaping these standards, corporate end users have to get directly involved quickly.

2. **The role of the government**. Proponents of the market would argue that there is no need for the government to be involved. If there is a demand, the market forces would see to it that such a technology is supplied. However in this case, market forces would drive each individual vendor to market their respective proprietary technology such that it renders the whole Grid useless due to lack of interoperability. Since Grid technologies are still in the nascent stages, the initial government funding to jump-start a revolutionary

technology would be necessary. Had the Defense Advanced Research Projects Agency (DARPA) or the National Science Foundation (NSF) not sponsored Globus, it would have been extremely difficult to gather a group of people with conflicting interests to sustain support in this project. Private funding can also be difficult to obtain as research efforts in areas of standards are deemed to be mundane and non-innovative (i.e. not "sexy"). (Schopf and Nitzberg, 2000) With the government in the picture, corporate sponsors would also be more willing to participate with the assurance that the interests of their competitors would not be unfairly advanced. However, there comes a time when the government has to wipe their hand off the project. When the Grid can sustain the momentum that the Internet generated in the commercial sector in the 1980s, it will be time for market forces to do their work.

6.2 Governance

If the global Grid becomes a reality, how should it be governed? With the Grid riding on Internet protocols, it is assumed that the Grid would be governed in the same way as the Internet.

6.2.1 Internet Governance

According to Johnson and Post (1997), there are four basic competing models for the governance of the global Internet. The global Internet can be governed by:

1. Existing territorial sovereigns,
2. Multilateral international agreements by sovereigns,
3. New international organizations, or

4. De facto rules that may emerge as a result of the decisions undertaken by the individual parties (users, system operators, etc) on the Internet.

He goes on to highlight the problems of the first three methods and explain how the Internet is capable of being governed by the fourth method through the decisions and conditions that each member of the Internet community pose. On a slightly different note, Gillett and Kapor (1997) pointed out how the Internet can be self-regulating through the design of the protocols. Lessig (1999) ties both ideas together by demonstrating that code (which are the protocols) are designed to incorporate the desires and wishes of the Internet community. All these suggestions of Internet governance point to a decentralized, self-regulating design.

6.2.2 Grid Governance

With the similarity of the Grid and the Internet as an infrastructure to deliver computation resources and information respectively, the Grid should ideally be self-governing and managed in a decentralized fashion for reasons that have been applied to the Internet. The reasons for self-governance that applies to both infrastructures are as follows.

- The difficulty in centralized management as discussed in Johnson and Post (1997) especially of existing territorial sovereign management.
- The loss of grassroot innovation if everything has to be agreed upon before being allowed. An example would be that the popularity of ringtone downloads by Japanese

iMode users was not expected by NTT DoCoMo and if iMode (mobile Internet) had been centrally managed, this seeming frivolous business model may not have made the cut.[6]

- The Internet protocol design supports self-governance and the Grid developers are learning from the success of the Internet and, designing standards and protocols that would support the Grid as a decentralized infrastructure.

The last point is especially critical for the Grid as well because unlike the Internet where majority of the participants are only consumers of information, all members of the Grid community can be both providers and consumers of resources. Being a provider on the Grid entails one to be able to exercise more control over where or who their resources should be allocated to, whereas consumers are governed by the 'rules' that providers set and can always choose not purchase from the provider who operates in a manner that is inconsistent with their values. If providers are not able to manage their own resources, few will be willing participants of the Grid.

How to Empower Grid Users

How can Grid developers design protocols such that Grid resource producers can easily maintain control over their own resources? A few of the challenges can be addressed here – the issue of decentralized control and the ease of use.

[6] The example comes from Scott Bradner (Senior Technical Consultant, Harvard University; IETF/ITU-T Liaison; Previously, Co-Director, IETF) in his keynote speech "Ad Hoc Networking, Resource Sharing and the IETF", May 8, 2003, Virtual Markets in Wireless Grids' project meeting, Syracuse University, NY.

