NEXT GENERATION INTERNET APPLICATIONS:

WEB SERVICES, PEER-TO-PEER & GRID COMPUTING

An Infocomm Technology Roadmap Report

Release February 2002

Infocomm **D**evelopment **A**uthority OF Singapore

Table Of Contents

List of Figures and Tables

Acknowledgements

We would like to thank the following organisations and individuals for their participation in interviews and the drafting of Track 2– **NEXT GENERATION INTERNET APPLICATIONS: WEB SERVICES, PEER-TO-PEER AND GRID COMPUTING** of the 3rd Infocomm Technology Roadmap (ITR-3) report:

1.	Dr Gunaretnam Rajagopal	BioInformatics Institute (BII)
2.	Mr Larry Ang	BioInformatics Institute (BII)
3.	Mr Wilson Lai	BioInformatics Institute (BII)
4.	Mr Thomas Yee	GridNode Pte. Ltd.
5.	Mr Pang Kang Ming	GridNode Pte. Ltd
6.	Mr Lawrence Chong	Hewlett-Packard Singapore Pte. Ltd.
7.	Mr Lim Teck Wee	Hewlett-Packard Singapore Pte. Ltd.
8.	Mr Pang Hee Kiang	Hewlett-Packard Singapore Pte. Ltd.
9.	Mr Seah Yu Hsin	Hewlett-Packard Singapore Pte. Ltd.
10.	Ms Lum Seow Khun	IBM Singapore Pte. Ltd.
11.	Mr Justin R Martin	IBM Singapore Pte. Ltd.
12.	Ms Rachel Soon	IBM Singapore Pte. Ltd.
13.	Ms Chua Horng Shya	Infocomm Development Authority of Singapore (IDA)
14.	Ms June Koh	Infocomm Development Authority of Singapore (IDA)
15.	Ms Evelyn Loh	Infocomm Development Authority of Singapore (IDA)
16.	Mr Alvin Lee	Infocomm Development Authority of Singapore (IDA)
17.	Ms Cynthia Yap	Infocomm Development Authority of Singapore (IDA)
18.	Dr Cheok Beng Teck	Institute of High Performance Computing (IHPC)
19.	Dr Chin Sai Kong	Institute of High Performance Computing (IHPC)
20.	Dr Chong Wai Keong	Institute of High Performance Computing (IHPC)
21.	Mr Randy E Hall	Intel Corporation
22.	Dr Milan Milenkovic	Intel Corporation
23.	Dr Guo Lih Shiew	Intel Technology Asia Pte. Ltd
24.	Dr Jit Biswas	Kent Ridge Digital Laboratories (KRDL)
25.	Dr Robert Deng	Kent Ridge Digital Laboratories (KRDL)
26.	Dr Pang Hwee Hwa	Kent Ridge Digital Laboratories (KRDL)
27.	Ms Chan Kar Wei	Microsoft Singapore Pte. Ltd.
28.	Mr Steven Cheah	Microsoft Singapore Pte. Ltd.
29.	Dr Teo Yong Meng	School of Computing, NUS
30.	Prof. Cai Wentong	School of Computer Engineering, NTU
31.	Prof. Francis Lee Bu Sung	School of Computer Engineering, NTU
32.	Prof. Stephen John Turner	School of Computer Engineering, NTU
33.	Assoc. Prof. Lim Ee Peng	School of Computer Engineering, NTU
34.	Asst. Prof Bertil Schmidt	School of Computer Engineering, NTU
35.	Mr Keith Budge	Oracle Corporation Singapore Pte. Ltd.
36.	Mr David Forden	Oracle Corporation Singapore Pte. Ltd.
37.	Mr Parthasarathy Subramanian	Oracle Corporation Singapore Pte. Ltd.
38.	Dr Wolfgang Gentzsch	Sun Microsystems
39.	Mr Philips Lai	Sun Microsystems Pte. Ltd.
40.	Dr Simon See	Sun Microsystems Pte. Ltd.
41.	Mr Tan Meng Wai	Sun Microsystems Pte. Ltd.

ITR-3 Track 2 Roadmap Task Force

IDA - Technology Direction

Mr Raymond Lee
Ms Lim Chay Yong
Ms Caren Choy
Mr Goh Eng Siang
Mr Ngin Hoon Tong

Dr Brian Chen
Chief Technology Officer
Infocomm Development Authority of Singapore

Important note: All market figures and market forecasts are quoted from leading analysts in this field, readers should note that there could be wide discrepancies between different analyst houses. However, our objective is to give an approximate information on market sizes to complement this report, which above all should be regarded as a technology trend report than a market forecast report. Readers should hence exercise caution in interpreting these market figures.

Executive Summary

INTRODUCTION

In this report on Next Generation Internet Applications (NGIA), the future landscape is visualised as a picture painted by the culmination of market drivers and a manifestation of key technology trends and feature the following characteristics:

- Virtualisation, where resources will be pooled from both closed local systems and open global access, over geographical boundaries, over ownership barriers and pooled from heterogeneous environments for unified access.
- An extension of control to Edge Resources towards a Decentralised and Distributed Architecture.
- The Ubiquitous Internet, an Internet accessed by a diversity of devices.
- A concept called immersion, characterised by the 2-way web, more media-rich, absorbing and engaging interactivity, and deepening functionality.
- A higher level of collaboration, facilitated by an increase in business2business, machine2machine and application2application interactions.
- An evolution of market and economic dynamics to a service and subscription based model.
- A paradigm shift in software development methodologies, adopting an increasingly agile and on-demand model.
- An underlying layer of standards for interoperability and interactivity between heterogeneous environments, pulling down barriers to openness.

In this arising eventuality, where computing and communications will be viewed as a single entity, and when technology trends and market forces are melded to form a platform for the NGIA, Distributed Computing, in particular Web services, Peer-to-peer (P2P) and Grid Computing will be viewed as engines of change.

Web Services

Web services are self-contained, self-describing, loosely coupled software components that can be described, published, discovered, and invoked over a network, and is based on a Service Oriented Architecture. Web services is expected to bring forth the next phase in the evolution of e-business, which is based on the concept that functionality can be represented as services, and these services can be dynamically discovered and integrated by using messaging through the network. The key benefits of Web services are that it enables businesses to implement applications significantly faster and it allows for potentially large savings in both implementation and maintenance costs.

Extensible Markup Language (XML), Simple Object Access Protocol (SOAP), Web Services Description Language (WSDL) and Universal Description, Discovery and Integration (UDDI) form the core technology for Web services. More higher layer functions, such as multi-party transactions models, authentication and security, dynamic service discovery, are required to deliver the promises of Web services. Development in these areas has already started and we expect standards to mature and industry to reach some consensus in the next year or so. Web services is likely to enter mainstream sometime towards 2004 to 2005.

Peer-To-Peer

P2P computing is re-emerging as a viable architecture for many applications after the phenomenal success in Napster. It refers to the sharing of computer resources and services by direct exchange between systems. It differs from the client-server computing model in that each edge system can act both as a client and a server in the context of some application. The updated definition of P2P is well captured by Clay Shirky, principal at The Accelerator Group, as *"A class of applications that takes advantage of resources like storage, cycles, content and even human presence available at the edges of the Internet."*

The initial drivers of this present P2P wave were mainly social. Applications like ICQ and Napster, that have brought P2P into the limelight, empowers its users, allowing them to easily create their own online communities and run them as they collectively choose. Enterprises begin to see the potential of P2P in many business applications such as collaboration, edge resources pooling and distributed network applications.

A lot of innovation in the P2P space today is driven by startup companies, such as Entropia and Groove Networks. In the area of standards, there has been little progress, so most P2P applications are not interoperable and have few things in common. The adoption of standards is however, essential for P2P to take-off in a big way, we therefore believe the adoption of standards through one of the following scenarios in the next 2 to 3 years:

- P2P Working Group agree on adopting certain best practices as standards, and/or
- Adoption of Web services standards, and/or
- Adoption of Sun's Juxtapose (JXTA) or Microsoft's .NET development platforms, and/or
- Emergence of de-facto standards due to market consolidation.

Grid Computing

The Grid is an emerging hardware and software infrastructure that serves as a distributed, robust and secure platform for access to high-end computational capabilities. Grid Computing's primary purpose is to pool together geographically dispersed resources and present it as a unified source for users to use as a metasystem. Examples of resources that are pooled together over a Grid are:

- Supercomputers and Clusters for grand challenge problems and processing-hungry tasks such as protein folding, high energy physics and earth observation
- Spare Cycles from installed systems banded into a Virtual Supercomputer
- Diverse Databases aggregated to form a Data Grid

This is extend into using Grid as computing platform, augmenting the local system and allowing the user open global access to resources on the world-wide Internet.

Grid Computing is a technology area that is being explored and advanced by the old-guard academic and research institutions such as National Centre for Supercomputing Applications at University of Illinois, Argonne National Labs, California Institute of Technology, European Organization for Nuclear Research to solve grand challenge problems and support national level projects such as Large Hadron Collider, Biotechnology, National Collaboratories and Petaflop computing power. The year 2001 saw Grid Computing on the receiving end of sizeable amounts of funding and national level projects such as TeraGrid Distributed Terascale Facility (DTF), United Kingdom (UK) National Grid, Dutch Grid, Korean Grid being launched.

Grid Computing as a field of study has evolved from the 80s. The differentiating factor now is firstly, an increase in commercial interest from the industry and secondly, the emergence of a de-facto standard -- Globus Toolkit, which is being adopted and explored by corporations such as IBM, Microsoft, Sun, Intel, Fujitsu and NEC.

Grid Computing has the potential to be adopted on a landscape sweeping scale, with diverse implementations in a permutation of Grids, such as national Grids, enterprise Grids, data Grids, sensor Grids, open market on-demand Grids for metered access to multi-vendor resources, hospital Grids and access Grids.

Currently, we see public awareness and the corresponding hype beginning. The current status of Grid Computing is still in early adopter and field trial phase. The propagators behind current Grids are mostly in the science and research community. Technology transfer to the industry and partnerships are taking place. These Grids (such as TeraGrid DTF, European Union DataGrid, UK National Grid) are expected to go live around 2003 to 2004. Forecasts from analysts such as Bear and Stearns, Deutsche Bank and IBM themselves, predict that the Grid will start to gain momentum in 2004 to 2005. Irving Wladawsky-Berger, Vice-President Technology and Strategy of IBM Server Group, predicts that Grid Computing will reach the commercial world in 2003 to 2004 or sooner. IDC predicts that the market for on demand computing will reach more than US$30 billion later in the decade. Bear and Stearns forecasts that distributed processing is expected to garner US$8 billion in 2004. The use of Grid will bring about a paradigm shift towards a service-oriented business model.

Conclusion

We believe that these technologies, Web services, P2P and Grid Computing will take root in the landscape, bringing significant change to the Internet environment. These distributed computing technologies will be manifested in broad reaching concepts such as Utility-like computing, Virtualisation, Services on Demand, the Ubiquitous Internet, and the Internet as a Computing Platform.

These technologies contribute to a change in the market and the technology user landscape. Software vendors move from a product and license fee structure to services platform provider and subscription-based model, resource providers evolve beyond selling hardware and services to be an utility provider of metered processing cycles. Data Centres become Grid Centres, providing access to data, computational and application resources. Skill sets also change in line with industry needs.

In this report, we look at the vision of the technology, its potential, its shape, form and usage, taking the views of analysts, industry leaders and researchers from labs around the world, undeterred by this nascent technology's current look and feel.

All the three technologies that we discussed in this report are at different stages of development and deployment, and will continue to evolve. They each face different challenges and still have many technical hurdles to overcome before materialising their vision and achieving mainstream adoption. Although some of the business models and usage pattern have yet to be established and proven, we see the merits of the technologies and their great potential in fulfilling many of the emerging market and business requirements, and the vision of NGIA.

1 Introduction to Next Generation Internet Applications (NGIA)

1.1 WHAT IS NGIA?

The Internet has become the de-facto communication platform connecting computers, devices, services and people across different geographical boundaries. Since its birth in the late 60s as US Department of Defence's (DoD) Advanced Research Project Agency Network (ARPANET), it has grown dramatically in size and sophistication. We have witnessed its revolutionary growth in the Web era in the 90s after the Internet first became "commercialised" in the beginning of the decade.

In Track 1 of this 3rd Infocomm Technology Roadmap (ITR-3), we presented our insights and vision of the Next Generation Network based on advanced photonics technologies and integrated Internet Protocol (IP) intelligence. In this second track of ITR-3, we envision the way Internet will be used above the underlying intelligent optical transport network, several years down the road. We term this the *Next Generation Internet Applications (NGIA)* – the emerging computing paradigm that defines how resources, data, application and services will be intertwined to meet the need of future activities over the Internet.

1.2 THE NGIA VISION

We advocate that the *NGIA* is about weaving together devices, services, businesses and even people on top of an intelligent, robust, secure infrastructure into an intricate web, which will stretch the boundaries of the Internet and bring the level of collaboration to a more sophisticated and immersive level. We believe that NGIA will be characterised by:

1. Virtualisation, where resources will be pooled from both closed local systems and open global access, over geographical boundaries, over ownership barriers and pooled from heterogeneous environments for unified access.
2. An extension of control to Edge Resources towards a Decentralised and Distributed Architecture.
3. The Ubiquitous Internet, an Internet accessed by a diversity of devices.
4. A concept called immersion, characterised by the 2-way web, more media-rich, absorbing and engaging interactivity, and deepening functionality.
5. A higher level of collaboration, facilitated by an increase in business2business, machine2machine and application2application interactions.
6. An evolution of market and economic dynamics to a service and subscription based model.
7. A paradigm shift in software development methodologies, adopting an increasingly agile and on-demand model.
8. An underlying layer of standards for interoperability and interactivity between heterogeneous environments, pulling down barriers to openness.

We will explore the key drivers and trends that contribute to the NGIA manifestation in the subsequent sections of this chapter.

We believe three technologies, namely ***Web services, Peer-to-Peer (P2P), and Grid Computing*** will be the dominant movements shaping the development of NGIA. The following chapters of this report will present the key developments in these three technological areas and how we perceive the course of development will be like in the next few years. We start by discussing in greater details the NGIA vision and its characteristics.

1. **Virtualisation, where resources will be pooled from both closed local systems and open global access, over geographical boundaries, over ownership barriers and pulled from heterogeneous environments for unified access.**

   Geographical boundaries will be blurred as associated latencies are reduced to insignificance. You will be able to use geographically dispersed resources as if they were on your local system. You will find no differences between your local hard drive and those from your distributed storage providers. The model of outsourcing will evolve to a on-demand environment, where the access to distributed resources will be as secure, reliable and dependable as (or even surpass) using centralised and local resources. Virtual Computers will be the norm, and gigabit applications will gain mainstream acceptance.

2. **An extension of control to Edge Resources towards a Decentralised and Distributed Architecture.**

   The current Internet has everything concentrated at the network centre. Content, processing, and storage are hosted on servers, while PCs are used mainly as presentation-level devices. The next stage of evolution will see the unravelling of the centre, where the edge devices no longer function as simple consumer of content and computing resources, but become more participative in nature. Edge resources will become more dominant, interacting and collaborating with fellow peers and network centres alike, thereby creating a more decentralised and distributed architecture.

3. **The Ubiquitous Internet.**

Through Internet enabled devices, the network will be everywhere. The scope of the Internet will spread beyond the circle formed by PCs and web-sites today, to other devices and appliances.

- In the Next Generation Internet (NGI), most computing devices will be connected to the network 24 hours a day and 7 days a week. This will happen in phases.
- The PCs are already permanently tethered to this network and to each other. Many mobile phones and Personal Digital Assistant (PDA) are also entering this cloud.
- Following this is the spread of this cloud to our everyday life as seen in mobile access in our in-car computing system, network access in our household appliances and embedded devices scattered in sensor networks.

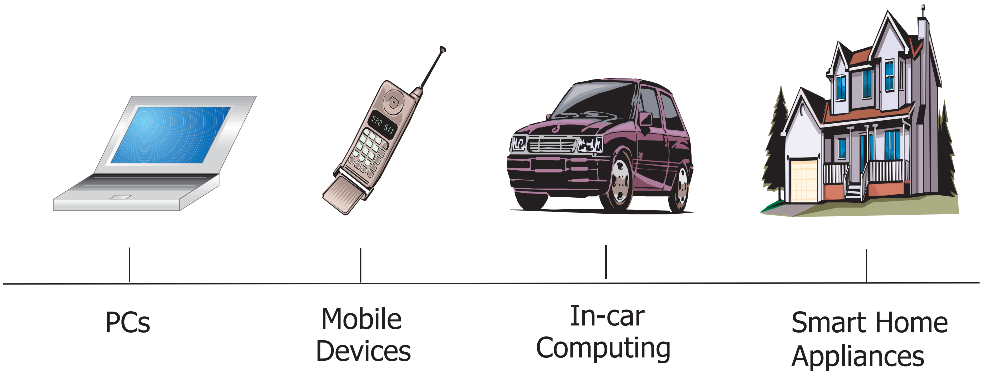


Figure 1. Ubiquitous Internet through a diversity of equipment

This swarm of computing devices extends the reach of the Internet beyond the domain of the desktop. In a reinforcing cyclic effect, the more devices, application and users that there are in the network, the more useful and valuable it will be, leading to even more devices, users and applications wanting to enter this cloud – a manifestation of Metcalfe's Law.

4. **A concept called Immersion, characterised by the 2-way web, more media-rich and engaging interactivity and deepening functionality.**

   This diversity of computing devices and their differing configurations will change the nature of the Internet. From catering to a network of PCs accessing it through a keyboard and browser, it will stretch to accommodate other devices that want access. Voice portals and navigation through speech commands, intuitive interfaces with Natural Language Processing increase the multi-modal nature of the network. The Internet will evolve beyond the web with its minute animations and decorative graphics into an electronic atmosphere that will surround and immerse the user in interactive entertainment. The 2-way web will be manifested in a broadening of the Internet from a unidirectional broadcasting medium to a bi-directional interactive medium.

5. **A higher level of collaboration, facilitated by an increase in business2business, machine2machine and application2application interaction.**

   In this cloud of devices, applications and users bound to each other, they will find a computing world where they are not limited to local resources and where they do not need to work alone. Imagine a scenario where a university wants to build a working model of how a human body works. Scientists from around the world contribute their algorithms and working models in different data formats and simulation sequences. These heterogeneous resources are pooled and compiled over data Grid, sent to a processing power provider to generate a 2-hour high quality interactive simulation, then sent to a storage provider to be saved. Teachers from around the world then access this e-video to conduct biology lessons.

   A user with a desktop computer with 1-gigabit processing power can tap the resources of a collective network to pool together teraflops of processing power.

   In the NGI, the roles and responsibility of a device or an application is multifaceted. On a normal basis, your desktop PC is a dumb client, control and monitoring being handled by the server. But when your company's engineering department performs their monthly design simulation run, your desktop PC is turned into a node in a enterprise wide processing farm, contributing your spare cycles for use by this engineering department. From being a resource user, you also morph into the role of a provider of services.

6. **An evolution of market and economic dynamics to a service and subscription based model.**

   The vision of Web services, P2P and Grid Computing promises a future where computing will be more decentralised, distributed and service-oriented. Grid computing will be used as an Internet computing platform to enable metered, utility-like access to a Grid of resources. These resources may be a pool of supercomputers, a P2P network of PCs, data warehouses or software tools scattered around the world. Access to these diverse resources, whether through a browser, a mobile device or transparently through an Operating System (OS) will be metered using a variety of economic and accounting models such as auction and subscription.

7. **A paradigm shift in software development methodologies, adopting an increasingly agile and on-demand model.**

   The nature of software development will change from local teams of development in a closed environment and proprietary solutions to an open concept of exposing code functionality to other departments, business partners, vendors and the open public. Service-Oriented Architecture (SOA) and the Services-Oriented Development of Applications (SODA) are two related concepts that best describes the characteristics of delivery and development of software services for this new paradigm. These two models will significantly shorten software development life cycle, reusability will increase, and static code structures will evolve to be more dynamic.

8. **An underlying layer of standards for interoperability and interactivity between heterogeneous environments, pulling down barriers to openness.**

   This motley mass of varying devices plugged into the Internet and attached to each other via a multiple of network, be it wireline, wireless, the public mobile network or the office Local Area Network (LAN), will form a complex and intricate and heterogeneous cloud, and all this will work over IP. This aptly describes the climate of the NGI, where a growing acceptance for heterogeneous platforms, interfaces, networks, systems and applications is prevalent. Be it languages such as Microsoft's C#, Sun's Java, old world Common Business Oriented Language (COBOL) or the classic C, there will be a myriad of languages. Processors and internal architectures range from Intel and Motorola to Nokia. Resource Providers may be the traditional server or a peer machine.

   And above all, all these myriad of technologies will work with each other using standards and community efforts. Industry players will still try to be dominant players, but consumer demands and Metcalfe's law will drive interoperability.

1.3 KEY MARKET DRIVERS AND TRENDS

This section discusses the key market drivers and trends that contribute to the shape and manifestation of NGIA.

1.3.1 From a hierarchical centralised corporation to a dispersed dynamic organisation.

Corporations are on a globalisation drive, which results in a pattern of an organisation distributed geographically and whose work is co-ordinated through electronic communications.

"Your backyard is someone else's front room." -- this insight from Peter Drucker highlights that support and secondary functions in your organisation may be outsourced to be someone else's core line of business. For example, in this model, production is outsourced to a contract manufacturer, who is viewed as a close business partner and an external department.

This gives rise to the concept of *Virtual Organisation,* which can be viewed in two ways:

- Single Global Corporation - an organisation distributed geographically and whose work is co-ordinated through electronic communications
- A gathering of Separate Legal Units - a network of independent companies linked by IT to share skills, costs, and access to one another's markets. They aggregate together and provide a set of services and functionality, so to the outside world, they appear to be one organisation.

With the widespread and increasing adoption of this model, we see that a corporation function becoming increasingly distributed.

1.3.2 Increase emphasis on inter-enterprise process workflow integration

In the past decades of computerisation within corporations, emphasis has been on Management Information Systems (MIS), office automation and workflow (e.g. email) and employee productivity (PC tools). However, many of the inter-organisational administrative functions and workflow are still manual. Today, it is common for an enterprise to have a wireless LAN, the sleekest laptops and cutting edge enterprise software, but send printed-out invoices to business partners using snail-mail. Businesses are now re-orienting to emphasise on efficiency and effectiveness of their process and information flow, both within the organisation and with external business linkages, and their back-office function integration. Integration of supply chains across different businesses through automated business process from purchase order initiation, through procurement, manufacturing and logistics are top priorities in many businesses.

Computerisation is now moving from within the enterprise to the gaps between organisations. The IT aspects *within* an organisation (such as ERM, accounting, employee productivity) are beginning to be saturated, and the industry is looking at the opportunities in computing in the work areas where organisations interact with one and another such as purchasing, selling, distribution and employment.

1.3.3 Move towards a Service Oriented Model

Revenue Pattern apparent - The market is beginning to shift towards services as the revenue from this area gains weight. Analysts such as IDC and Bear Stearns have highlighted that the cost of software and maintenance of storage is about 4 times the cost of the actual hardware assets. Gartner Dataquest forecasts that the global IT services market to reach $1.4 trillion in 2005. IBM highlighted that more than 40% of its $90 billion in revenue was derived from their Global Services division.

Service Oriented Model speeds up implementation - In order to use an application, an organisation needs to lay in a whole support hardware, software and manpower infrastructure. For example, an email application system will require the organisation to purchase the hardware, the email server software as well as the numerous client licenses. In addition to this core set-up, the organisation will also need to have administrators to monitor usage and perform maintenance, set usage policies, set up protective systems such as virus detection.

The Service Oriented Models aims to eliminate the need for users to make capital expenditures for servers and software licenses, and move to fast and as-needed access of computing base on subscription or metered usage fees.

Technology has a Short Life Span - With the rate of innovation, technical products have increasingly short life spans. In order to be competitive, the organisation must continuously tear out infrastructure and upgrade their systems. This scenario is becoming increasingly dissatisfactory to consumers. For example, the mature PC market, we see consumer leasing their machines from vendors, or arranging some form of upgrade deal.

1.3.4 Use of technology as strategic & competitive advantage

From Amazon to e-brokerages, sustainable business models built on a technology platform are surfacing. In this trend, technology is shifting from its role as an enabler for productivity to a core business function. With this growing importance comes the added responsibility of reliability, robustness and security.

One example of use of technology for strategic advantage is TSMC (Taiwan Semiconductor Manufacturing Co) and their Virtual Foundry initiative. The Virtual Foundry service, an Internet application, enables online collaboration for design, monitoring of IC chips production and supply chain management.

Another example is e-brokerage pioneer Charles Schwab, whose online brokerage model is built on the Internet. Competing in this sector are firms like ABN AMRO and Morgan Stanley, who are utilising technology as a competitive edge in providing the user with a range of service offerings at low cost per transaction.

Celera, an online information business, provides information based on the human genome and related biological and medical information, charging for access to their proprietary databases and use of proprietary Bioinformatics tools, is another business model dependent on technology and the Internet.

1.4 KEY TECHNOLOGY TRENDS

In the undercurrent of this converged landscape are key technology trends that will shape, affect and influence the success of NGIA, as outlined below.

1.4.1 Extensible Markup Language (XML)

XML is a metadata language for defining semantic tags to describe a document's structure and meaning. It allows various industries and professions such as music, medicine, and semiconductor, to develop their own domain-specific markup languages.

XML has enabled almost anything on the net to be treated as an object tagged with meaningful information, enabling the web to become an organised space rather than a big mess of information. It has great impact on how applications and platforms across the Internet will interact and share data. Its use will penetrate to almost all aspects of the Internet, ranging from web pages publication, search engines, data interchange, messaging, software agents, to machine-to-machine interaction. Systems are becoming more context-aware. In the NGI, Search Engines will differentiate between you searching for the latest rice prices or Anne Rice's novels.

1.4.2 Network Ubiquity and Everything-Over-IP

Network will be ubiquitous, built using the multiplicity of network technologies. Wireless & wireline, public and private are converging to packet-based data networks. IP has become the universal transport protocol for data traffic. Network interoperability and connectivity are non-issues in the modern connected world. Advances in optical networking and photonics, wireline and wireless broadband access will continue to deliver greater bandwidth, network intelligence, Quality of Service (QoS) capability and ease of maintenance.

1.4.3 Everything Digitised

Assiduously, the objects around us in the business and home environment are going digital – digitised voice, video, music, graphics, books, library, medical images and others. Knowledge is increasingly being stored digitally. Digitisation is paving the way for many new possibilities – information processing, storage, data mining and analysis, search and index, and others. It brings about new uses such as interactive TV, online purchase of e-books, enabling search for videos based on a screen shot or find a song using a few notes. It will require support layer such as Digital Rights Management (DRM), micro-payment, and accounting mechanisms.

1.4.4 The Chip beyond the PC

Chips with powerful processor and memory storage are increasing being embedded in consumer devices, like cars, cell phones and home appliances. This has brought about intelligence, memory and storage function to a multitude of edge devices. Software services can be pre-loaded or downloaded dynamically over the Net, enriching the functions of these devices. According to Forrester Research, the world only has about 450 million PCs, but the *extended Internet,* including embedded chips, will increase the Internet population by billions of nodes.

1.4.5 Open Source and Standards

An important trend is open community efforts to ensure interoperability and the ability for Internet users to connect to each other in a Metcalfe-ian network. Open source and standards are considered to be open community driven efforts.

Several factors are contributing to the strength of this key trend.

- Customer allergy to vendor lock-in continues - They will want the NGI to be open, global in nature and accessible over the different implementations, such as web sites being built using differing vendors products.
- Vendors are wary of dominating proprietary platforms - Standards are being driven as components of a common open platform strategy.
- Learn from experience – the industry has learnt that the global spread of the web is fuelled by the adoption of standards such as Uniform Resource Locator (URL), Hyper-Text Transfer Protocol (HTTP), Hyper-Text Markup Language (HTML).

1.5 THREE KEY APPLICATIONS IDENTIFIED FOR NGIA

In viewing computing and communications as a single entity, and taking into consideration the characteristics for the NGIA, we identify Web services, P2P and Grid Computing as its key pillars.

Web Services

- The promise of Web services lies in the expected costs savings, speed of deployment and reusability. It involves a wave of change encompassing the business models, user paradigms and a change in development methodology to SODA.
- An overview of Web services, its status and key trends are highlighted in Chapter 2.

P2P

- P2P garnered attention when Napster's adoption rate exceeded that of email and instant messaging (IM), and when the viral growth caused copyright authorities to shut down their servers. Now P2P characteristics of decentralisation, self-healing and autonomy is vaulted to address some of the client-server woes.
- An overview of P2P is highlighted in Chapter 3.

Grid Computing

- Grid Computing's potent combination of a utility-like access and large-scale resource pooling capabilities are anticipated to give rise to a new class of applications. Grid Computing is driven by the academia and is still in early adopter phase. However, its vision of turning the internet into a computing platform, to a service oriented model delivering instant, reliable access to resources dispersed globally gels perfectly with the future landscape of global and virtual organisations characterised by large amount of inter-enterprise collaborations.

- An overview of Grid Computing is highlighted in Chapter 4.

2 Web Services

2.1 OVERVIEW

2.1.1 Definition

Web services are self-contained, self-describing, loosely coupled software components that can be described, published, discovered, and invoked over a network. They can dynamically locate and interact with other Web services component on the network to provide a service without intervention. Web services comply with certain industry standard specification, and is platform and language neutral.

Web services is based on a Service-Oriented Architecture[1] (SOA) (See *Figure 2*), which comprises three entities and three fundamental operations. The three entities are: ***Service Consumer, Service Provider,*** and ***Service Broker.*** They interact with each other using the three fundamental operations: ***Publish, Locate*** and ***Bind,*** to provide a loosely coupled computing paradigm.

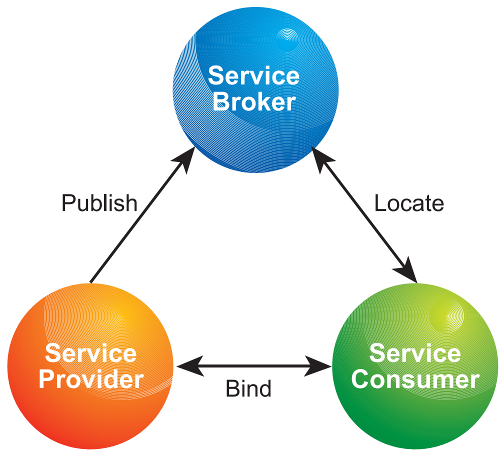


Figure 2. Service-Oriented Architecture

A basic transaction in the SOA normally consist of the following steps:

- A Service Provider advertises its services by publishing them through a Service Broker.
- A Service Consumer hunts for a particular service through the Service Broker.
- The Service Broker locates the Service Provider for the Service Consumer.
- The Service Consumer sends requests to the Service Producer.
- The Service Provider responds to the requests of the Service Consumer, binding them together for the duration of the transaction.

[1] Note that the Service-Oriented Architecture is not unique to Web Services, and can apply to services that are unrelated to the Internet or Web deployment.

2.1.2 Evolution of Applications Integration

A substantial amount of Information Technology (IT) resources is spent each year on developing and maintaining integration points between applications. Three types of application integration have evolved over the years:

- **Simple replication -** simple duplication of all or portions of an application, creating a single master with multiple slaves.
- **Data integration -** describes the use of tools that move data from one data source to another.
- **Function integration -** involves one application programmatically invoking code that lies in another application. This includes Application Program Interface (API), Remote Procedure Call (RPC), Java Remote Method Invocation (RMI) and Common Object Request Broker Architecture (CORBA).

Various methods exist for replicating all or portions of an application for applications integration. This includes installing multiple instances of the application either manually or in an automated fashion with software distribution tools. Data integration can be done through batch data transfer, data replication or Extract, Transform and Load (ETL) solution. ETL tools provide custom data extract, file transfer and data loading from various applications to data warehouses.

Most Enterprise Application Integration (EAI) tools today are predominantly based on data integration. With EAI tools, changes in a data element in the source application trigger a small extract that is sent and updated in the target application while applying the target application's business logic. The limitation however is the lack of real-time transactional capabilities.

Multiple attempts at function integration have been developed that allow one application to invoke an API of another application. Various options include Distributed Computing Environment RPC, CORBA, and RMI in Java. However, these technologies have their own component model and platform environment, and were relatively difficult for programmers. As such, they failed to gain significant market share and momentum to succeed.

2.1.3 Enter Web Services

The recent development of Web services technology delivers significant promise for application integration. Web services technology provides a standardised method of publishing and subscribing to software services that are available over the Internet. A Simple Object Access Protocol (SOAP) interface can be layered on top of a new or existing piece of code to allow it to be accessed by other applications over the Internet.

Applications can locate these services using Universal Discovery, Description and Integration (UDDI) and determine the interface definition of the service using Web Services Description Language (WSDL). In essence, Web services technology can be thought of as an evolution of function integration technology such as CORBA.

A major advantage of Web services over CORBA is that Web services do not require integration of an Object Request Broker, which can be very complex. The underlying transport protocol behind Web services is over HTTP, so it is fully compatible with firewalls and the Internet.

2.1.4 Types of Web Services

Implementations of Web services will fall within a spectrum ranging from simple to complex. At the starting end of the range, a simple Web services involves request and response between two services, and is more informational in nature. In general, Simple Web services mean that customers, suppliers, or trading partners can obtain information from an organisation's back-end business systems. Some examples are weather reports, flight schedules, stock quotes, credit card validation and news feed.

In order for businesses to obtain the full benefit of Web services, true business to business collaboration requirement will go beyond informational Web services. For instance in the case of a secured supply chain marketplace, where buyers and suppliers collaborate and compete for orders and the fulfilment of those orders, numerous document exchanges will occur, including request for quotes, returned quotes, purchase order requests, purchase order confirmations, delivery information and more. Transactional Web services for Business-to-Business (B2B) and Business-to-Consumer (B2C) are required to meet the needs. Other examples of transactional Web services include airline reservations and purchase order/ fulfilment.

At the other end of the spectrum, much more complex Web services for business linkages and process integration at a workflow or process level will be required.

2.1.5 Why is Web Services Important?

Web services is expected to bring forth the next phase in the evolution of e-business. This phase is derived from the concept that functionality can be represented as services, and these services can be dynamically discovered and orchestrated by using messaging through the network. As a technology, Web services is capable of providing considerable benefits to organisations and consumers as it can impact the functionality of an application in the following ways:

- Easily embedded. This increases application features, while reducing complexity.
- Functionality from discrete applications can be consolidated, reducing duplication. This simplifies application development and maintenance.
- Organised into teams or locations. Functionality has the capability to be organised into teams or locations. This allows for organisations to be more competitive.
- Range can be widened. Given this larger choice of functionalities to choose from, consumers are in a better position to select the best features for the job required.
- Can be easily outsourced. This allows organisations to focus on their core business more easily.

Apart from the technical capabilities of Web services, there are several important business reasons why their use could be applicable:

- Lower costs. Investments in software that are Web services enabled are protected because of the open standards based approach, allowing for large savings in cost. These cost savings are can be realised in both existing and future applications.
- Strong Vendor Support. It is rare that vendors have agreed to support unifying standards for a new technology but Web services is one such technology whereby the

standards are almost universally accepted, leading to its fast adoption.

- Interoperability with other e-business protocols. RosettaNet and electronic business extensible Markup Language (ebXML) are widely accepted by their own groups of users. On the integration of Web services with these protocols, these users will be able to integrate with one another.
- Ease of use. Web services is a technology that now comes with a large range of available tools. As a result, business users are able to piece together several Web services to deliver new services.
- Development productivity. As there is now a wide variety of tools available for Web services and they are based on open-standards, developers are easily able to quickly develop complex applications.

2.2 FACTORS DRIVING ADOPTION

We believe that three key factors would drive Web services adoption and they are described below:

2.2.1 Decrease in Product Life Cycles

In today competitive world, the ability to respond quickly to changes in business is often the key factor that affect success or failure of an organisation. Product life cycles are decreasing, due mainly to the high expectations of the customer and much changes in an organisation's process requirements. Enterprise Systems that require substantial implementation efforts to effect changes are beginning to fail in keeping up with the business changes. Web services with its ability to adapt to such changes, in some great way, would be able to help companies adapt to the continual changes in business requirements.

2.2.2 Outsourcing of Non-core Functions

Outsourcing requires some aspect of business interfaces and process integration. However, different legacy process and systems among outsource service providers and its clients (multiple) and suppliers often pose great obstacles. The cost of interfacing is costly and outsourcing contracts would have to be large in order to justify such effort, putting smaller organisation at a disadvantage. Web services could solve this problem, as it is able to bind different systems together with its standards. Furthermore, with the implementation costs low and its faster implementations, smaller organisations can now enjoy the benefits of outsourcing. With Web services, outsourcing integration would be much easier, faster and cheaper, allowing more organisations to enjoy the benefits of this trend.

2.2.3 Integration

The Web is currently filled with HTML documents, which are designed for human interactions. To facilitate e-commerce, interactions based on browsers and HTML were implemented and back office functions such as credit card authorisation and secure transactions were being integrated as the technology matures. The next step in the evolution of e-commerce is the integration of supply chains across different businesses in order to increase efficiency and the bottom line. For this end to end integration to happen, the logical processes within enterprises have to be made portable in order to easily integrate them with the other enterprises. Additionally, a business level services oriented architecture has to be established in order to integrate supply chains across businesses. Web services possesses these characteristics and as such, they are being promoted as the next step in the evolution of e-commerce transactions.

2.3 WEB SERVICES ARCHITECTURE AND TECHNOLOGIES

2.3.1 Architecture

It can be seen from the definition of Web services that interoperability is the key element that must be present for different Web services component to interact with each other. Web services uses three dominant XML standards, WSDL, UDDI and SOAP to accomplish its basic service publication, location and binding functions. SOAP is an XML-based messaging protocol that provides a standardised enveloping mechanism for communicating document-centric messages and remote procedure calls using XML. SOAP messages support the publish, locate and bind operations in the Web services architecture. The document that describes the Web services is the WSDL document. These documents may reference other WSDL documents and by referencing the URLs of other Web services, it is able to request for their services. WSDL defines the interface and mechanics of service interaction and is the de-facto standard for XML-based service description, forming the minimum service description required to support interoperable Web services. The UDDI business registry features an indexed list of the URLs of the WSDL documents and a description of the Web services. As such, UDDI performs the registry function of Web services by providing the service of locating Web services. *Figure 3* depicts how these technologies are related in the Web services architecture.