To maintain control over one's resources, Grid resource producers should be able to influence the following:

- The type of resources for 'sale'
- The quantity available
- The availability period (time)

A resource producer should be able to make his/her above preferences known in a simple manner. One suggestion to do this would be for the potential resource producer to install a piece of software for Grid resource management that can be easily obtained (such as the SETI@Home software which can be readily downloaded and installed) from the relevant Grid that one belongs to. As previously discussed in Section 5.2, it will be very difficult if not impossible for any individual seller to operate without an intermediary who can provide a neutral marketplace for which all transaction parties can converge to. The type of resources to be made available on the Grid are very standardized items and once the preferences are made, the software should be able to detect what kind of computer system one is running. For example, the potential producer goes through a checklist of resources that can be "rented out" and checks off that he/she is willing to rent his/her computational resource on the Grid. Then the software will run a check to determine that the system that the producer is operating from, e.g. a Pentium IV processor of 2.4 GHz. Quantity is not really a factor here as it is assumed that the Grid will use all or none of the idle computational cycles. However in the case of storage, one should be able to decide how many gigabytes/terabytes/petabytes of space to allocate for Grid purposes. Time wise, computational cycles will only be used for the Grid after the software has detected a period of non-activity from

the owner and once the owner becomes active (through a mouse click or typing), all Grid transactions should be aborted. All these should run in an unobtrusive manner, in the background, without requiring further input from the owner once the preferences are set. The use of trusted software agents to carry out the transaction processes according to the pre-set preferences would aide in the transparency of the transaction process without the owner physically getting involved. (See Buyya, 2002 for more details on how trusted agents can be used.) Subsequent changes to preferences can also be easily made through the amending the preference checklist. Having such simple designs in place can really empower the resource producers, allowing them a sense of control and at the same time, make their participation on the Grid hassle-free.

By making the resource producers go through such a process assumes a default OFF position to participate on the Grid. In cases of the leisure users, this would prevent unwanted cases of "intrusion" into their personal systems. This would ensure that all resource producers are willing participants and not in any way disadvantaged by their ignorance or lack of knowledge to protect their systems. However in the case of the business users, a default ON position is not necessarily unethical or unfair as the resources ultimately belongs to the enterprise and are being put to use for the benefit of the enterprise through the enterprise Grid.

6.3 Cultural/Non-Technical Organizational Barriers

In order for Grid computing and utility computing to enjoy widespread adoption, there must be a market for it. This is especially so in the case of utility computing for the business end users. However there are many non-technical organizational barriers to the adoption of utility computing. These inhibiting factors can be attributed to culture. Culture can defined as *a pattern of shared basic assumptions that the group learned as it solved its problems of external adaptation and internal integration, that has worked well enough to be considered valid and, therefore, to be taught to new members as the correct way you perceive, think, and feel in relation to those problems.* (Schein, 1997) When new ideas challenge the status quo such as utility computing threatening to replace current IT outsourcing methods, people tend to fear what the new changes may bring and be resistant to the changes. The following subsections explore the various organizational issues that may arise in an enterprise when utility computing is implemented. These have been termed as *organizational politics* in a market study report by Platform Computing (Platform, 2003). Policy recommendations are also suggested for overcoming each barrier.

6.3.1 Loss of Control (Or Access to Resources)

This issue topped the list of concerns listed by respondents of the market survey carried out by Platform Computing (Platform, 2003). When employees are used to having resources only at their disposal, persuading them to share the resources (even though it is really an exchange for access to a larger virtual pool) worries them. They are concerned that a shared infrastructure might not allow them to access resources that they already have when they need them, directly affecting their work performance. This is especially true for certain employees who have access

to unique resources that are pertinent to their jobs and asking them to share these resources worries them immensely. However if there is a shared pool of resources in addition to their dedicated resources, employees are not as skeptical. This illustrates that people want to hold on to what they have probably due to the fear of the unknown – the promised larger virtual pool of resources is still a vision unless they are proven wrong.

Recommendation

This is a common misconception among the employees that they will be surrendering access to their resources and thus lowering their productivity level. However with the proper training and consumer education by service providers, employees will come to understand that they can be make use of advanced Grid technologies to easily set up policies and rule mechanisms to control external access of their resources (see Section 6.2.2 on *"How to Empower Grid Users"*).