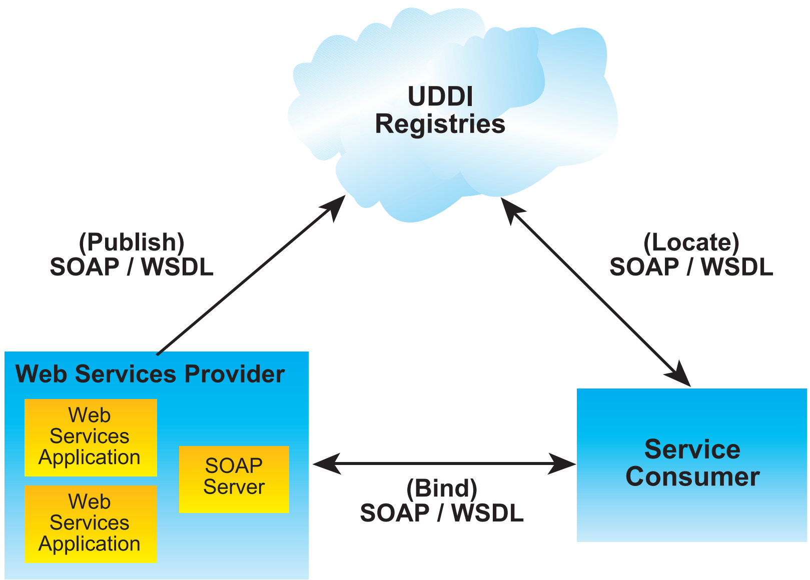


Figure 3. Web Services Architecture

2.3.2 Technology Stack

Until recently, software has been implemented in a monolithic way whereby a complete solution is usually provided by one vendor. This leads to applications that have difficulties in changing with the environment. However, with the arrival of Web services, software applications can now be represented by a collection of related technologies. At a bare minimum, any Web services entails a connection between two applications in which queries and responses are exchanged in XML over HTTP. *Figure 4* below depicts the technology stack that make up Web services.

Higher Level Layers	Business Rules and other extensions
	Other emerging standards and protocols
	Web Services Flow Language (WSFL)
Core Layers	Universal Description, Discovery, and Integration (UDDI)
	Web Services Description Language (WSDL)
	Simple Object Access Protocol (SOAP)
	Extensible Markup Language (XML)
	Common Network Protocols (HTTP and others)

Figure 4. The Web Services Technology Stack

There are two broad layers in the Web services Technology Stack, the Core Layer and the Higher Level Layer. Today, the core layers that define basic Web services communication have been widely accepted and likely to be implemented quite uniformly. The rest of the components within the Higher Level layers that define specific aspects of business processes remain an open question, and it is possible that divergent approaches will emerge.

The development of generally open and accepted standards is a key strength of the coalitions that have been designing and defining Web services architecture. At the same time, these efforts have resulted in a dizzying array of jargon and acronyms.

At the lowest level, we have the Common Network Protocols because Web services must be network-accessible to be invoked by a Service Consumer. HTTP is the de-facto standard protocol for Internet-available Web services. Naturally, other Internet protocols like Simple Mail Transfer Protocol (SMTP) and File Transfer Protocol (FTP) can also be supported. For Intranet-available Web services, Message Queuing (MQ), COBRA and others can be used depending on the application requirements. The following sections will provide an overview of the various layers of the Web services Technology Stack.

2.3.3 Core Layer of the Web Services Stack

Web services is composed of 4 basic standards: XML, SOAP, WSDL and UDDI.

XML. XML is a standard supported by the World Wide Web Consortium (W3C). It is a text-based mark-up language specification. XML defines portable structured data, while differs from HTML, which describes presentation and data.

XML can also be used to define other data descriptive languages, such as vocabularies, messaging protocols and mark-up grammars. XML has been accepted as the standard for data representation in Web services and for describing data types with XML Schemas.

The current version of XML is XML 1.0, which is widely used and there has been no major revisions so far. Late in 2001, the XML Core Working Group has published the first working draft for the XML 1.1, which seeks to address the Unicode character set and line ending issues.

SOAP. SOAP is an XML-based protocol to exchange information in a distributed environment. It defines a messaging protocol between requestor and provider objects. As such it allows the requesting objects to perform a remote method invocation on the providing objects in an object-oriented manner.

The SOAP specification was co-authored by Microsoft, IBM, Lotus, UserLand, and DevelopMentor. Subsequently, the W3C XML Protocol Workgroup was created, comprising over 30 participating companies. In most implementations of SOA by vendors, SOAP is the basis for distributed object communication

SOAP has recently been referred to as the *Services-Oriented Architecture Protocol* due to its use in SOA implementations, even though SOA does not define a messaging protocol. Another strength of SOAP is that it is a vendor neutral standard, allowing implementations to be independent of platform, programming languages and OS, among others.

SOAP version 1.1 is currently in use and version 1.2 has recently been submitted and is still in the working draft stage. SOAP 1.2 seeks to remove ambiguity on how messages can be processed. It also comes with improved namespacing of elements, incorporating W3C's standard method for namespacing.

WSDL. WSDL is an XML-based language describing the functions of a Web services and how to invoke it. Initial work done by Miscrosoft on SCL (SOAP Contract Language) and IBM's NASSL (Network Accessible Service Specification Language) have produced WSDL 1.0. WSDL allows service providers to describe the format of requests and response messages for RMI. WSDL 1.1 was submitted to the W3C in early 2001.

WSDL addresses Service Description Language (SDL) and Interface Description Language (IDL) independent of the underlying protocol and encoding requirements. WSDL provides an abstract language for defining the published operations of a service with their respective parameters and data types. It also addresses the definition of the location and binding details of the service.

UDDI. The UDDI specification was first defined by IBM, Microsoft, and Ariba to define a way to publish and discover information about Web services. The latest version, UDDI 2.0, has just been released. This specification consists of a core information model for UDDI registries and a description of UDDI API specification. The core information model provides four basic information about a web service: business information, service information, binding information and service specification.

Businesses can dynamically connect to services provided by external business partners through UDDI interfaces. There are 3 kinds of UDDI, the white, yellow and green pages.

Although in the near term, there are no plans to support full-featured discovery (e.g. geography-limited searches or bidding and contract negotiation supported by vendors like eLance), UDDI is expected to be the platform of higher level services that are supported by other standards. Further into the future, UDDI also plans to support complex business logic, such as those in hierarchical business organisations.

2.3.4 Higher-Level Layers of the Web Services Stack

Web Services Flow Language (WSFL). WSFL is an XML language to describe Web services compositions. It is layered on top of WSDL. WSDL describes the service endpoints where individual business operations can be accessed, while WSFL uses WSDL to describe service interfaces and their protocol bindings.

WSFL constructs a framework in which service providers and consumers can use to describe the business logic required to assemble various services into an end-to-end business process. This framework allows users to assume the roles of the business process. WSFL is currently in development, and is sponsored by IBM.

XLANG. XLANG is an "XML business process language", which enables developers to aggregate applications as components in a business process. XLANG is a "notation for the specification of message exchange behaviour among participating Web services" and it supports the automation of the business processes.

XLANG is expected be the basis of automated protocol engines which are able to track the status of process instances and help enforce protocol correctness in message flows. WSDL specifies some fundamental concepts and these are the basis for defining an XLANG service. An XLANG service description is a WSDL service description that comes with an extension element that describes how the service behaves as part of a business process. The XLANG service behaviour may also rely on simple WSDL services to provide fundamental functionality to implement the business process. XLANG is considered as Microsoft's answer to IBM's WSFL specification and these two competing standards are fighting for a spot to become the de facto Web services standard.

ebXML. Before Web services, there have been efforts to define a standard XML format for exchanging e-business related information. This standard is focused on Electronic Data Interchange (EDI)-style information exchange. Organisation for the Advancement of Structured Information Standards (OASIS), the group developing ebXML, recently adopted SOAP as a key element of its specification and this may lead vendors to incorporate ebXML into their overall Web services strategies. ebXML was adopted by United Nations Centre for Trade Facilitation and Electronics Business (UN/CEFACT) standards body in May 2001. The ebXML initiative is intended to allow businesses to locate each other, define trading partner agreements, and exchange XML messages in support of business operations, automatically, over the Internet. As there are many similarities between ebXML and SOAP/ WSDL/ UDDI, some convergence is already taking place with the recent adoption of SOAP in the ebXML transport specification. The ebXML architecture defines:

- Business processes and related messages and content
- A registry and discovery mechanism to publish business process sequences with related message exchanges
- Company profiles

- Trading partner agreements
- A uniform message transport layer (mapped to SOAP with multipart MIME attachments)

ebXML is an EDI based standard which aims at addressing the exchange of information for the Internet. EDI is conventionally deployed on expensive dedicated networks, which are tightly coupled, inflexible architectures, thus requiring significant technical expertise. ebXML is currently seen as a competitor to UDDI in the registry functionality due to it being the incumbent standard. However, in the future, we report UDDI to be the standard for Web services registry.

Business Rules. Additional elements that support complex business rules have be implemented before Web services can automate truly critical business processes. Some of these business process include:

- Management – allow sharing of management and usage information through standardised means.
- Non-repudiation – allow for legal non-repudiation needs of message senders and receivers (signed receipts, secure time stamping, etc.)
- Orchestration – describe the full collaborative messaging roles between groups of end-points collaborating on an activity (e.g. multi-party collaboration).

Eventually, mechanisms will be available to the various business processes and they will come in two forms. Some of them will be in standards, while others will come as value-added solutions from independent software vendors.

2.4 MARKET OVERVIEW

Many major vendors such as HP, IBM, Microsoft, Oracle and Sun have announced their Web services architecture. Only Microsoft bases its architecture on Windows based platform, with the rest basing their architecture on Java. Beside the major IT vendors, several smaller companies are also building Web services development frameworks. These players include BEA System's Web Logic, The Mind Electric, which makes Glue, a Java-, XML- and SOAP-based platform for constructing Web services; Eltegra, which produces Exadel, an XML component server and platform; WebCollage and its Syndicator product; and Bowstreet, a startup that produces the Business Web Factory and has announced co-development agreements with both Sun and HP. Even SAP, the Enterprise Resource Planning vendor, has come up with its own concept of Web services.

2.4.1 Market Forecast

The Stencil Group[2] has observed that the hype of Web services is now way ahead of adoption and has identified the middle of 2002 as the tipping point in which adoption will keep up with the hype. Following this milestone, in 2003, the Stencil Group expects the

[2] The Stencil Group, The Stencil Scope: How Web Services Will Beat the "New New Thing" Rap, Jun 2001.

true impact of Web services will emerge and it will far exceed any previously anticipated or perceived impact. It is here that Web services will move from being an emerging technology to being an accepted method of software development and distribution.

According to Gartner's Technology Hype Cycle Trend Analysis, Web services is just reaching the Peak of Inflated Expectation and is on its way to the Trough of Disillusionment. They forecasted that it would be 2003 before Web services starts on the Slope of Productivity. These forecasts imply that it would take about two years before Web services become widely implemented. Gartner Dataquest has forecasted the Web services market to be worth US$1.7 billion by 2003.

It is expected that the first implementations of Web services would be in the B2B arena as Web services can realise great returns from B2B implementations. Next in line, B2C organisations would start to implement Web services for their transactional applications. The sectors most likely to be implemented with Web services in the B2C arena are most likely to be in logistics, billing and shipping as these sectors would find significant benefits in Web services.

Based on these forecasts and the fact that the largest vendors are all jumping into the Web services bandwagon, most vendors are confident that Web services are expected to have a strong impact to both end users and organisations when it matures and becomes accepted in the next 18 to 24 months. In support of this expectation, IDC has also forecasted that Web services will have a substantial impact on the enterprise software market in the next 12 to 24 months.

2.4.2 Factors affecting the Market

Although market forecasts are optimistic about the adoption of Web services in the near future, there are some factors that may inhibit the proliferation of Web services:

- **Maturity of Service-Oriented Architecture.** To achieve the full potential of Web services, the idea of purchasing and consuming services over the Internet must first gain acceptance. Presently, technology and legal issues are still waiting to be resolved. For example, it is unlikely that service providers will publish vast amounts of data onto public UDDI directory without any security standards in place. In addition, there are the issues of trust and accountability. Basic Service Level Agreements (SLAs) must be defined so that consumers can be assured of a respectable level of QoS. Legal guidelines must also be in place to punish rogue service providers who cannot deliver on their SLAs. Therefore, the acceptance of Web services will be hindered if related issues are not sufficiently addressed.

- **Incumbent technology.** One key market for Web services is the EAI segment. Some organisations that are quick to embrace IT have already jumped into the previous bandwagon of EAI. As they have just invested in EAI application to integrate their disparate systems, they are unlikely to invest further on any technology that would replace their expensive incumbent technology, unless they have a critical need for it. As such it would be a few years before these companies would migrate to Web services.

- **Spending freeze.** The current global slowdown has caused many organisations to withhold spending on technology that are not in critical need. As such, many may adopt a wait-and-see approach on this new emerged technology.

2.5 FUTURE TRENDS AND DEVELOPMENT

For Web services to succeed, there are many technical areas that have to be pursued and developed, many of which are related to the open, hostile environment in which they must survive. This section highlights some key areas, such as transactions, security and service discovery, where the industry will channel its resource for further development.

2.5.1 Transactions

One key aspect of e-business is the management and control of complex and distributed multi-platform transactions. Without a multiple transactional model, Web services can only function in simple point to point transactions. Presently, software platforms, like Microsoft's COM+ and Sun's Java 2 Platform Enterprise Edition (J2EE)/ Enterprise Java Beans (EJB) platforms, implement their own specific transaction model, like XLANG and EJB Container respectively.

Two groups were working towards the standardization of transactional protocols:

- **XAML.org** (Transaction Authority Markup Language) and
- **OASIS**

Launched in Oct 2000, XAML is an industry initiative that aims to produce a specification for the XML message interfaces and interaction models of Web services to support the co-ordination and processing of multi-stage transactions on the Internet. XAML initiative has quietly ceased operation after XAML partners have agreed to discontinue their work.

The Business Transaction Protocol (BTP) Technical Committee within OASIS began its work in March 2001 based on some initial proposals from BEA, Choreology and HP. The Pre-final Draft 0.9 version of the BTP Specification was published in Oct 2001. Two other revised versions were published in Dec 2001. It specifies communication protocol bindings for interdependent workflow among multiple trading partners to ensure reliable business transactions. Two-phase commit is used to ensure transaction integrity. It is well suited for use in a Web services environment as it is designed to allow transactional coordination of services offered by multiple autonomous organisations. OASIS tries to speed the adoption of this protocol by avoiding unnecessary dependencies on other standards.

Microsoft appears to follow its own path with its XLANG specification, which it labels "work in progress". IBM is expected to come up with its own specification, Web Services Endpoint Language (WSEL), which works with the WSFL to describe non-operational characteristics of service endpoints, such as quality of service properties. WSEL is currently under development.

It remains to be seen how the transactional standards would consolidate. It is evident, however, that there will be a shakeout in these standards otherwise true Web services business transactions would not be feasible. The OASIS BTP initiative appears to be promising, however, the initiative lacks membership of key vendors, such as IBM and Microsoft. Microsoft, with its XLANG, may fare well in this standards war, but it would be too soon to tell if XLANG would be adopted by the other vendors for transactional purposes.

2.5.2 Security

Security is yet another inherently complex issue in any distributed environment. Security framework and solution to address end-to-end authentication, authorisation, confidentiality and integrity must be formulated. Besides, there must be mechanisms for audit and non-repudiation in Web services usage and transaction.

Today, the lowest common denominator seems to be simple transport level authentication. This just allows an application to identify the Web server from which the current request originated. There is no way of identifying and authenticating the user (individual or enterprise) of the Web service, which makes it difficult to prevent unauthorised access of exposed Web services and the underlying business service implementations. Users should be able to define policy-based authorisation and authentication of Web services using a GUI but this level of sophistication is not available today.

Digital signatures provide a convenient, yet secure and private means to identify a signer and therefore it is considered to be an adequate process to authenticate SOAP messages. However, the SOAP specification does not address security issues but allows for them to be implemented as extensions. Hence, digital signatures make particular use of message extensions to SOAP, which are located in the SOAP header. Not all the current SOAP implementations provide support for digital signatures.

In December 2001, W3C launched its XML Key Management Activity based on the XML Key Management Specification (XKMS) submitted by a group of vendors[3] in March the same year. The purpose of the Working Group is to develop an XML application/protocol that allows a simple client to obtain key information such as certificates, management or trust data from a web service. The current XKMS defines a set of protocols for distributing and registering public keys. It is made up of two major components:

- XML Key Information Service Specification (X-KISS) - defines a protocol for a trust service resolving public key information contained in XML Signature (XML-SIG) elements.
- XML Key Registration Service Specification (X-KRSS) - allows a client of the service to delegate all or part of the tasks that are needed to process elements.

XKMS has garnered strong industry support and was perceived as a mandate for a second generation PKI standard. "Pre-standard" XKMS-enabled services and developer's toolkits are already offered by VeriSign and Entrust. Ratification of the XKMS by W3C is expected to go beyond 2003.

[3] VeriSign, Microsoft, webMethods, Baltimore Technologies, CitiGroup, Hewlett Packard, IBM, IONA Technologies, PureEdge submitted the XML Key Management Specification to W3C, and Reuters Limited.

The XKMS complements the proposed standard for XML-SIG jointly developed by W3C and Internet Engineering Task Force (IETF). This joint Working Group aims to create an XML compliant syntax for representing the digital signature for web resources and procedures for computing and verifying such signatures. The specifications were submitted for approval as IETF Draft Standard and W3C Full Recommendation in mid 2001. Several vendors such as Entrust, VeriSign and IBM have XML Signature tools available today.

Another important standard is the Security Assertion Markup Language (SAML) from the OASIS Security Services Technical Committee. It is an XML security standard for exchanging authentication and authorisation information. Such functions complement the XKMS's web-based Public Key Infrastructure (PKI) by providing non-repudiation in Web services transaction.

Security in Web services is still rudimentary and individual vendors have different approaches to it, adding to the incompatibility of Web services of different security standards. As there is currently a lack in efforts to develop unifying security standards for Web services, this is one area that has to be further developed before reliable Web services based applications can be implemented universally.

2.5.3 Towards Application Integration

Web services as a technology offers a platform-neutral approach to application integration. Its key strength lies in the loosely coupled service-oriented architecture, standard-based communication model, and the use of lightweight and firewall-friendly technologies such as SOAP and HTTP. This eliminates the tight integration and proprietary communication used by many EAI solutions today. We expect Web services to impact the EAI market significantly when the transaction technologies of Web services become more sophisticated. As the technology is today, with its request/response nature, it is able to achieve just a subset of EAI functionality in the near term. Even then, it enables quick and non-complex integration. Moving on, EAI vendors will incorporate Web services interface into their product lines. The current proprietary communication mechanisms will be replaced by XML and SOAP, thus enabling their products to be open-platform. This trend will also surface in many enterprise software packages in general, such as ERM, CRM, SCM and data warehouse. Instead of using third party connectors and adapters from EAI vendors, these software will have Web services interfaces to enable easier integration in customer sites.

In order for Web services to be used for true application integration where one application can be made available within the context of another without actually duplicating the application itself, several extensions must be made to the series of Web services specifications. Web services currently lack a mechanism to encapsulate a user interface, which is important in order to package an application and embed it into another application.

For binding at the interface level, a new standard, Web Services User Interface (WSUI) initiative was announced by the Epicentric Inc. in June 2001. This working group attempts to standardise a display and user-interface layer for Web services, allowing for them to be embedded into web sites as end-user application. In Oct 2001, OASIS announced the formation of the Web Services Component Model (WSCM) Technical Committee to define a Web services interface and XML vocabularies to allow companies to deliver Web services to end users via a range of channels such as the browser, the portal, and embedded

Web application. Epicentric Inc. plans to submit the WSUI specification to this Committee, highlighting its intent to collaborate with the industry leaders in the WSCM Technical Committee.

In addressing a gaping hole in Web services technology, WSUI/ WSCM holds the promise to enable Web services to implement true application integration, creating new value, yet reducing implementation costs and these benefits should not be ignored. What is now needed is the momentum to ensure that the future WSUI/ WCSM standard becomes matured enough to feature the required tools to make that happen.

2.5.4 User Identity

Identity forms a key component for Web services because it sets the foundation for individual consumers and businesses to maintain and exchange information securely. The ultimate vision is to achieve ubiquitous single sign-on, decentralised authentication and open authorisation from any device connected to the Internet.

Presently, there are three major competing solutions to this, Microsoft's Passport, AOL's Screen Name Service and The Liberty Alliance Project. Passport is a centralised user authentication service purporting to allow easy and secure authorisation of users to participating web sites. Currently, Passport has a "Passport wallet" that can store credit card and address information for a user's Passport account, for the purpose of doing express purchase online. Other future enhancements to Passport have also been announced by Microsoft. Passport is an integral part of .NET My Services, formerly HailStorm, with most of the its users coming from Microsoft's Hotmail account users. Considering the sensitivities involved in having a single identity database owner, we believe Microsoft may eventually adopt a Federated Passport approach.

On another front, America Online has also rolled out a new unified sign-on service across its properties and partner sites. Codenamed "Magic Carpet" and currently promoted as the "Screen Name Service", it allows its users to sign in with a single click and seamlessly browse sites supporting the new technology. This service derived its identity namespace from AOL, AOL Instant Messenger (AIM), and CompuServe 2000 accounts.

Another competition to Passport comes from the Liberty Alliance Project, formed by an industry consortium comprising industry leading companies, like America Express, AOL Time Warner, Hewlett-Packard Company, Mastercard International, Nokia, NTT DoCoMo, RSA Security, and Sun Microsystems. The Liberty Alliance Project aims to create an open, federated solution for network identity. Unlike Passport, where all the identities and related information are stored on Microsoft's servers, members of the Liberty Alliance Project manage their own end user database and ensure that the use of critical identity information is managed and distributed by the appropriate parties, rather than a central authority. AOL Time Warner has also joined this Project. However, collaboration efforts are required to create the necessary open standards on privacy, security, and trust for the Liberty Alliance to succeed. Presently, Liberty has no specifications to implement yet, but Passport and Screen Name Service is already deployed. Hence, the current implication for this project is mainly of strategic importance.

2.5.5 Services Discovery

The Service oriented architecture of Web services requires a dynamic discovery of services online. UDDI and WSDL are the first attempts towards this goal. UDDI, as it is designed today, is intended mainly to provide technical discovery of services. That is, it allows a program or programmer to find and locate information about Web services published by businesses in Business registries. Advanced search that allow programs to locate a certain product or services based on specific criteria such as price and contractual requirements is not under the current scope. WSDL is also unable to provide much semantics beyond a simple description of the services such as data types, service operation and binding. Together, these two specifications provide a framework for service discovery.

UDDI.org has plans to extend the functionality of UDDI beyond technical discovery. New functions will be incorporated to address the ability to locate product and services, and incorporate hierarchical organisation support. We expect more specification to be done is this area to provide a full-fledge service discovery required to support the true service oriented architecture.

2.5.6 Others

There are other challenges that have to be met before Web services can be used within a typical business environment.

Accountability deals with such issues as duration of access and use of Web services, pricing model and method of payment for Web services. Also related to accountability is the issue of change of ownership and standards regarding the specific use of Web services such as service reuse. Such issues will be resolved in due time as the industry manages to consolidate on definitive practices, protocols and mechanisms.

Reliability concerns issues of measuring and communicating Web services QoS. Such issues include the degree of downtime, and contingencies on a Web services being down, such as deciding between locating other services or waiting for the service in question to be available again. The widespread enforcement of de facto service quality standards can help alleviate such problems.

Scalability deals with the ability of Web services to be sized up from the small applications that they are now to becoming enterprise systems by using scalability mechanisms such as load balancing and fault tolerance. Some possibilities include the potential development of a Web services application server in order for Web services to handle the load of larger applications. Nevertheless it is likely that such limitations will be addressed as the technology matures.

2.5.7 A step toward Semantic Web

Web services will become smarter by being context-aware and being able to share such context with each other in order to provide consistent customised and personalised services and interfaces to a consumer. Standardisation would be required across industry for such context to be widely implemented, for example, a standard XML vocabulary for context descriptions, as well as framework for the sharing such context is required.

Further on into the future, the Semantic Web would be another potential technology in this space. Berners-Lee describes the Semantic Web as “a web of data that can be processed directly or indirectly by machines”. Semantic Web is an extension of the current Web, it is suppose to bring structure and meaning to the content of Web pages, giving machines and software agents the ability to process them. Currently, W3C are looking at establishing standards, like Resource Description Framework (RDF) and RDF Site Summary (RSS) 1.0, to form the initial building blocks of the Semantic Web vision.

Web services is a very first step towards the actualisation of the “Semantic Web”. Currently, Web services is only a platform that provides interoperability between software components. However, some vendors have considered embedding context awareness into Web services to provide some rudimentary form of web intelligence. We expect intelligent Web services to take more concrete form in 3 to 5 years, before it eventually achieve the vision of Semantic Web.

3 Peer-To-Peer

3.1 OVERVIEW

3.1.1 Definition

Peer-to-peer (P2P) computing is the sharing of computer resources and services by direct exchange between systems[4]. It can be best explained in terms of the client-server computing model (see *Figure 5*). In a client-server model, a central server, normally a powerful computer, is dedicated to the managing of disk drives, printers, applications, network traffic or other computing resources, while a client, normally a less powerful personal computers, relies on the server for the centralised computing resources. The key concept underlying P2P computing is that each edge system can function both as a client and as a server, making "peer" the best term to describe the role and relationship of these edge system. The ability of a peer to perform both client and server functions is the central feature of P2P. Note that the client-server computing model and the P2P computing model can co-exist and are not exclusive of each other.

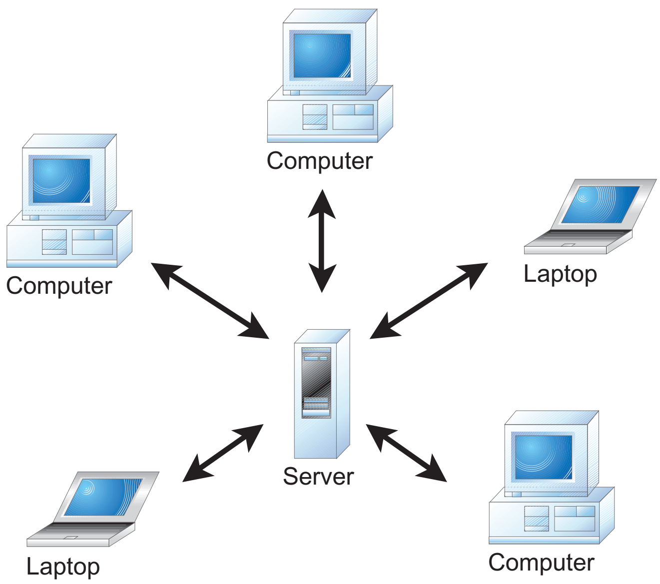


Figure 5. Client-Server Computing Model

In general, there are two P2P computing models, a pure decentralised P2P model and a brokerage P2P model. A pure P2P computing model is one that does not make use of any central server (see *Figure 6*). Each computer or electric device dynamically discover other peers on the network and communicates with them directly. The greatest strength of this model is its robustness as there is no single point of failure. However, decentralised systems are in practice a lot more difficult to scale, due to issues like peer discovery.

[4]Peer-to-Peer Working Group, what is peer-to-peer?, http://www.p2pwg.org/whatis/index.html.

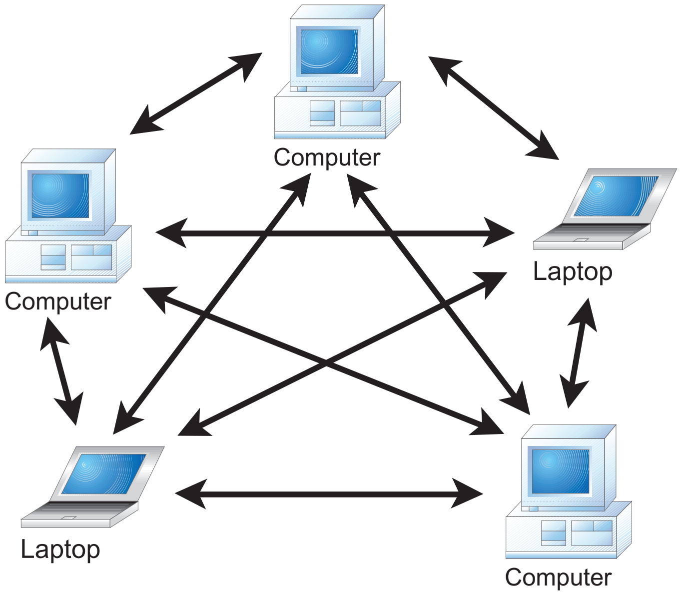


Figure 6. Pure P2P Computing Model

The brokerage P2P computing model is a hybrid between the pure P2P computing model and the client server computing model (*see Figure 7*). It involves the use of a central server, normally for ease of peer discovery. However, the central server can also do content lookup, and even content storage.

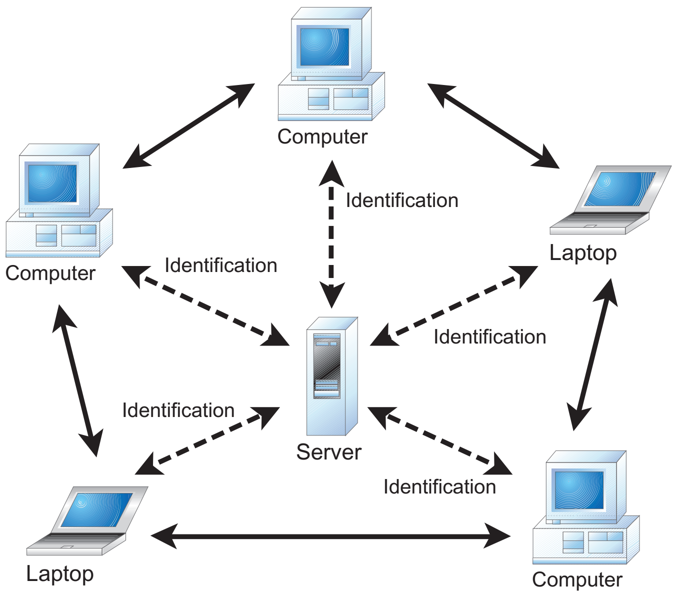


Figure 7. Brokerage P2P Computing Model

The concept of P2P is not new. The traditional meaning of P2P networking involves computers in a workgroup configured for the sharing of resources such as files and printers. Although one computer may act as the file server or Fax server at any given time, all computers in this workgroup could host those services on short notice. All computers in this workgroup will typically be connected using a LAN and run the same networking protocol.

The updated view of P2P computing involves a shift in emphasis from the hardware to the software applications. In addition, users are no longer restricted to dedicated LANs and the Internet has become the network of choice. A good definition of this updated P2P view was proposed by Clay Shirky. It is defined[5] as:

A class of applications that takes advantage of resources like storage, cycles, content and even human presence available at the edges of the Internet.

In the same article, Shirky also identified two primary characteristics that separate P2P applications from other applications:

1. P2P applications must treat **variable connectivity** and **temporary network addresses** as the norm because the need to access and take advantage of decentralised resources at the edges of the network means operating in an environment of unstable connectivity and unpredictable IP addresses.
2. Nodes at the edges of the network must have **significant or total autonomy.**

3.1.2 Properties of P2P

The decentralised computing environment provided by P2P give rise to many appealing properties.

Dynamic and Self-Organising. Any member of a P2P group is free to join or leave because there is no central authority to restrict membership. However, a P2P community may choose to define collectively its acceptable rules of conduct and policies. This is especially true for P2P communities in a corporate environment, where different levels of management hierarchy will lead to different access permissions.

Enduring. Online P2P communities are also enduring because members can relocate geographically without affecting their contacts and relationships within the online community. It is therefore feasible to have an identity that is not tied in to a physical node and a particular IP address.

Resilient. A P2P application has to survive in a very fluid operational environment because it needs to assume that all peers are dynamic and not connected all the time. The decentralised nature of the P2P architecture makes it resilient and fault tolerant by ensuring that there is no single point of failure. Exception will be the hybrid P2P computing model, where the central server, which is normally used for peer discovery and content lookup, is the single point of failure. However, the failure of any one peer will normally not lead to the collapse of the whole system.

Collaborative. The sharing of resources across nodes via direct exchanges creates decentralised environment that is conducive for online collaborative projects and joint activities. These may be structured projects using the P2P environment for software product development by a virtual team or even an assembly of participants who engage in online gaming.

[5] Clay Shirky, What Is P2P… And What Isn't?, http://www.openp2p.com/pub/a/p2p/2000/11/24/shirky1-whatisp2p.html, 24 Nov 2000

3.2 FACTORS DRIVING ADOPTION

The P2P computing model found its trigger for wide acceptance during the few months spanning the end of 1999 and the beginning of 2000. The trigger was Napster. With its popularity, other P2P projects, like Freenet, Search for Extra-Terrestrial Intelligence (SETI)@Home, and Great Internet Mersenne Prime Search (GIMPS), received increased visibility by the public at large. We believe the following factors will drive the adoption of P2P:

3.2.1 Empowering End-users

The initial drivers of this present P2P wave were mainly social. Applications like ICQ and Napster, that have brought P2P into the limelight, empowers its users, allowing them to easily create their own online communities and run them as they collectively choose. Most importantly, these communities are self-organising, dynamic, enduring, resilient, and collaborative. We believe this will continue to be a main driving force behind P2P.

3.2.2 Natural fit for Ad-Hoc Workgroups

With Napster and ICQ gaining good traction, enterprises also begin to see the benefits of P2P and how a P2P networking environment also matches the behaviour of modern work cultures. In the past decade, company boundaries have grown wider, and the need to tie customers and suppliers together to provide an integrated end-to-end solution have become greater. Companies have increased their reliance on temporary workers and consultants, while depending ever more on ad-hoc and cross organisation workgroups. P2P tools, therefore, provide a natural fit for this type of working environment.

3.2.3 Complementary developments in other technology Areas

In addition, several essential developments have reached a stage that makes widespread use of P2P possible. Firstly, the number of Internet-connected computer users have reached critical mass. The ubiquity of connected computers has enabled more and more anywhere, anytime access to the net and its resources. The increased popularity of electronic wireless devices also offers greater opportunities for P2P computing.

Secondly, the increased in communications bandwidth due to the advancement of broadband access technology and optical networking technology have made it possible to move large amounts of data and rich media content from one location to another easily. In addition, technology advancement in the processing power and storage capacity of computers have given the edge computers the ability to handle the extra services required in a P2P environment, breaking out of the old client-server computing model.

3.2.4 Low barrier to Widespread Adoption

Last but not least, P2P applications do not require any significant changes to the average users' usage behaviour. The tools that many P2P applications require is a desktop computer together with the net, which is widely available.

3.3 P2P APPLICATIONS

The majority of P2P applications can be grouped under two main categories: Collaboration and Resource Utilisation (see *Figure 8*).

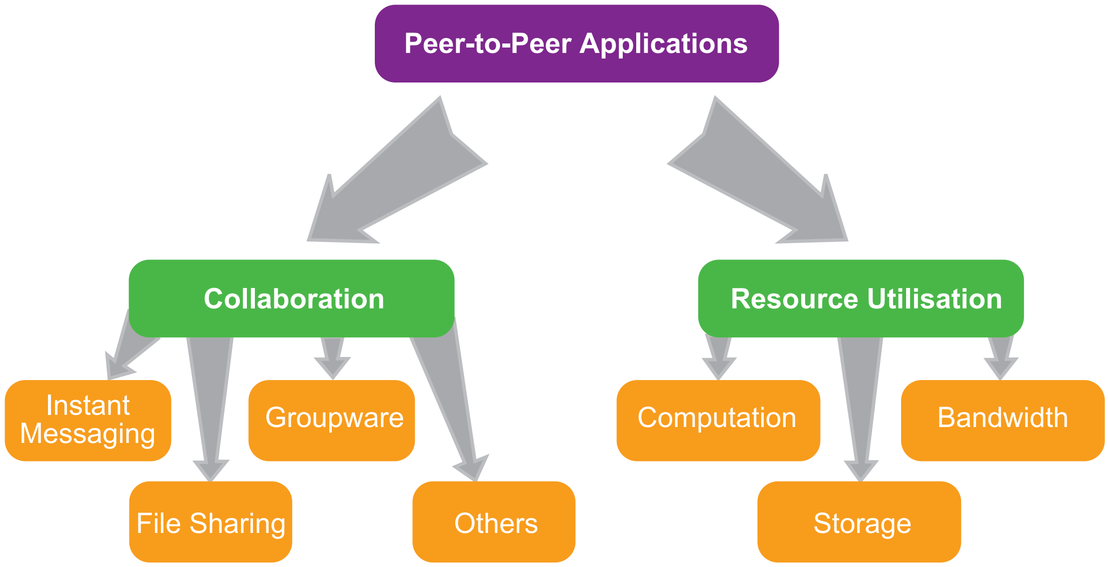


Figure 8. P2P Application Categories

Collaboration. Applications that allow users to access, modify and exchange information in real time. This group of applications can be viewed as the ***new use of P2P*** computing model, whose main attraction is the active communications and ad-hoc information sharing among online communities. The success of ICQ and Napster are often cited as the key developments sparking this P2P collaborative space.