6.3.2 Staff Displacement

Another issue with the problem of loss of control is that the IT department may itself be resistant to changes. Utility computing changes the ownership of resources from the user to the service provider. This is effectively removing the responsibility of owning and maintaining IT resources from the IT employees over to the service providers. IT employees may deem this as a threat to their jobs and thus refuse to adopt utility computing for fear of being displaced.

Recommendation

Automation also brought about the same fear as assembly-line workers worry about their jobs being replaced by machines. The process of automation may have eliminated some jobs but they have also created others. In the same way, with the proper training by service providers, internal IT staff can move on to new value-added positions that utility computing brings. Some transfer of IT staff from the enterprise to the service provider may also be necessary since the internal staff are familiar with the IT needs of the enterprise and can better assist the service providers in meeting the needs of the enterprise.

6.3.3 Risk Adversity & Fear of the Unknown

Generally, people who are happy with the way things are being run, will not like changes. Changes bring uncertainty and uncertainty can bring either good tidings or bad news. As the saying goes, "if it isn't broken, don't fix it". As the complexity of IT has grown immensely over the years, changing something can sometimes be an arduous task. The culture has always to stick with tried and tested methods that have been proven to do the job. If current IT provisioning models work fine, IT staffs are risk adverse in changing it for something that has yet to yield results. It is not only the IT staffs that are in question here. Decisions makers such as the CTO (Chief Technology Officer), the CIO (Chief Information Officer) or any CXOs (Chief Executive Officers) for that matter, may not be keen to take unnecessary risks in trying something new.

Recommendation

As utility computing is still in its nascent stage, consumer education is extremely important to dispel any confusion that end users may have due to the media hype. As with any decisions that are made, it has to be strategic. By presenting proofs of ROI (Return on Investment) and cost effectiveness, it would definitely excite any CXO into seriously considering the implementation of utility computing. On top of that, educate them on the changes that will come with the new implementations so that they are prepared and know how to deal with them when the time comes. Having an internal champion who is in the position to make decisions would really help forward the case of service providers.

Careful management of the risks involved is very critical. Implementing the new system in a way that is palatable with the enterprise such as seeking their opinions and inputs before proceeding and doing a phased implementation or establishing a pilot programs are ways that can put enterprise at ease with this new and ambitious project.

6.3.4 Trust (Or The Lack Thereof)

Despite the many proponents utility computing touting its benefits, there are always the skeptics who still see this as a fantasy. There are also others who would not deem a third-party trustworthy enough to entirely handle the IT needs of the enterprise. This lack of trust in utility service providers can stem from issues of security and privacy, which have yet to be properly addressed. This general lack of distrust can also arise from the perception that an outsider who

does not know the business processes of the company well enough, cannot possibly provide the kind of IT resources on a utility basis to adequately meet the needs of the enterprise.

Recommendation

The issue here revolves around the proper education of the end users to establish trust in utility computing system and, the viability of the service provider and the technology. To dispel myths and hype, service providers have to produce proof of ROI and set reasonable goals that utility computing can achieve for the enterprise. The education should be targeted at decision makers so that service providers can establish an internal champion for their ideas.

Technology-wise, security issues are currently being hashed out at both GGF and Globus. It would not be long before a viable solution can be found to overcome this problem. Even as security issues are being are thrashed out, enterprises can still implement utility computing as a private utility that skirts over these issues.

6.3.5 Dependency

According to another set of results from an IDC survey (Tapper, 2003), dependency is another non-technical inhibiting factor to the implementation of utility computing. Enterprises must ultimately surrender the ownership of resources to service providers and enterprises are afraid of the lock-in by service providers or being left stranded if the sole service provider does not perform up to their expectations.

Recommendation

Either market forces or the government would see to it that the utility computing market does not turn into a monopoly. Given a reasonable number of players in the market and the ease of switching to another service provider due to the interoperable nature of the Grid computing, enterprises should not afraid of being stranded.

7 CONCLUSION

Although Grid technologies are still developing, the huge potential of the Grid cannot be dismissed. This is especially so for the commercial sector, particularly in the area in IT provision. The Grid is envisioned as a platform that can radically change the way IT resources are being provisioned through utility computing. Thus it becomes of grave importance to IT industry players to be aware of the impact to them, in order that they are prepared for the changes and not be caught off-guard.