Resource Utilisation. Applications that allow enterprise to leverage available processing power, disk storage and bandwidth capacity within a distributed network. This concept dates back at least to two decades and can be viewed as the ***traditional use of P2P*** computing model. Back in 1982, researchers at Xerox PARC were already building a series of applications around 100 Alto computers linked over an Ethernet network. The earlier efforts were however, focused on collecting resources from a LAN and usually worked on Unix-based workstations. In the mid-1990s, several projects emerged to use the Internet as a coordination medium for volunteer computational efforts. The most significant being those used to crack encryption schemes. In 1996, researchers at University of California Berkeley formed the SETI@Home project to enlist Internet users in the search for extra-terrestrial intelligence and this is often cited as the key development sparking the P2P resource utilisation space.

We will discuss in greater details significant application categories that have emerged in these two groups over the next few sections.

3.3.1 Collaboration

Three significant categories have emerged in this space, they include:

- Instant Messaging (IM)
- File Sharing
- Groupware

Note that some applications span multiple categories; for example, Aimster makes use of an IM platform to do file sharing.

Besides the above three categories, there are other fast emerging P2P collaborative applications, one of the more promising being software-interaction. This group of applications use P2P as the means for exchanging data between a group of applications, allowing different software systems to be linked to each other. Through software-interaction, changes of data at any point in a network of software systems will be propagated to the rest of this network, ensuring that all information in this network of software systems are current and accurate. This group of applications are especially useful for EDI and B2B systems. For example, Engenia Software's RelationshipWeb is a web-enabled collaborative platform specifically targeting B2B solutions, which also interoperate well with existing categories of B2B software such as CRM and ERM. However, we note that Web services are also targeting the same space. Hence, we believe that P2P companies that are already in this space will incorporate Web services technologies into their systems.

Other more prominent P2P collaboration applications that will not be discussed in this report include online gaming, e-Learning and knowledge management.

IM. IM can be viewed as a computer-based technology that sits between synchronous communications (conferencing, real-time voice or text chat) and asynchronous systems (e-mail and fax). Its huge appeal lies in the fact that it carries the immediacy of a phone call and does not carry the long threads like e-mail. Apart from instant text message capability, recent IM systems integrate other services such as e-mail, voice-over-IP, video conferencing and mobile phone Short Messaging Services (SMS).

According to the Wall Street Journal, IM has grown faster than any communications medium so far in history, attracting 80 million users by mid-1999. Currently, META Group estimates the total IM worldwide population at around 150 million clients. With millions of IM subscribers out there, advertisers and vendors are trying to find ways to market and sell to a virtually untapped market. Two very important assets are involved here:

- Identity. The value of an identity datastore lies in its size. A community with a huge identity datastore will generally have a greater wealth of information and computing resources to share (Metcalf's Law) when compared to a community with a smaller datastore, thereby attracting greater participants. Size, therefore, begets growth and market leaders have natural monopolies as a result. Vendors with a huge user identity datastore and strong adoption may license other companies access to its datastore, making it an invaluable asset.

- Presence. Presence allows for real-time (or at least user-time) authentication and access. It also allows for real-time notification and gives applications the ability to take advantage of availability as it arises. For example, presence can be used to determine whether the your buddy is 'online', 'offline', 'video-capable', 'out-of-office', 'don't disturb' or others. This can be invaluable as it allows you to get work done in bursts. However, presence stream consumes bandwidth and processor power, so users will run as few as possible so as not to slow down their computers and therefore will be inclined to stick with their favourite IM application. Therefore, independent service vendors (ISVs) will more likely build their applications on top of the dominant presence stream, making it more valuable.

America Online (AOL) is the current leader in the IM market with over 100 million users and has been very protective of this market, reluctant to interoperate with other IM players and to allow other ISVs to built services on top of its platform. Given their huge lead, we believe it is very difficult for other players to succeed in this IM space unless they already have a huge customer base to leverage on. Microsoft is the exception that may replace AOL as the IM market leader because they can leverage on their desktop OS monopoly by bundling an IM client into the MSN Explorer and Windows OS. Time will tell whether AOL is able to fend off this challenge.

File Sharing. The consumer file-sharing space has so far been an exciting area of intense technical innovation, but sadly poor business case. Most of the file-sharing systems can be described as one of the following categories (see *Figure 9*).

	Brokerage	**Decentralised**
Close	• Aimster • CenterSpan Communications (Scour Exchange) • Napster	**FastTrack Protocol** • Grokster Ltd. • KaZaA • MusicCity Networks (Morpheus)
Open	**Open Nap Protocol** • File Navigator • Autonomous Zone Industries (Mojo Nation)	**Freenet Protocol** **Gnutella Protocol** • Free Peers (Bearshare) • Lime Wire

Figure 9. Categories of file-sharing systems

In the ***brokerage case,*** one or more central server is usually used to maintain an updated list of the files that are available from the users currently online. User queries are run against one or more of the index servers, search results are returned to the requestor and files are transferred directly between peers. The most popular example using this model is Napster. In the ***decentralised case,*** there are no central index server. Users search the system by querying neighbouring peers, who in turn forward the queries to their neighbouring peers if they do not have an answer, and so on. Each peer on the network is responsible for processing incoming queries against its list of locally available content and reporting results back to the requestor. Files are transferred directly between peers. The most popular example using this model is Gnutella.

There are a variety of custom protocols for communication between peers and among peers and index servers. For ***close systems***, these protocols are propriety and unpublished. For ***open systems,*** they are non-propriety and freely available, therefore multiple vendors may produce interoperable applications. An interesting trend to note here is that popular closed systems tend to become open systems through reverse-engineering efforts. For example, the open source community has created the OpenNap Protocol, which is a open source Napster Server Protocol.

The first wave of P2P file-sharing systems were non-profit low cost content distribution networks. Besides the possible cost of setting up and maintaining index servers, the content, storage and distribution are all provided by the users. Currently, the most prevalent money generating models are through banner advertisements and revenue arrangements involving the distribution of companion programs. However, this field is already very crowded and plagued by numerous legal restrictions. To circumvent these issues, companies like Napster, has evolved to become a licensed media distribution company through the use of various digital rights management technology. Other players like KaZaA and MusicCity Networks (Morpheus) have adopted decentralised technical architectures. It remains to be seen whether the decentralised approach will effectively stall legal actions and whether users will find the regulated systems attractive. With the right price and features, regulated systems may still remain viable in terms of user adoption and revenue. Nevertheless, if the evolution of Yahoo! Portal is a good guide, then we can expect only one or at most two survivors in 3 years time.

Another feasible business case will be the enterprise file-sharing model. Most of the difficulties encountered in the Internet file-sharing models, like security and resource ownership, are lessened or absent when applied to the enterprise. In addition, an "Enterprise Napster" model is also very appealing due to its lack of structure and simplicity, that is the users only need to select the folders to share and the P2P software will do the rest. The key difference between the Internet consumer world and the business enterprise environment is that the latter consists of many overlapping groups and different levels of management hierarchy. An enterprise P2P file-sharing system that supports user definition of groups and search capabilities within those groups will address this need. NextPage is the clear leader in the enterprise file-sharing space.

Groupware. Groupware, or collaborative software are applications that facilitate inter-organisation communications. Some of the core functions of groupware include file-sharing, search, calendaring, scheduling, notifications and alerts.

Compared to other sectors of P2P, like IM and file sharing, this sector is the farthest behind in both number of products and breadth of adoption. There are several reasons for this:

1. Groupware requires the integration of many features unlike other P2P collaborative products.
2. Groupware requires a good graphical user interface for ease of use and this requires significant programming investment.
3. Groupware disrupts the current enterprise networking model of security and centralised control. Therefore, it will face significant resistance from the enterprise IT departments.

Due to the difficulty of creating and maintaining a sufficiently useful and user friendly product, coupled with the difficulties of selling into enterprise, a rapid consolidation is expected in this space within the next 12 months.

Traditional groupware solutions rely on client-server installations and as a result suffer from the many difficulties of centralised resources, like version control issues between data on the desktop and data on the server, and duplication of computing resources between individual PCs and shared servers. With the promise of unstructured data sharing, no single point of failure, and easy end-user installation and customisation, P2P groupware is expected to solve or at least mitigate most of these difficulties. However, a sufficiently useful P2P groupware is required before it can compete with the more mature groupware like Lotus Notes, making the development of a P2P product both difficult and time consuming. Presently, Groove Networks is a clear leader in this sector, both in customer base and venture funding. In the next 12 months, companies with good technology features but no sales traction is likely to be acquired by or merge with companies that have existing customer bases. Further down the road, we can expect more P2P groupware features to be embedded into other software applications when they become more mature.

3.3.2 Resource Utilisation

P2P resource utilisation breaks down into three obvious application categories:

- Computation
- Storage
- Bandwidth

Note that any application will probably use more than one of these types of resources; for example, a storage application will need network capacity to transfer stored data to the other peers.

Computation. This refers to the aggregation of Computer Processor Unit (CPU) time over a network of computers, and works best for applications that require intensive simulations or analyses and can be easily split into component blocks to run independently on many computers. These P2P computation systems allow the efficient use of processor downtime, harnessing underused resources to perform tasks that may otherwise require a supercomputer. In addition, the technology can also be exploited to allow different processing devices (such as servers or PCs) to be linked together to share processing resources. Some of the application areas of this model includes:

Application Areas	Description
Financial	Derivatives pricing, risk analysis and fraud detection.
Life Sciences	Biotechnology and pharmaceutical applications like drug target discovery, phylogenetic analysis, immune system simulation and protein folding.
Energy	Oil field analysis that help oil companies drilling in new fields find the best methods to extract the maximum amount of oil from a deposit.
Transportation	Scheduling of aeroplane routes and optimisation of container-ship management.
Telecommunications	Network simulation and design.
Semiconductor	Exhaustive testing of new micro-processor designs.

Storage. Ranging from a simple distributed file system, whereby users can share files with one another easily and quickly, to a complex distributed storage system, where storage components are modular and distributed across geographies to create a clustered, wide area storage network. The main value proposition of P2P storage systems lies in its ease of maintenance because these systems need not be centrally managed and upgraded. For a more complex distributed storage system, there is the addition advantage of geographical dispersion, whereby data can be cached at locations near the source of demand. Some application areas of a distributed storage application includes:

Application Areas	Description
Distributed Databases	Databases with complex matching algorithm can achieve greater lookup speed using distributed computing. E.g. DNA databases, ERM applications and data mining.
Distributed Search	Each peer is responsible for creating or discovering its own data for indexing resulting in faster updates of content changes.
Streaming Media	An extension of Content Delivery Networks like Akamai and Digital Island, where edge computers are used instead of servers.

Bandwidth. This group of applications operates on machines distributed over the network, accelerating data delivery and updates, preventing network traffic from concentrating at a central location, and minimising the cost of operating existing network services using distributed resources across geographies. Some of the distributed-network application areas include:

Application Areas	Description
Load Testing	Provides a simulation of high network load or distributed denial of service attack. This is useful for business reliant on their Web performance.
Quality of Service Testing	A diagnostic tool that provides for continuous testing of a Web site in order to measure and assure its performance from multiple locations on the Internet.
Content Indexing	Search engines like Google can use a P2P network for detecting changes to Web sites so as to increase the search engine's reach and freshness of data.
Network Mapping and Analysis	Using machines scattered across the Internet to create an accurate and up-to-date network traffic map. Useful for companies like Akamai that rely on efficient network connections to provide high-quality content delivery.
Security Monitoring	Using machines at various points of the network to detect, monitor and identify malicious activity.
Distribute Content Delivery	Distributed content platform that allow one peer to serve content to another peer close to it, improving performance and easing the load on the central content server.

Computation and Storage Intensive Applications. Despite the limelight given to public Internet based P2P systems offering solutions to computation and storage intensive applications like SETI@Home, these projects unfortunately do not have a strong business case due to issues related to attracting participates to donate their computing resources. Current, most of these companies are targeting the enterprise environment because by running these applications behind the customer's firewall, the issues of security are more well managed. In addition, the enterprise using the P2P application owns all the computing resources, so the issues of attracting user's donation of computing resources also disappears. The greatest value proposition of using these P2P applications are cost savings and the maximisation of spare computing resources, thereby eliminating the need to do additional hardware investments.

Presently, Platform Computing is the strongest player in this area with nine years of track record, a large customer base of over 600 companies, and healthy profits. Platform targets Electronic Design Automation (EDA), life sciences, Manufacturing Design Automation (MDA) and financial modelling. They have two products: the ***LSF Suite,*** a load sharing system that matches supply and demand of computational resource cycles, and ***SiteAssure Suite,*** which monitors the health and availability of computational resources.

Most of the other players in this field like Entropia and United Devices are start-ups that are still dependent on venture capital funding. Presently, Entropia is one of the most well-funded company in this space and has US$30 million through its third round funding in Jan 2001. Entropia Grid, its enterprise software solutions focus on life sciences (bioinformatics), computational chemistry, financial services and web testing. It has launched several high-profile demonstration projects, such as FightAIDSatHome and drug discovery with Bristol Myers Squibb. On the other hand, United Devices (UD)'s MetaProcessor technology, based on SETI@Home, is a secure distributed platform that allows customers to access more computing power by aggregating the unused or idle processing power, storage resources, and network access of individual computers on a network. UD's main customers are biotech companies and web application service providers. They are working to develop distributed computing applications in the areas of life sciences, web site load and stress testing, Internet QoS monitoring services and research science.

Most of these companies are also viewed as the first wave of commercial Grid Computing companies, because their systems allow the linking of different edge computers to create a single virtual supercomputer. The concept of Grid Computing is however not restricted to the use of P2P, both client-server and P2P computing model can be used to link systems ranging from PCs to mainframe supercomputers. We will discuss Grid Computing in greater detail in the next Chapter.

Bandwidth Intensive Applications. On the other hand, most P2P companies offering solutions to network bandwidth intensive applications target the public Internet environment. The niche markets that these companies have entered, like load testing, QoS testing and network monitoring, offer geographical dispersal as their greatest value proposition. For example, a network monitoring P2P company is able to offer solutions that are capable of monitoring performance reaching the last mile, as well as the rest of the network.

However, we have also noted that these P2P start-ups are unable to match the incumbents in terms of robustness and features. For example, the load and quality-of-service testing markets, and the search-engine market, are currently dominated by Keynote and Google respectively, making the barriers-to-entry for newcomers high. Therefore, P2P companies with interesting solutions but no revenue streams are likely to be absorbed by these incumbents eventually.

3.4 MARKET TRENDS

3.4.1 Market Forecast

Currently, the P2P market is very much driven by startup companies. According to PeerIntelligence, there are approximately 186 P2P companies at the end of June 2001. In terms of business opportunities, Aberdeen Group predicts that the total revenue of the P2P market is expected to hit US$930 million by the end of 2004, from US$20.6 million at the end of 2000.

3.4.2 Future Outlook

We see the future P2P markets in the following four segments.

Winners Emerged. There will probably be two to three big players in each of the categories mentioned. At the present state of the IM industry, AOL is the clear leader.

However, Microsoft is emerging as a strong challenger because given its dominance in the desktop OS market, we believe that Microsoft can easily bundle an IM client into its MSN Explorer and Windows OS to make it the de-facto IM client. If the Internet Browser wars is a good indication, then Microsoft has a good chance of replacing AOL as the market leader.

In the public Internet consumer file-sharing space, we expect this space to follow the evolution of the Internet portal market, where only 1 to 2 big players will survived. Answers to whether a fully decentralised system is able to stall legal actions and whether users will continue to use regulated systems will become clearer in the next 1 to 2 years. The enterprise file-sharing space is equally unclear but NextPage currently demonstrates huge potential to become a significant player.

Groove Networks seems to be a safe bet in the groupware category given their experienced software team and the number of years that they have spend in this area. However, these companies will be under pressure to decide between building up large professional services staffs to do customisation work at the risk of inflating their burn rate, or outsourcing the work to third-party professional services firms such as Accenture and Viant in order to stay focused on building out their technology. These companies will also be potential acquisition targets for large technology firms like Microsoft, Sun, IBM, and Oracle.

For the resource utilisation market, Platform Computing appears to be the clear winner. Given their head start, they have built up a solid customer base that we believe other start-ups cannot compete. Significant challengers may however come from IBM and Sun, big technology vendors that have significant research in the areas of Grid Computing.

Niche Players. For each of the P2P application category, we do believe there will still be several small players offering solutions in specific vertical markets, especially those with very formal rules of document handling and version control such as law firms or drug manufacturers. These companies will be able to generate good margins but may have trouble achieving wide distribution because of the customised nature of their solutions. They will ultimately be take-over targets for service providers or consultancies with expertise in those sectors.

Extensions on Current P2P Applications. In the coming years, we expect new services to be offered on top of the IM and file-sharing platform. In particular, enhancements of IM functionalities and e-mails will make them good enough P2P collaborative groupware. We believe this to be the greatest category in P2P collaboration space because these applications have the advantage of providing incremental changes to the current usage experiences of the already established IM and e-mail user base, unlike the deployment of P2P groupware.

Customised Consulting Projects. We expect system integrators and consultancies to offer custom solutions to their clients, incorporating (or even building on top of) various P2P collaboration and resource utilisation solutions. Existing enterprise software packages, principally CRM, ERM and Knowledge Management (KM) software will take on many of the collaborative aspects of P2P groupware, thus reducing the need for separate groupware packages. Resource intensive software applications will also incorporate P2P resource utilisation components so as to better utilise any idle and spare computer resources in the enterprise. These projects will not be products and will be sold as consulting contracts instead of software licenses. In addition, there will also be opportunities for third parties to extend or enhance existing software solutions, like building additional tools to work within Groove.

3.5 FUTURE TRENDS AND DEVELOPMENT

The P2P market is dominated by start-ups. Presently, there has been little progress in the adoption of general standards in P2P, so most P2P applications do not interoperable with each other and have few things in common. However, we believe that the adoption of general standards is necessary for P2P applications to proliferate, which we believe will appear within the next 2 to 3 years.

Currently, 2 industry consortiums have come together to establish certain P2P standards. Major software vendors, like Sun and Microsoft, have also introduced their P2P development platforms through Juxtapose (JXTA) and .NET. In addition, the adoption of Web services standards by the P2P community is also likely because the Web services standards establishment process is occurring at a much faster pace. However, even if there is little progress in the above three efforts, we can still expect the birth of de-facto standards through market consolidation. In this section, we will discuss in detail the current development in establishing standards and predict the likely development platforms.

3.5.1 Standardisation Efforts

Presently, there are two industry consortiums in the P2P space, New Productivity Initiative (NPI) and Peer-to-peer Working Group (P2PWG).

New Productivity Initiative (NPI). NPI was formed in Nov 2000 by ANSYS, Aurema, Cadence, Compaq, Hewlett Packard, Platform Computing and Teraport. It is an international, industry-lead, working group formed to define and develop an industry specification reference model and a layered set of open APIs (application programming interfaces) for Distributed Resource Management. The NPI Technical Working Group has submitted the Distributed Resource Management Reference Model document to the technical community at large for critical review. Priorities for standardisation will emerge as peer feedback is incorporated into analysis and design activities.

NPI is in the process of establishing a working group within the Global Grid Forum. NPI has also established a close technical working relationship with the principals of the Globus Project.