The author predicts that the there will be five major players in the new utility computing ecosystem. They are the *Grid resource supplier*, the *Grid infrastructure supplier*, the *utility service provider*, the *re-seller* and the *end user*. Further industry analysis reveals that there are new roles for current players in the traditional IT provision industry and opportunities for new entrants in this new ecosystem. The author hopes that this thesis has shed some light on the characteristics of each role to help industry players better understand the requirements of the new roles and enable them to see how and where they would fit in this new ecosystem. Current players in the IT provision industry would have to decide which of the above roles to play in this new utility computing ecosystem and to re-define their market strategies accordingly. This thesis also suggests that players in the telecommunications sector, who want a share of this growing pie, can enter as new entrants to the field by leveraging on their strengths.

Policy-wise, some of the challenges examined in this thesis are Grid standardization for inter-operability, decentralized Grid governance to encourage Grid resource sharing and the proper consumer education to ensure that cultural/organizational barriers would not inhibit the

commercial adoption of Grid computing. The author suggests that all interested parties to the Grid computing market be involved in Grid standards bodies to avoid any bias of the standards towards any one party, particularly for the commercial end users. As for Grid governance, the initial governmental intervention is deemed as inevitable as a jump-start a promising project financially. However as the commercial sectors start to get involved, market forces would be at play. The author also suggests that Grid governance be extended to every end user in a simple user-friendly way since participation on the Grid requires a lot of control and trust in the system that can only be achieved if one believes that they are in control of their own resources. The last policy issue highlights the need for consumer education to transcend the cultural/organizational barriers to Grid computing and recommends ways of doing so.

As Grid technologies mature further and more Grid capabilities become a reality, there are more to be even more issues to grapple with. Future research would have to include areas of security, economics and jurisdiction. Security is a huge concern for enterprises that are looking seriously into Grid computing and intensive technical research in this area is underway to afford end users greater levels of protection. Technical research into the areas of metering and usage monitoring are also important to realize the Grid potential of pay-as-you-use paradigm. The laws too must evolve to cope with the various changes to IT provisions across institutional or geographical boundaries.

As the Internet makes the transition from the wired to the mobile and wireless, the Grid too can expand into the wireless domain. Future research into the viability of the wireless Grid would be appropriate as microprocessors become small enough to be embedded into everyday devices and

wireless technologies become more prevalent and affordable. Although there are many limitations to integrating wireless devices onto the wireless Grids, some research work has shown that these limitations can be overcome and the wireless devices can prove to be invaluable assets to the wireless Grid. There is definitely a lot of potential for growth in this area as the NSF has funded a project a wireless project titled "Virtual Markets in Wireless Communication and Computation Grids". (http://wirelessgrids.net/)

With the immense interest of the various sectors in Grid computing and the on-going research in the above-mentioned areas, the Grid may live up to its hype and deliver its promise as the next revolution in computing in the near future.

REFERENCES

1 Becker, D. (March 21, 2003), " PlayStation 3: The Next Generation", http://news.com.com/2100-1040-866288.html , CNET News.com.

2 Buyya, R. (2002), "Economic-Based Distributed Resource Management and Scheduling for Grid Computing", Doctoral Thesis, School of Computer Science and Software Engineering, Monash University, Melbourne, Australia.

3 Caldwell, B. (December 11, 2001), "Outsourcing Uptake Will Surge as IT Utility Matures", Research Brief, Gartner, Inc., Stamford, CT.

4 Caldwell, B. (November 14, 2001), "IT Outsourcing Services Growth Continues in North American Market", Dataquest Alert, Gartner, Inc., Stamford, CT.

5 Claunch, C. and McCoy, D. (January 15, 2003), "Predictions to Watch in 2003", Note Number: COM-19-2110, Research Note, Gartner, Inc., Stamford, CT.

6 Couture, A. (May 07, 2002), "Looking Back on SSPs and Ahead to Storage on Demand", Gartner Dataquest Perspective, Gartner, Inc., Stamford, CT.

7 Devine, A., and Holmqvist, S. (2001), "Mobile Internet Content Providers and their Business Models: What Can Sweden Learn From the Japanese Experience?", Master Thesis, Industrial Engineering and Management, The Royal Institute of Technology, Stockholm, Sweden.

8 DiCenzo, C. (June 20, 2001), "What Continues to Drive Storage Growth?", Market Analysis, Gartner, Inc., Stamford, CT.

9 Eriksen, L. (August 06, 2003), "Why Utility Computing Will Succeed Where ASPs and Outsourcing Failed", http://www.utilitycomputing.com/news/355.asp , Utility Computing, London, United Kingdom.

10 Eriksen, L. (July 30, 2003), "Will the Real Utility Computing Model Please Stand Up", http://www.utilitycomputing.com/news/342.asp , Utility Computing, London, United Kingdom.

11 Ferreira, L. et al (December 2002), "Introduction to Grid Computing With Globus", IBM Redbook SG24-6895-00, IBM Corporation, Armonk, NY.

12 Foster, I., and Kesselman, C. (eds) (1999), *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, San Francisco, CA.

13 Foster, I., Kesselman, C., and Tueke, S. (2001), "The Anatomy of the Grid", *The International Journal of High Performance Computing Applications*, 15(3), pp. 200-222.

14 Gabler, J. (February 24, 2003), "An IT Utility Is a Logical Approach for Providing IT Services", Note Number: TU-19-1450, Gartner, Inc., Stamford, CT.

15 Gabler, J. (February 24, 2003), "Flexibility and Satisfied Users With IT Services Utility Model", Note Number: COM-19-0808, Gartner, Inc., Stamford, CT.

16 Gentzsch, W. (August 23, 2001), "Grid Computing: A New Technology for the Advanced Web", White Paper, Sun Microsystems, Inc., Palo Alto, CA.

17 Gentzsch, W. (October 1, 2001), "Grid Computing: A New Technology for the Advanced Web", White Paper, Sun Microsystems, Inc., Palo Alto, CA.

18 Gerrard, M. (May 08, 2001), "Is it Time to Deregulate Your IT 'Utility'?", Note Number: DF-13-3013, Gartner, Inc., Stamford, CT.

19 Gillett, S., and Kapor, M. (1997), "The Self-Governing Internet: Coordination by Design", *Coordinating the Internet*, Kahin, B., and Keller, J. (eds) (1997), The MIT Press, Cambridge, MA.

20 Gurpreet, D. (eds) (2002), *Social Responsibility In the Information Age: Issues and Controversies*, The Idea Group Publishing, Hershey, PA.

21 Haas, L., and Lin, E. (March 2002), "IBM Federated Database Technology", IBM Corporation, Armonk, NY.

22 Hagel, J., and Brown, J. S. (2002), "Service Grids: The Missing Link in Web Services", http://www.johnhagel.com/paper_servicegrid.pdf

23 IBM Global Services (January 2002), "E-Business on Demand: The Next Paradigm", #G510-3002-00, Somers, NY.

24 Jagannathan, S., Srinivasan, J., and Kalman, J. (2002), *Internet Commerce Metrics and Models in the New Era of Accountability*, Prentice Hall, Upper Saddle River, NJ.

25 Johnson, D., and Post, D. (1997), "And How Should the Net Be Governed?: A Mediation on the Relative Virtues of Decentralized, Emergent Law", *Coordinating the Internet*, Kahin, B., and Keller, J. (eds) (1997), The MIT Press, Cambridge, MA.

26 Kahin, B., and Keller, J. (eds) (1997), *Coordinating the Internet*, The MIT Press, Cambridge, MA.

27 Kerstetter, J. (June 23, 2003), "Business Software: Comes The Revolution", BusinessWeek, US Edition.

28 Klingenstein, K. J. (1999), "Middleware: The Second Layer of IT Infrastructure", *CAUSE/EFFECT Journal*, Volume 22, Number 4.