Peer-to-Peer Working Group (P2PWG). P2PWG was formed in Aug 2000 by Intel Corp., Endeavors Technology, Blue Falcon Networks and Avaki. Many members of NPI are also members of P2PWG. P2PWG was organised to facilitate and accelerate the advancement of infrastructure best-known practices for P2P computing. Currently, the P2PWG standards activity is complementary to the work of the NPI, focusing on a lower level of infrastructure standards such as communications, security, identity, presence, etc. These infrastructure issues are viewed as essential building blocks that will help spur the adoption of P2P once standards can be established, whereas minimum value can be gained through the use of propriety solutions. Presently, five working committees have been formed to address the areas of:

- NAT/Firewall
- Taxonomy/Vocabulary
- Security
- File Services
- Interoperability

3.5.2 Adoption of Web Services Standards

A recent O'Reilly survey revealed the following:

	0% to 10%	10% to 20%	20% to 30%	> 30%
Technologies Used by P2P Companies Surveyed	.NET JXTA Hailstorm	XML-RPC UDDI	SOAP	XML (slightly over 60%)

Slightly over 60% of the companies are using XML and a number of companies are beginning to incorporate features of Microsoft .NET and Sun JXTA in their efforts. More important is the adoption of Web services standards, like SOAP and UDDI, by a significant proportion of P2P companies. With the increased adoption of Web services, we can expect more P2P companies to eventually adopt Web services standards if no consensus is reached by the P2P working groups.

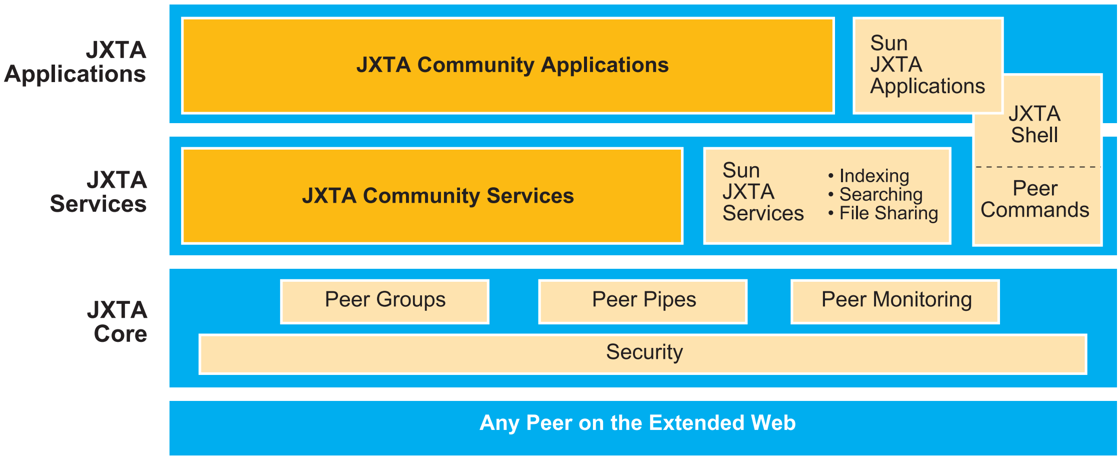
3.5.3 De-facto Standards

In the same study, it was also revealed that most P2P firms see themselves as infrastructure players that provide development platforms, but most of them merely provide APIs or SDKs for their applications silos. It is our opinion that true infrastructure players need to achieve broad industry support.

Presently, with the setting up of a "interoperability working committee" in the P2PWG, the industry consensus point towards allowing market forces to determine the eventual winners in the P2P market. The products and platforms of these eventual winners will then become the de-facto standards for each application area. Although, there are no clear winners yet, except may be for Groove Networks in the collaborative groupware space, we expect the scene to be clearer in the next 12 to 24 months.

3.5.4 Key Platforms and Development

SUN's JXTA started as a research project at SUN Microsystems under the guidance of Bill Joy and Mike Clary to address the P2P space. Its strategy is one of complete openness and standards conformance, from design to XML-based protocols to open-source implementation. While JXTA is fundamentally a specification that is language independent, SUN is pushing for Java to be the best way to create cross-platform and cross-device applications. JXTA's design is derived from the OSI network stack and consists of the JXTA Applications Layer, JXTA Services Layer and JXTA Core Layer sitting on top of any electrical devices (see *Figure 10*).



Courtesy of Sun Microsystems

Figure 10. JXTA Software Layer

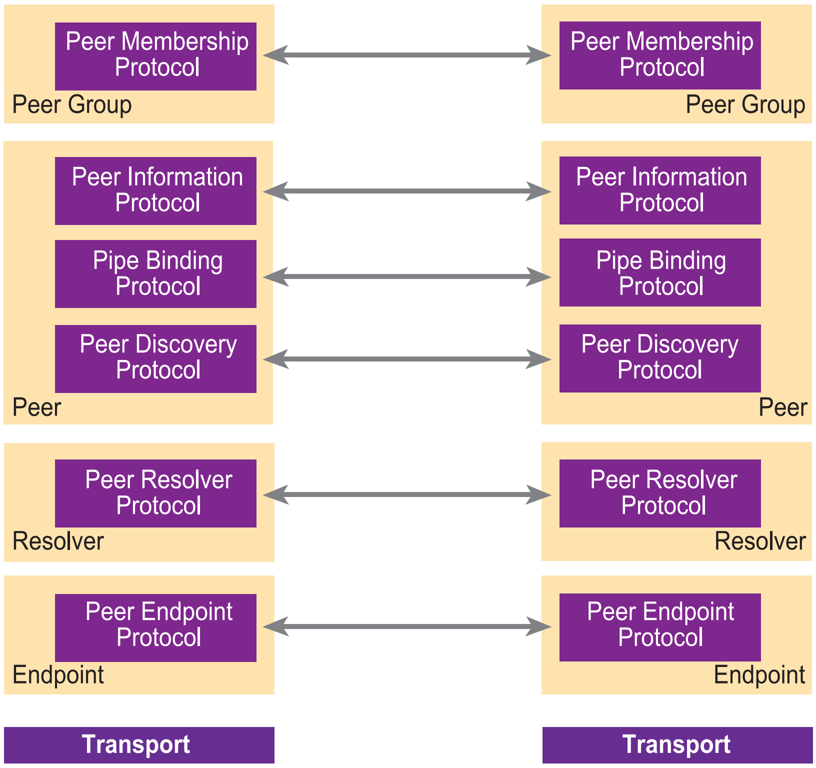
The JXTA Core, which provides support for P2P services and applications, has three main components:

- **Peer Group** consists of mechanisms for establishing, joining and leaving peer groups.
- **Peer Pipes** consists of mechanisms for establishing inter-peer communication channels.
- **Peer Monitoring** consists of mechanisms for handling access control, metering, and other surveillance functions.

The JXTA Services expand upon the capabilities of the core and also facilitate applications development. Facilities provided in this layer include mechanisms for searching, sharing, indexing, and caching code and content to enable cross-application bridging and translation of files. Finally, JXTA Applications are built using the services and core components.

Straddling the boundary between peer services and applications is the JXTA Shell, an application built on top of the JXTA Core that can be used both as a useful development environment for debugging and as a demonstration of some key features of the JXTA platform.

Initially, six protocols are defined to address the core P2P functions governing discovery, organisation, monitoring, and intra-peer communication (see *Figure 11*). Additional protocols can be included in the future to boost functionality. A common XML-based messaging layer binds the protocols to the appropriate underlying network transports.



Courtesy of Sun Microsystems

Figure 11. JXTA Protocols

.NET is Microsoft's platform for developing next generation software that connects the world of information, devices and people in a unified, personalised way. While heavily marketed as the platform for developing XML Web services, its infrastructure includes all the necessary tools to develop P2P applications. Presently, Microsoft Messenger XP is an example of a P2P application built on some of the P2P .NET bits underlying Windows XP. NET's messaging is built upon IETF standards, Instant Messaging and Presence Protocol (IMPP) and Session Initiation Protocol (SIP). IMPP provides simple IM, presence and notification. SIP lies at the heart of .NET's direct P2P signaling, messaging, telephony and events notification. Identity is provided by Passport.

Groove Networks was founded by Lotus Notes creator Ray Ozzie. The company's flagship product, Groove 1.0 is a groupware application that enables groups of collaborators to form in a decentralised ad-hoc fashion. In addition, there is Groove Development Kit, which is the basis of tool development for integration with Groove. With over US$117 million in venture capital and an experienced software team, Groove has created a full-feature groupware application that can be used as a development platform.

3.6 CHALLENGES

The future of P2P is promising but not without obstacles. In this section, we look at the most likely challenges that will dampen the adoption of P2P applications.

3.6.1 Network Administrator and Internet Service Providers (ISP) Resistance

The biggest obstacle to the adoption of P2P comes from the network administrator and ISP due to the lack of control of network resources. For example, the popularity of Napster as a file-sharing has led to huge increase in network bandwidth usage, forcing many US university administrators to block it. ISPs also prohibit home users from running "servers or server like devices", and running a P2P application, like Napster, essentially transforms a computer into a server when it uploads music files to its peers. In addition, many network administrators tend to view the P2P networking environment as the "perfect" propagation ground for virus, further enhancing P2P's negative image.

3.6.2 Cautious Investment Climate

The burst of the "dot.com" bubble and downturn in the economy has greatly reduced the amount of funding available for P2P. According to Aberdeen Group, P2P technology industries have gathered approximately US$380 million in funding through end of 2000, which is small in comparison with the amount lavished on previous trends. On a more positive note, most of the current well-funded P2P companies have a solid business plan and show greater promise to succeed unlike earlier "dot.com" start-ups. However, this does not mean that all of them will succeed.

3.6.3 Significant Value Proposition required

With the exception of distributed computing, most P2P applications have yet to clearly demonstrate their superiority over their more established client-server counterparts. For example, collaborative groupware products like Groove still needs to demonstrate its effectiveness over the traditional and more popular data sharing methods, like e-mail attachments and database. For the most part, IT managers distrust resources not under their immediate control and need a high degree of impetus to rely on them. The value proposition of many P2P applications is still primarily based on price but in a business setting, other factors like the ease of maintenance, reliability and scalability may affect its adoption.

3.6.4 Standards and Interoperability

As discussed in Section 3.5, there is a lack of standards adoption for P2P companies and this has significantly reduced the developmental progress of P2P technology. However, we believe one or more of the following scenarios will occur within the next 2 to 3 years:

- P2P working groups agreed on adopting certain best practices as standards, and/or
- Adoption of Web services standards, and/or
- Adoption of Sun's JXTA or Microsoft's .NET development platforms, and/or
- Emergence of de-facto standards due to market consolidation.

4 Grid Computing

4.1 OVERVIEW

4.1.1 Definition

The Grid is an emerging hardware and software infrastructure that serves as a distributed, robust, service oriented and secure platform for access to high-end computational capabilities. The Grid pools large-scale, heterogeneous resources from geographically diverse locations into an ensemble and presents them as an integrated and single metasystem. These metasystems created may have different OSs, platforms, architectures and data formats. Applications are run on top of this layer of resources that form the metasystems.

At the physical layer, a metasystem is a pool of resources such as desktop PCs, high performance supercomputers, storage, instruments and databases connected by a high-bandwidth network to serve as a virtual computer. This metasystem is differentiated from a simple network of computers by a software layer, often called middleware, which converts a pool of independent resources into a virtual machine. This virtual machine should be as simple to use as the machine on the user's desktop, differences should be transparent, and should allow easy collaboration between people located anywhere in the world.

The Grid is the sum of networking, computing, and data storage technologies needed to create this seamless, balanced, integrated computational and collaborative environment.

4.1.2 Type of Grid Communities

A variety of Grids will be created to serve different groups of people and varying types of requirements, incentives, sharing policies and degrees of geographical dispersion. This section highlights the cross-industry utilisation of Grid, and its potential to be adopted on a grand global scale, from industries and businesses to non-profit oriented organisations such as governments and research institutions.

At a simplified level, there are four types of Grid Communities, namely the National Grids, Private or Enterprise Grids, Virtual Grids and Public Grids.

National Grids

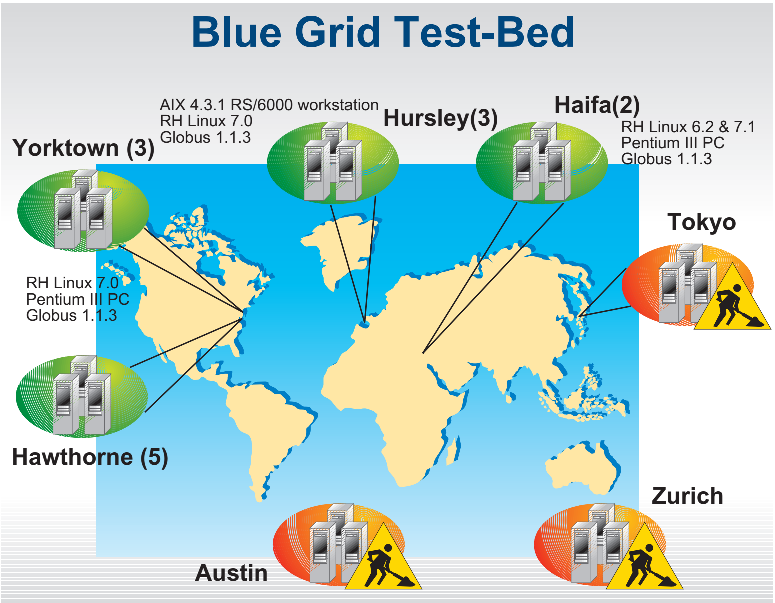
National Grids are built to enable collaboration on a national level and exist primarily within a country boundary. National Grids are funded by government agencies. The community for National Grids consists of official, planners and scientists concerned with problems and usage of Grids for government purposes.

One example is the **UK National Grid,** which is being built to link up 9 sites across UK. The UK National Grid is classified as a national Grid because the main sites for this Grid are within UK's country boundaries, the users are the UK research community, and the funding is from the UK government. The primary purpose of this Grid is to provide computing capabilities for the nation's researchers and to encourage collaboration.

Private Grids

Private or Enterprise Grids are built for the private use of an organisation. A Grid is considered private if a single organisation (or group of partner organisations) funds it, owns it, builds it, uses it, and the physical sites of that Grid are spread across that organisation's various localities. Private Grids are used by staff within an organisation or a group of organisations, with a common purpose and functionality to share resources. Examples of Private Grids are enterprise Grids, organisational Grids, Grid Intranets. It may include inter-enterprise groupings such as supply, purchasing and selling chains.

One example is the **Blue Grid** that IBM is building as an internal Grid linking IBM R&D labs in US, Israel, Switzerland and Japan. The Blue Grid Project is IBM's internal testbed, and they will be studying and developing Grid protocols and services including Globus.



Courtesy of IBM

Figure 12. Blue Grid Test Bed

Virtual Grids

Virtual Grids are created for a group of people who have similar interests. This is unlike the private Grid, which is considered as Grid version of Intranet. The virtual Grid is used by people with a common profession, need or interest, such as scientists and researchers in a particular field of study but from geographically dispersed locations to share resources and collaborate.

One example is the **European Union (EU) DataGrid** currently being constructed in European Organization for Nuclear Research (CERN). The EU DataGrid was initiated because of Large Hadron Collider (LHC) scientific project at CERN, which is expected to be operational in 2005. This LHC is expected to produce in excess of one petabyte of data a year, and the world-wide High Energy Physics community is planning to access this data for research. The primary sites and users are in Europe, and users from the US and Japan are among the others that are expanding the scope of this Grid.

Public Grids

Public Grids are being explored to support a market for on-demand computing. The market of selling computing resources on utility model does not exist yet. We see some early efforts in selling processing power by the hour and use of storage by the byte.

One example of a public commercial Grid project is the **IBM Grid Computing Initiative.** IBM is investing US$4 billion to build 50 new data centres to augment their existing 175 data centres around the world. This distributed collection of sites, when pooled together will power the IBM commercial Grid. This is classified as a public Grid as it is funded by a company with the intent of gaining customers and revenue.

4.2 EVOLUTION OF GRID COMPUTING

This section gives the reader an overview of the history circa 1970-1990 using a few selected examples in 10-year blocks, levering off with a more detailed look at the past 2 years. The main notable events outlined in 2000 and 2001 are meant to give the reader a view into the main trends. This view into the current state of Grid computing is further expounded in the following section.

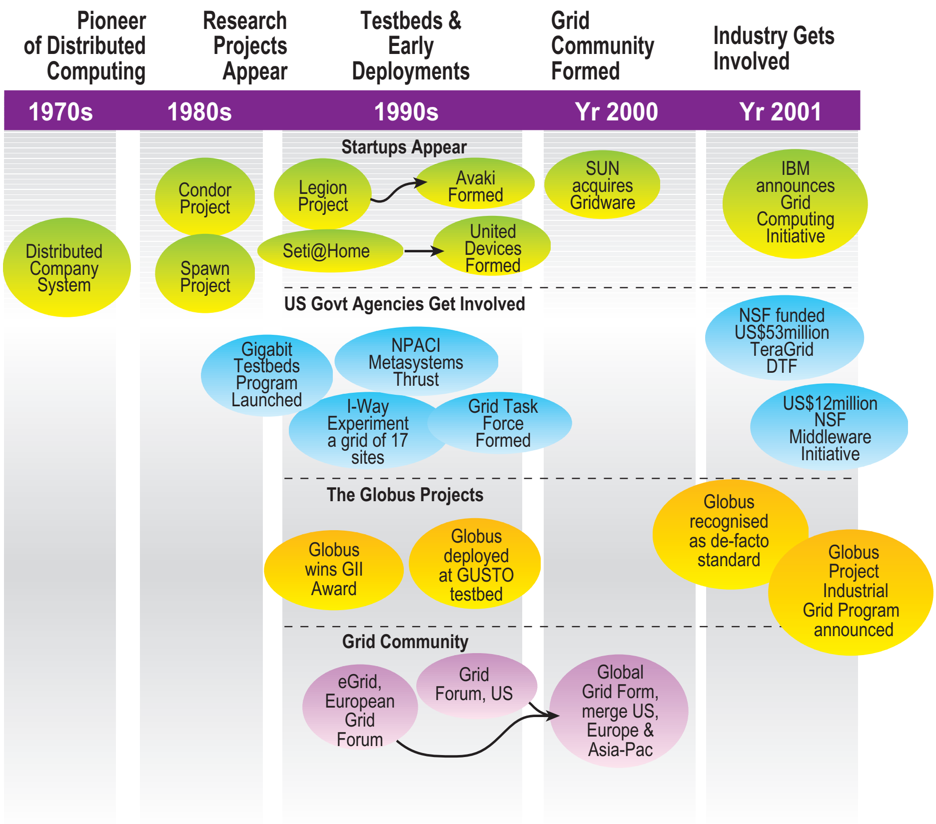


Figure 13. Evolution of Grid Computing

4.2.1 Notable Event in the 1970s

Grid Computing as a field of study has existed since the 1970s. The ***Distributed Computer System*** was a distributed timesharing system developed at the University of California, Irvine. The system consisted of multiple minicomputers coupled by a ring communication system.

This project is viewed as an early effort towards Distributed Computing, a type of computing in which different components and objects comprising an application can be located on different computers connected to a network.

4.2.2 Notable Event in the 1980s

The 1980s saw a multiple of *academia initiated distributed computing research projects* being started. One example is the Condor project which searches for spare processing cycles. It was proposed in 1985 at the University of Wisconsin-Madison, and started as a continuation of two research projects: Remote Unix and Distributed Resource Management. The Condor project was started to utilise the computing power that was lying unused during periods of inactivity. The Condor team developed a software system that puts inactive computers back to work. Condor scavenges for processing power that would otherwise be lost.