29 Lessig, L. (1999), *Code and Other Laws of Cyberspace*, Basic Books, New York, NY.

30 MacKinnon, B. (2003), "Commercial Computational Grids: A Road Map", ACM: Ubiquity, Issue 14 (May 27 - June 2, 2003), http://info.acm.org/ubiquity/views/b_mackinnon_1.html .

31 Mattson, P. (2000), "Application Service Provider - A Business Plan", Master Thesis, Sloan School of Management and School of Engineering, Massachusetts Institute of Technology, Cambridge, MA.

32 McKnight, L., and Bailey, J. (eds) (1997), *Internet Economics*, The MIT Press, Cambridge, MA.

33 Phan, T., Huang, L., and Dulan, C. (2002), "Challenge: Integrating Mobile Wireless Devices Into the Computational Grid", *Proceedings of the 8th ACM International Conference on Mobile Computing and Networking (MobiCom '02)*, September 25-27, 2002, Atlanta, GA.

34 Philipkoski, K. (January 19, 2001), "A Super Computer Collaboration", http://www.wired.com/news/technology/0,1282,41306,00.html , Wired News, San Francisco, CA.

35 Pindyck, R., and Rubinfeld, D. (eds) (1998), *Mircoeconomics*, Prentice Hall, Upper Saddle River, NJ.

36 Platform Computing, Inc. (March 2003), "Market Study Report: The Non-Technical Barriers to Implementing Shared Computing in a Commercial Environment", Markham, Ontario, Canada.

37 Platform Computing, Inc. (November 2002), "A Guide to Harnessing Grid Computing in the Enterprise", Markham, Ontario, Canada.

38 Rold, C., and Berg, T. (February 07, 2003), "Sourcing Strategies: Relationship Models and Case Studies", Note Number: R-18-9925, Gartner, Inc., Stamford, CT.

39 Schein, E. (1997), *Organizational Culture and Leadership*, Jossey-Bass, Inc., San Francisco, CA

40 Schopf, J., and Nitzberg, B. (2000), "Mobile Internet Content Providers and their Business Models: What Can Sweden Learn From the Japanese Experience?", TR #CS-00-05, Computer Science Department, Northwestern University, Evanston, IL.

41 Shapiro, C., and Varian, H. (1999), *Information Rules*, Harvard Business School Press, Boston, MA.

42 Sullivan, T., and Scannell, E. (October 06, 2000), "Big Guns Rattle ASP Battlefield", http://archive.infoworld.com/articles/hn/xml/00/10/09/001009hnsoftserve.xml , InfoWorld Media Group, San Francisco, CA.

43 Tapper, D. (March 2003), "Utility Computing: A Look at Demand-Side Needs for On-Demand Computing", IDC #28864, IDC, Framingham, MA.

44 Tapper, D., and Goepfert, J. (May 2003), "Wall St Is Bullish on Utility Computing", IDC #29373, IDC, Framingham, MA.

45 The Economist (June 5, 2003), "Software's Jolly Iconoclast", New York, NY.

46 The Economist (May 8, 2003), "Déjà vu All Over Again", New York, NY.

47 The Economist (May 8, 2003), "The Fortune of the Commons", New York, NY.

48 UtilityComputing.com (May 27, 2003), "An Opportunity for Telecom Growth - Grid Computing", http://www.utilitycomputing.com/news/296.asp , Utility Computing, London, United Kingdom.

49 Villacastin, R., "Managing On-Demand Computing", Marketing Presentation Slides, http://www.ca.com.my/marketing/presentation.html , Computer Associates (CA), Kuala Lumpur, Malaysia.

50 Weissman, J., and Lee, B., "The Service Grid: Supporting Scalable Heterogeneous Services in Wide-Area Networks", Department of Computer Science and Engineering, University of Minnesota, Twin Cities, MN.

51 Wolski, R., Plank, J., and Brevik, J. (April 2000) "G-Commerce -- Building Computational Marketplaces for the Computational Grid", University of Tennessee Technical Report UT-CS-00-439.

52 Wolski, R., Plank, J., Brevik, J. and Bryan, T. (2001) "Analyzing Market-based Resource Allocation Strategies for the Computational Grid", *The International Journal of High-performance Computing Applications*, Volume 15, Number 3.