4.2.3 Notable Events in the 1990s

For Grid Computing, the 1990s are notable for four things:

Intellectual Capital moves from Lab to industry.

Firstly, intellectual capital from several research projects that started in the early 90s spun off into commercial projects in the late 90s. Two notable examples are the commercialisation of the Legion project into Avaki and the extension of the SETI@Home project into MetaProcessor of United Devices.

Government Funding gets serious.

Secondly, government funding agencies started to pay more attention to the concept of Grid Computing and Metacomputing. Some examples are:

- In the late 80's and early 90's, the Gigabit Testbeds Program[6], a collaborative effort involving both industry and the research communities, was established by Corporation for Networking Research Initiatives (CNRI) with funding from National Science Foundation (NSF), Defence Advance Research Projects Agency (DARPA) and industry. Six distinct testbeds collaborations were created.
- I-Way, a grid of 17 sites was created and demonstrated at the SuperComputing 95 Conference. The Metasystems Thrust Area was formed under the National Partnership for Advanced Computational Infrastructure (NPACI).

[6] Gigabit Application is the former label for Grid Computing. More information about Gigabit Applications may be found in a 1992 paper written by Charlie Catlett "In Search of Gigabit

Globus gains prominence.

Thirdly was the emerging prominence of the Globus project as a Grid architecture. In the 1990s, the Computational Grid concept was developed by Ian Foster and Carl Kesselman and applied to scientific and engineering computing. They co-created the Globus Project, which was deployed and tested on the GUSTO (Globus Ubiquitous Supercomputing Testbed Organisation) testbed. GUSTO, which spans the globe and is one of the largest computational environments constructed received international recognition.

Globus is used in several projects such as the National Aeronautics and Space Administration (NASA) Information Power Grid (IPG) and the Synthetic Force-Express (SF-Express) project. SF-Express provides distributed interactive simulation applications using Globus services to couple 13 supercomputers at 10 sites, and hence to achieve a record setting 100,000 vehicle simulation in early 1997. This goal was initially targeted for 2002.

There were other alternatives such as Legion, which since has spun off into the commercial venture Avaki, and the University of California, San Diego project AppLeS, now renamed as Grail.

Grid Community forms.

Fourthly, in the late 1990s, the Grid Computing Community came together to form the Grid Forum in the US and eGrid, the European Grid Forum.

4.2.4 Notable Events in the year 2000 & 2001

In the year 2000-2001, there are 5 main trends across the Grid Computing arena.

Collaboration in Grid Community increases.

Firstly, we see more partnerships and alliances forming.

- Grid Forum, a primarily US-based grassroots organisation of individuals developing, deploying, or using Grid technologies, combined with eGrid - the European Grid Forum, and leaders from Asia Pacific, to create Global Grid Forum (GGF).
- The EU DataGrid at CERN originally planned to use their own architecture and middleware to construct their Grid, announced that they would collaborate with the Globus team in US to use and build upon the Globus toolkit.

Commercial activity and Industry involvement intensifies.

Secondly, more commercial activity has emerged among incumbents such as IBM, SUN, Intel and Oracle. Besides participating in various Grid projects, some have acquired Grid computing start-up companies. Venture funding also begins for distributed computing.

Public awareness starts and Hype begins.

Thirdly, the year 2000 may be viewed as the year that the awareness about Grid Computing begins to move to the general public. Popular media such as Red Herring, London Times, New York Times, BBC and the Economist began covering Grid Computing. The very successful SETI@Home project that asked the public to donate spare cycles over the Internet, was another contributing factor.

Government internationally commits to Grids.

Fourth is more Government commitment to Grid Computing:

- The US government is funding several projects in the Grid Computing field, which include the TeraGrid DTF and the NSF Middleware Project.
- The UK is building the UK National Grid and also launched the UK Grid Support Centre, a helpdesk and support centre for people who are developing Grid applications or deploying Grids in UK.
- Japan is active in Grid Computing and is regarded as the leader for the Asia-Pacific Grid (APGrid). Korea initiated a Korean Grid Forum, and announced plans to launch a national Grid.
- Europe has several Grid Computing activities happening now. Netherlands is building a Dutch Grid. Germany's "Uniform Interface to Computing Resources" (UNICORE) project is developing technologies to establish a computational Grid between the German High Performance Computing (HPC) Centres. France allocated 15 million French Francs for the French Grid Initiative. The Joint Committee of the Nordic Natural Science Research Councils has decided to set up a working group to study requirements for a Nordic Data Grid Centre.

Globus – The de-facto standard for Grid Computing.

The fifth trend discernible is the growing support for Globus. Globus is now regarded as the de-facto standard for Grid construction.

- In Nov 2001, the Globus Project Industrial Grid Program[7] was launched.
- Many commercial companies have adopted Globus Toolkit and plan to develop an optimised form of it for their platforms.
- Like RedHat that provides a commercial version of Linux, Platform Computing announced that it would collaborate with Globus to provide a commercially supported version of the Globus Toolkit.

4.3 FACTORS DRIVING DEMAND

4.3.1 A PetaFLOP by 2010

One of the reasons why US is building Supercomputing Grids is to reach petaflop computing power by 2010, a direction set by funding agencies for the HPC community.

There are mainly three ways in which the HPC community is working towards achieving this goal -- through shared memory architecture, multi-processor machines, and clusters. Grid Computing technology is used to scale these clusters up to high processing power and to manage it. This is driving Grid application areas in Distributed Supercomputing and High Throughput Computing (HTC).

[7] The Globus Project "Industrial Grid Program", is a program launched in 2001 to increase commercial support for the Globus toolkit.

4.3.2 To Solve Grand Challenge Problems

There exist several grand challenge problems that call for advances in computing infrastructure and capabilities. Grids play a certain role in this, particularly in problem areas that are too big for the biggest supercomputer to handle.

Selected examples of grand challenges are given below.

- Protein Folding. Protein Folding is regarded as the next frontier in Biotechnology. Researchers that aim to simulate the process of protein find that it is beyond the current limits of technology.
- The LHC. The LHC experiment at CERN that is expected to start generating data in 2005 is one of the reasons for the building of the EU DataGrid. The LHC is a particle collider for High Energy Physics experiments that is expected to produce 6-8 petabytes of data a year, and will require massive computing power to analyse and process this data. More information is included in the following section on Distributed Supercomputing Grids.

4.3.3 Driven by Government Funding and Academia Research Efforts

A healthy amount of funding is being poured into Grid Computing. From the US, we see US$12 million for a Grids Centre, US$57 million for Scientific Discovery through Advanced Computing (SciDAC) program, US$53 million for the TeraGrid deployment and The U.S. Department of Energy's investment of US$30 million for "national collaboratories". In Europe, £120m for the UK National Grid has been awarded.

4.3.4 Scavenging idle cycles & Utilisation of Edge Resources

With the increasing power of workstation and office productivity, combined with the current deployment of one machine per user, we see an increase in spare cycles available for utilisation. This is especially interesting, as the utilisation of idle cycles may be viewed as a scavenging method of cost savings, vs. capital expenditure in processing power. This area of grid computing is utilised to save costs. This is driving the Grid application area in High Throughput Computing.

4.3.5 Episodic Requirements and On Demand Computing

Grids are being built as an external source of computing resources for episodic requirements. This is both driven by and will facilitate a rise of event driven tasks, one off needs and periodic computing. Organisations will not have to build to over-capacity as a buffer for sudden surges in needs, they will be able to connect to a external provider for augmentation. One example is IBM Blue Grid which is envisioned as a computational service centre, to provide a source of computing power that may be accessed on demand. The platform aims to be dependable, reliable, robust, and secure, with metered usage and access to this grid is being explored either through a portal or transparently through their OS. This is driving Grid application areas in On-Demand Computing.

4.4 GRID COMPUTING APPLICATIONS

There are five main types of Grid Computing applications: Distributed Supercomputing, High Throughput Computing, Data Intensive Computing, On Demand Computing and Advanced Collaborative Environments (ACEs). The following section gives a definition to each application category and discusses the trends and development in each area.

4.4.1 Distributed Supercomputing

DEFINITION. Distributed Supercomputing is *for problems that are too big* for a (given) single computer to solve, so the Grid aggregates processing power from a collection of providers. The ensemble of providers ranges from all the supercomputers in a country to all the workstations in a company. This computing power may be from supercomputers or clusters.

There are several Supercomputing Grids being built to provide HPC power. Examples of these are:

- The TeraGrid Distributed Terascale Facility (more detail on this project below)
- The UK National Grid
- The Dutch Grid
- The Korean National Grid

EXAMPLE. The TeraGrid Distributed Terascale Facility (DTF).

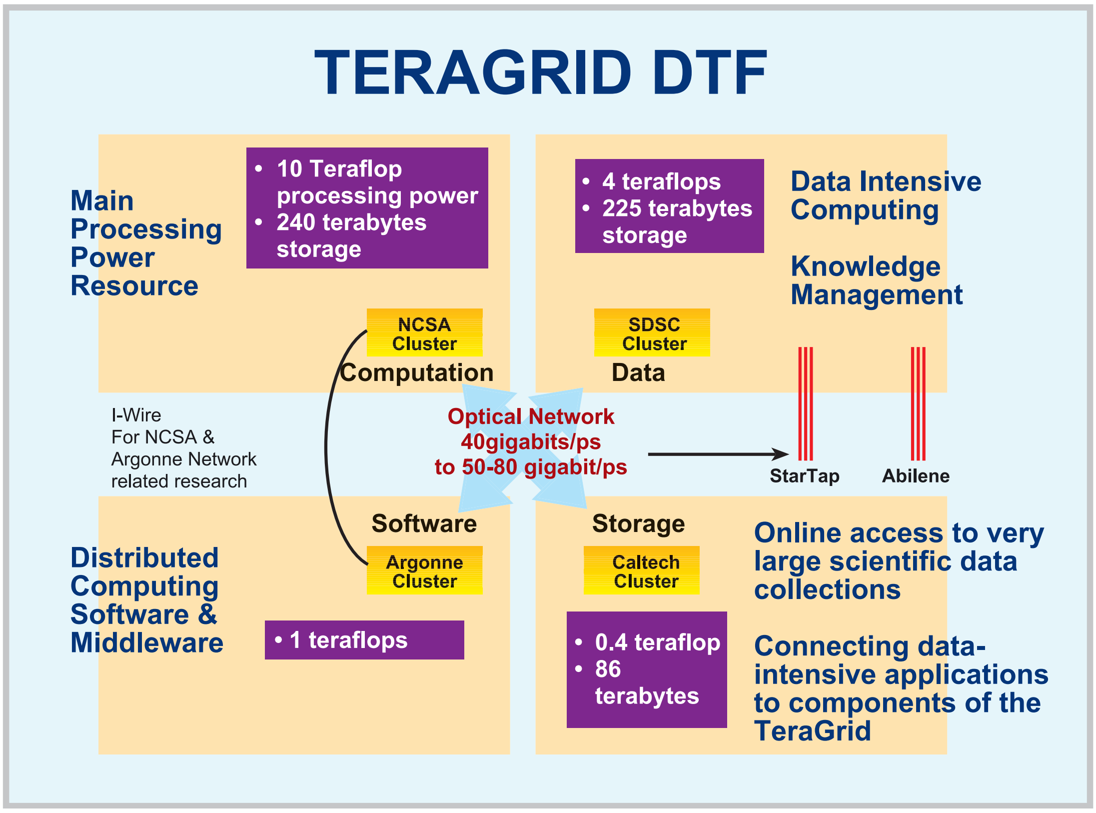


Figure 14. The TeraGrid Distributed Terascale Facility distributed in four sites

About The TeraGrid DTF Partners

The TeraGrid DTF project is a collaboration between the four research institutions listed below. The four partners are also the sites where the resources of the TeraGrid will be distributed.

- National Centre for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign
- The San Diego Supercomputer Centre (SDSC) at the University of California at San Diego
- Argonne National Laboratory (Argonne)
- California Institute of Technology (Caltech)

TeraGrid DTF Resources

Each site will be both a self-sufficient centre capable of operating on its own and a component of the TeraGrid. Together, these four clusters will form a single, balanced TeraGrid that supports both traditional supercomputing and a growing number of Grid-intensive applications with 13.6 Teraflops of computing power.

Cluster main source of processing power - The DTF will consist primarily of clustered IBM servers based on Intel Itanium processor interconnected with Myricom's Myrinet[8].

Network Infrastructure - The clusters will operate as a single distributed facility, linked via a dedicated optical network that will initially operate at 40Gbps and later upgrade to 50-80Gbps.

It will be connected to four other networks:

- Abilene, a network that links more than 180 research institutions across US.
- STAR TAP, an interconnect point in Chicago that provides access to and from international research networks.
- Consortium for Education Network Initiatives in California's California Research and Education Network-2, an advanced high-speed network that connects institutions in California.
- I-WIRE, an optical network that provides Argonne and NCSA additional bandwidth for related network research initiatives.

The Dispersion of Resources between the Four Partners

- NCSA will be the main source of processing power, with 10 teraflops of processing power and 240 terabytes of storage.
- SDSC will have 4 teraflops of processing power and 225 terabytes storage. SDSC will lead the TeraGrid data and knowledge management effort.
- Argonne will have 1 teraflop of processing power and will mainly be the centre for advanced distributed computing software, high-resolution rendering and remote visualisation capabilities.
- Caltech will have 0.4 teraflops of processing power and 86 terabytes of storage. Caltech will focus on providing online access to very large scientific data collections.

[8] Myrinet is a packet communication and switching technology that is widely used to interconnect clusters of workstations, PCs, servers, or single-board computers.

KEY OBSERVATIONS

- Support for Open Source. The TeraGrid and the UK National Grid are two high-profile Grids that are using open source Globus and Linux. The trend for increasing support for open source is expected to gain in strength.

- Grid Projects to watch. One of the reasons why Grid Computing is gaining attention is because of the high profile Grid deployment such as the TeraGrid DTF, the UK National Grid and the EU DataGrid. From the timeline outlined below, we can see that these Grid projects are still in planning and implementation phase and will only be operational in 2003-2005. These projects should be monitored to determine the usage trends and acceptance, application requirements and research work being done as they will be influential technology direction setters.

- Don't wait till these Grids are launched before taking action to pursue Grid Computing. Forecasts from analysts such as Bear and Stearns, Deutsche Bank and IBM themselves, predict that the Grid will start to gain momentum in 2004-2005. Irving Wladawsky-Berger, VP Technology and Strategy of IBM Server Group, predicts that Grid Computing will reach the commercial world in 2003-2004 or sooner. The local research institutions and industry should take action in this 2-3 year window.

Table 1. Key Milestones in Major Grid projects

Overview of Grid Milestones						
	2000	**2001**	**2002**	**2003**	**2004**	**2005 <**
TeraGrid		Fund Awarded	Base set of services operational IBM to deploy clusters in fall 2002	Testing Fall 2003, Grid operational		
EU DataGrid	Funds Awarded	Sept, Final Approval from CERN Project Start	Phase 1 starts Prototype			Phase 2 Production Version LHC start generating data in 2005
UK National Grid		Funds Awarded		Grid Operational		

4.4.1 High Throughput Computing

DEFINITION. In this application area, the main objective is to search a network of workstations and machines for idle capacity (spare cycles), and pool this processing capability into a unified source for utilisation. The pools of machines that provide the idle cycles are distributed across a network. This network may be a closed office LAN, an inter-enterprise collaborative network, or the public Internet.

The HTC resource management system is constantly on the look out for idle cycles, and adopts the mentality of a scavenger. The current heterogeneous computing landscape adds to the complexity of the resource management system.

EXAMPLE. The Legion Project and its commercial Spin-off, Avaki Corp

The Legion Project was initiated at the University of Virginia in September 1993. It is a metasystem or Grid OS, the software infrastructure that creates a single, coherent, virtual machine from a system of heterogeneous, geographically distributed, high performance machines.

Summary of activities:

- The Legion project has since been spun off into a commercial project. Applied Metacomputing was formed in 1998 to commercialise Legion, and was renamed to Avaki in 2001.
- Avaki released its flagship product Avaki 2.0 which extends the capability from computational grid to data grid across a network of machines.

Avaki provides services for application support, system management services for managing, monitoring and control this pool of resources. It also includes core services to transform this collection of resources into a single virtual machine, as well as protocol adapters to work across heterogeneous environments.

Why is HTC also called P2P? As discussed in the P2P Applications section, resource provisioning is an emerging field in P2P. HTC is a form of resource provisioning in CPU time where resources are scavenged predominately from edge resources such as PCs, hence the concept of P2P. Although the main contributors are PCs and workstations, other higher end machines are also utilised. Legion is operating a similar network for NPACI. The NPACI-net consists of heterogeneous machines with various capabilities and Oss such as IBM SP-3s, HP V2500s, Sun HPC 10000s, Compaq Alpha Clusters, as well as desktop based Intel PCs.

KEY OBSERVATIONS

Business Model not proven yet. The business model of getting the public to contribute spare cycles, aggregating and then selling CPU time from this pool is still not proven yet. The emergent such as United Devices, Avaki, Entropia, who launched their companies with platforms that pooled CPU time from the public internet, are starting to release enterprise version of their products. This is a move from selling CPU time to selling software solutions.

Academia and research organisations, and heavy processing cycles users such as the biotech community and finance and electronic design automation users will be the pioneer adopters of HTC. Adoption will increase as applications are ported over to this grid platform.

Still immature. Secondly, these public-cycles solutions are single use. For example, when the public downloads a SETI screensaver or downloads a "Search For Cancer Cure" client to contribute cycles, these cycles are used to solve one problem. These clients will need to evolve to be more generic and multi-purpose and have higher security. And in view of the possible potential, the current systems are primitive. The technology is still evolving, and the next generation is expected to be more sophisticated.

Java is expected to be the dominant client and job submission mode that will use its cross platform interoperability and applet sandbox technology for security and portability.

Billion dollar market size predicted. Analysts expect the model of utilising computing power from the public Internet to be increasingly important. United Devices announced in Oct 2001 that their aggregated pool of cycles have reached a "sustained operation" of 65 teraflops. Compare this to the US$53 million NSF funded TeraGrid DTF, which is projected to have 13.6 teraflops. When the technology is strengthened and the business models mature, this may prove to be a highly disruptive and potent technology. The utilisation of spare cycles is expected to gain prominence as a source of processing power. Bear and Stearns forecasts that the distributed processing market, a similar market space as HTC Grid Computing, is expected to be worth US$8 billion in 2004.

4.4.3 Data Intensive Computing

DEFINITION. The Remote Data Access Working Group (RDAWG) within the Global Grid Forum defines data Grids as "inherently distributed systems that tie together data and compute resources."

The typical profile for a Data Intensive applications is petabytes of data that require teraflops of computational power for processing, over geographically distributed sites and having thousands of users requiring access. High-speed networks are typically used to connect these sites and transport data between collection, storage and processing sites.

The major difference between a Data Grid and a computational Grid is the specialised infrastructure provided to applications for storage management and data access. In a computational grid the applications implement their own storage management schemes rather than using Grid provided services.

Data Grid Applications. There are two kinds of Data Intensive applications; ***Distributed Processing*** where by the application is moved to the site where the data is stored, and ***Distributed Caching,*** where the data is moved to the supercomputing facility for processing.

Data-Intensive applications have two main goals. Firstly, to gain access to geographically distributed data sources, of particular interest to the RDAWG are systems that provide special capabilities for providing access to large (terabyte-sized) remote data sets. Secondly is to synthesise new information that is maintained in geographically distributed repositories, digital libraries and databases. These are the software systems and environments that provide mechanisms for accessing distributed and heterogeneous data resources. These software systems are often used to assemble Data Grids or to provide essential Grid functionality to data intensive Grid applications.

EXAMPLE. NPACI's Bioinformatics Infrastructure for Large-Scale Analysis

The NPACI is building an infrastructure for large-scale analysis of biomolecular data. The Large-Scale Analyses project uses data manipulation, analysis, and visualisation to integrate data from distributed molecular structure databases. The Data Grid consists of a federation of databases including the Protein Data Bank (PDB), GenBank, and the Molecular Dynamics Trajectory Database.

- PDB – is a worldwide repository for the processing and distribution of 3-D biological macromolecular structure data. The Protein Data Bank is operated by the Research Collaboratory for Structural Bioinformatics (RCSB) and is funded by the NSF, the Department of Energy, the National Institute of General Medical Sciences and the National Library of Medicine.
- Genbank - is the National Institutes of Health genetic sequence database, an annotated collection of all publicly available DNA sequences. GenBank is part of the International Nucleotide Sequence Database Collaboration, which is comprised of the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at the National Centre for Biotechnology Information.
- Molecular Dynamics Trajectory Database – this is lead by the University of Houston, working on metadata and tools for comparisons cluster analysis, global optimisation, and other operations in their databases.

Each database connects to the Data Grid infrastructure to enable uniform access and authorisation mechanisms to the various collections. Legion is involved as an academic partner, and it is assumed that the technology, if proven successful, will be translated to commercial products.

KEY OBSERVATIONS

Globus as Data Grid building block. The Globus team released version 2 of their Globus Toolkit, which includes facilities for constructing Data Grids. Globus is expected to emerge as an important feature in building Data Grids, and competing technologies and building kits will develop integration bridges to interoperate with Globus.

From vendor specific to a common standard. The current landscape in data technology such as database management systems, meta-models, query tools, analysis and information discovery facilities, are vendor and domain specific. We expect the launch of various Grid projects will spur collaboration in these technologies, which is likely to lead to common meta-models, standards and a higher level of interoperability between database technologies and products.

Possible Data Grid. A suitable project would be the human genome databases. There are several different groups mapping the human genome, called Single Nucleotide Polymorphisms (SNPs). The human genome contains an estimated 10 million SNPs and only less than half are detected. It is distributed over several different databases as several parties such as the SNP Consortium and Celera Genomics are mapping it. Researchers in the biotech community require access to the human genome databases. A Data Grid infrastructure that connects these different databases will increase ease of use and improve the quality of analysis.

4.4.4 On Demand Computing

DEFINITION. On demand applications use the Grid to gain access to resources for short-term requirements. These resources cannot be cost effectively or conveniently located locally. These resources may be computation power, software, data repositories, or specialised sensors. This is the effort that is pushing towards utility-like or Grid-like computing where users experience reliable, consistent, instant and inexpensive access to resources around the globe.

While Distributed Supercomputing's concern is the need for more power and higher performance, and High-Throughput is use of idle resources, On-Demand Computing's main concern is cost effectiveness for applications that require processing that is dynamic and short term. The main emphasis for On-demand computing is Costs-Performance concerns, and having "instant" and easily accessed computing power at uncertain times and for uncertain periods when need arises. On Demand Computing defines the way the resources offered on the Grid is used and accesses. The actual resources that are accessed may be HPC time, HTC time, instruments, data sources, or collaboration facilities.

EXAMPLE. IBM E-Sourcing

IBM has launched a project, the Grid Computing Initiative, to provide any desktop with on-demand computing access via the desktop. IBM is investing US$4 billion to build 50 new data centres to augment their existing 175 data centres around the world.

This aggregation of data centres, when pooled together will power the IBM commercial Grid and offer its customers utility-like computing over the Internet, which IBM is marketing as e-Sourcing.

To manage this complex environment, IBM has embarked on the Autonomic Computing program, in which e-Liza is one of its projects. They will be exploring technical issues such as system identity, self-healing, self-managing and self-protection technologies, as well as optimisation, environment awareness and context awareness.

KEY OBSERVATIONS

The PC and Browser will not be the only access client into the Grid. Current research activities for client access into Grids are focusing on the PC/Browser combination. A trend is emerging to bring the access beyond PCs. For example, when users send processing jobs to the Grid, he can check on the status of the job (how much processing has been done, expected completion time, any anomalies or errors.) using an internet-connected device. For processing jobs that takes weeks, having easy access to system information frees the worker from being chained to the desk and offers greater mobility.

Access to Grid will shift to the OS. Technology for accessing Grid functionality will be included as a feature in OS. For example, Microsoft has announced that they have contributed funds to the Globus project to explore ways in which XP may be used as the client access platform into potential Grids being built. Sun Microsystems has also indicated in a whitepaper[9] that their "central Grid Engine will disappear, partly, as an

[9] Wolfgang, Gentzsch, Sun Microsystems, Inc., DOT-COMing the Grid: Using Grids for Business http://www.sun.com/software/gridware/article.html

integrated component into the local OS, and partly being replaced by intelligent mobile agents, which enable a P2P, and self-healing environment with potentially infinite compute power available on-demand."

Leveraging Web Services technologies in Grid computing. The Grid community's interest in Web services technology is expected to increase. SDSC is exploring the intersection of the commercial Web services and scientific-oriented and engineering-oriented Grid services. They are interested in using Web services to construct access portals and application portal into the back end of Grid resources. A major focus for the next version of the Globus Toolkit (in particular Globus Resource Allocation Manager-2) is the use of SOAP and WSDL and the proposed Web services Workflow/choreography services. Globus is exploring J2EE and .Net now. UDDI Web services registries could also list Grid services.

Need for Universal Interface – Portals as the first step. Universal accessibility of On-Demand Computing requires a universal user interface to provide a common way of interacting with the on-demand systems across the different providers and users. Portals are the first step towards such an interface. A selection of research projects in Grid portal are listed below:

- The Grid Portal Toolkit (GridPort): The NPACI's toolkit enables application developers to build Web-based portals to the computational grid. It hides the complexity of managing grid resources to allow users to concentrate on the scientific components of the problem and build new computational portals rapidly.
- PUNCH: An on-going research project from Purdue University that turns the www into a distributed computing portal. It serves as a computing platform to manage and broker resources among end-users, application services providers, storage warehouses and CPU farms.
- Open Portal Interface Environment (OPIE): The NCSA's OPIE project is an advanced web interface environment that can be used to combine many external, and internal web-accessible data sources for user interaction. OPIE also provides environment customisation by providing a web data interaction system that allows individual users to specify the data they want to see and the way they want to view it.
- Science Portal Project: This is a NCSA/NSF funded project to build a problem-solving environment for Grid Computing.
- In the industry, Sun released the Technical Compute Portal, a partnership between SUN and iPlanet, which provides a browser interface into its clustered processing power.

Decision Making Aid – Faster processing at higher cost or cheaper but slower execution. Cost Performance trade-off decisions in on-demand computing imply that the computing system be able to decide how and where to execute a process. Intelligence may be incorporated into the system by the use of agent technology or other forms of artificial intelligence or rule based systems.

4.4.5 Advanced Collaborative Environments

DEFINITION. Collaborative applications are concerned with enabling and enhancing human-human interactions and enable shared use of resources. Grid Collaboration application usually involves need for big computational power and multi-user collaborative environments.

In this application category, Grid infrastructure is used to pool together humans over geographical distances in a realistic or enhanced environment. The infrastructure that the users sit on includes hardware and software that specialise in either realism or virtual reality. In realism, high processing capacity, sensors and collaborative tools are used to emulate a user environment as realistically as possible and supporting as wide a range of human gestures as possible. In enhanced environment, the Grid infrastructure simulates an environment that is not available in real life.

These ACEs take the level of human-human interaction beyond the normal levels of collaboration. For example, ACEs may be utilised to increase employee productivity working on a large project. ACE is being explored as an aid in engineering product design. The rationale for using ACE is that increasing the workforce to double does not halve the time taken to do the job, as the larger the number of employees, the higher the level of inefficiency and harder it is to work together smoothly.

EXAMPLE: IHPC ACE.

The Institute of High Performance Computing (iHPC) in Singapore has worked in the following ACE projects:

- Advanced visualisation using leading-edge Virtual Reality (VR) technology which enables life-like interaction with computer-based models. The MIRAGE (Modelling and Immersion in Real-time Advanced Graphics Environment) is a state-of-the-art facility that brings together HPC and advanced visualisation for research into areas involving large 3D data. Some of the visualisation tools used by the team include the semi-immersive ImmersaDesk™, and the fully-immersive CAVE™ (CAVE Automatic Virtual Environment).
- Real-time interaction and data manipulation, a real-time VR environment which enables a high level of interaction between users and the research data. An interactive and immersive exploration of data maximises the potential for insight that is otherwise not possible.
- Long-distance high-bandwidth virtual reality collaborations for international collaborations to investigate long-haul VR-to-VR communications. The aim is to use high-speed ATM-based networks to interconnect industrial and research institutions equipped with CAVEs™, ImmersaDesks™, and HPC resources by high-speed networks to support collaboration in design, training, scientific visualisation, and computational steering.

These capabilities were used in projects such as:

- A Jurong Town Corporation project to study the airflow within a proposed multi-storey factory complex, which involved the visualisation of computational fluid dynamics (CFD). The simulation results were imported into the CAVE™ for visualisation.
- A CFD simulation of the proposed Changi MRT Station was evaluated in the LTA project. The immersive visualisation of the complex airflow and temperature data was

uncluttered and intuitive, leading to an excellent understanding of the 3D information.

- A Motorola drop-test simulation where CAVE™ was used to evaluate the results of drop-test simulations of Motorola pagers. The results were presented in 3D animation. In addition to controlling the playback, the user was also able to digitally "slice" through the pager to examine the interactions between the interior components as the pager fell to the ground.

KEY OBSERVATIONS

Some ACEs are nearer term. Some are futuristic. As outlined above, there are many types of ACEs. The ones that are desktop and PC based are considered nearer term, while those involving teleimmersion and its accompanying VR helmet and CAVE sets ups are considered a niche area. For desktop-based technologies, it is comparatively easier to access and use and join in an ACE community. For ACEs that require specialised hardware set-up, the uptake may be slower. ACEs that support the visualisation of data sets and analysis data for group access will be the early drivers in this category.

For excellent visualisation, learn from Gaming industry. The PC game market drives the computer graphics industry, making gaming based display technology a good way to achieve very high-resolution displays with near-term technology. NPACI partners at Argonne, Minnesota and Princeton, as well as NCSA and SDSC, have built many variations of display technology for scientific visualisation based on gaming displays. It is very likely that alliances and partnership for technology transfer will be set up, particularly to pursue gaming quality graphics at gaming machine prices for ACE visualisation capabilities.

ACE for integrating Supply Chains, product design parties and Alliance Partners are nearer term opportunity areas. In the nearer term, ACE systems for areas such as engineering product design, supply chains integration, and bringing strategic corporate partners together are emerging. Possible industry model for ACE would be Taiwan Semiconductor Manufacturing Company and United Microelectronics Corp. of Taiwan's Virtual Foundry. This online collaborative environment allows engineers from chip companies and silicon wafer plants to collaborate on design and production tasks. It will also have Integrated Circuit chips production status monitoring, and enable internet-based supply chain management for the semiconductor value chain.

4.5 SUMMARY OF GRID APPLICATIONS

Grid Applications are not stand alone, may be blended with each other into a fuller Grid.

In the preceding sections, a brief overview of Grid application classes is given. The figure below provides a Grid Typology based on the intersection of application classes and Grid user communities. Using the figure outlined below, we show that many of the flagship Grids actually encompass many applications. For example, TeraGrid DTF and EU DataGrid both include all the five Grid computing applications.

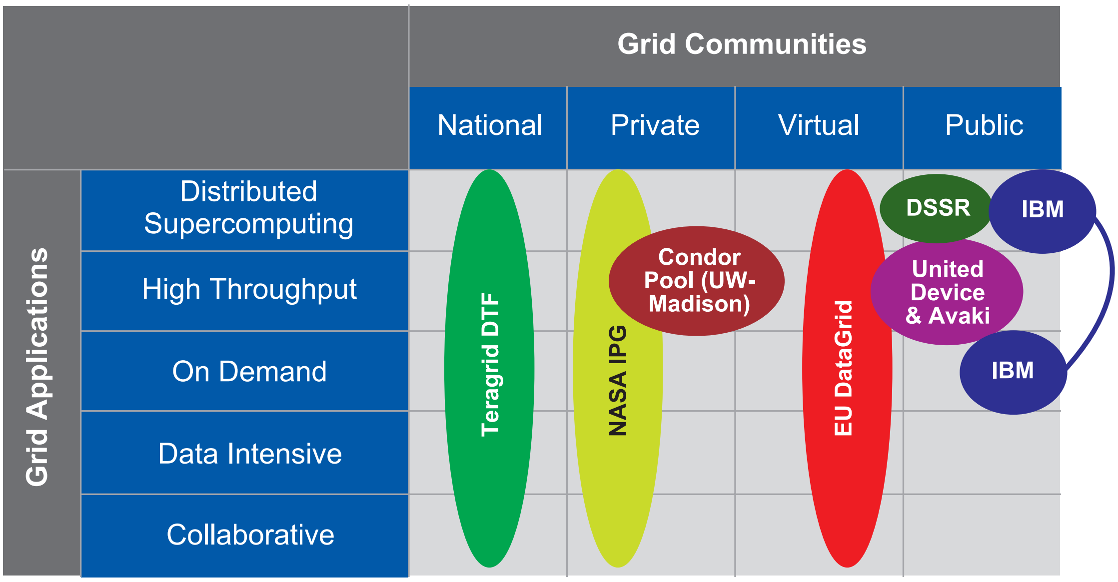


Figure 15. A Grid Typology using selected examples

A Look Ahead.

There are activities on several other Grid areas that are not highlighted here. Examples are sensor Grids, smart instrument or instrumentation Grids, agents Grids, e-learning and teleimmersion Grid for education, smart spaces composed of a Grid of displays.
Current Grids are working on pooling resources such as processing power, storage and databases to provide users with a unified source for access. Innovation and research is moving beyond processing power and storage. Other resources will emerge as Grid resources, such as sensors and MicroElectro Mechanical System, scientific instruments, and VR environments.

4.6 BUSINESS MODELS

Business Models are moulded by the product strategy, the competitive landscape, the value proposition to the customers, and in turns affects distribution channels, sales strategies and choice of strategic partners.

There are undercurrent trends in the industry towards commoditisation of technology, service-oriented and utility-like delivery. With the ever increasing shortening innovation cycles and short life spans of technology, customers are increasing being caught in the mire of being on the cutting edge of technology to increase their employee productivity and maintaining competitive advantage versus capital investments in technology that have shelf-lives of months.

Several points are important to consider:

Transition Pains - Firstly, efforts are being made towards using the Internet as a service delivery platform and as a sales channel. While offering the advantage of shorter sales cycles, lower cost of sales and new markets, the challenge of robust support structures for the Internet and new pricing methodologies have to be dealt with. Software companies undergoing a 180-degree shift from traditional product sales and license pricing structure to a service-fee pricing will experience initial difficulties.

Standards & Open Source lead to differentiation challenges - Secondly, marketplace acceptance towards standards and open source and wariness towards proprietary platforms leads to little opportunity for differentiation.

Services Revenue Pattern Apparent - Thirdly, the market is beginning to shift towards services as the revenue from this area gains weight.

4.6.1 Utility Model - The Various Variants

The utility model is a metered usage or "pay for what you use" approach. In Grid Computing, the utility model may be implemented by a straightforward charging for how much resources used or a more complex charging for software tools and services.

For **Computation Grids,** charging for the amount of processing power used may be measured by the processing hours or by the FLOPS (floating-point operations per second) that are utilised. One example of this is Debis Systemhaus Solutions for Research (DSSR). DSSR is a joint venture of DLR (German Aerospace Centre) and Debis Systemhaus. It offers its NEC SX-5 supercomputer power through an Internet e-commerce gateway. They charge 150 DM/US$90 for one CPU hour including advisory services. There are also emergents who are aggregating a pool of spare cycles and selling processing time on this unified source of computation. One example of this model is Parabon Computation, who sells its Grid processing time for US$6 per hour per 100 computers.

There is yet another model that charge for the processing jobs that you submit. Providers using this model may differentiate themselves from their competitors by having more powerful processing machines or more efficient processing algorithms. They are then able to process jobs faster and cheaper and increase the service level perceived by the utility customers.

For **Data Grids,** one model is the *Outsourced Databases* model where users do not want to manage their own data and invest in data mining tools. They send their data to a provider who charges for the amount of datab ase storage used, maintenance service, data backup and the use of discovery tools. In this model, when the user uses the information discovery tool to perform analysis on the data, the data is sent transparently to the provider back-end machines for processing.

Access to Proprietary Databases is another model for data Grids. Providers of such proprietary databases offer data access, use of discovery tools, mining tools, analysis tools and business logic algorithms. This access may be charged on a per-access usage or on a subscription basis. This is a model that Celera Genomics uses. Celera Genomics was the first to map the human genome, and they charge for access to their human genome databases and the use of their proprietary Bioinformatics tools. Analysis jobs are sent to Celera's back-end processing machines and the results are presented to the users. They are charging their pharmaceutical partners for subscriptions at the cost of US$5 million to US$15 million a year.

In a variant of the above Data Grid model, the provider uses a ***Aggregator*** business model where several databases (that usually do not belong to the provider) are merged into a data Grid and users are able to access that unified whole as a single source. An example of this is the NPACI's Bioinformatics Infrastructure for Large-Scale Analyses project. The project uses data manipulation, analysis, and visualisation to integrate data from distributed molecular structure databases, such as the Protein Data Bank, GenBank and the Molecular Dynamics Trajectory Database. They provide scanning and comparison algorithms for users to do analysis across all the federated databases.

4.6.2 Infrastructure Providers

Infrastructure providers are the building blocks providers, versus the domain specific commerce models. For example, building block providers sell the infrastructure layer that is used by data Grid builders in biotech or computational Grid builders for manufacturing. In this model, the provider sells the products that provide the underlying components and technologies for Grids. This is a mature business model, usually analogised to the gold-rush model of selling the picks and shovels versus the gold miners.

There are plenty of opportunities in Grid Computing for infrastructure providers. The incumbents appear to have a lead in this area; IBM and Sun provide high performance machines and storage resources; Intel's Itanium is predicted to be the dominant chip powering the clusters; Oracle being the dominant player in databases and traditional network vendors providing the networking infrastructure.

The middleware, which is the key layer that provides the Grid-like behaviour to Grids sees more activity for emergents. Such as Platform Computing, who plan to commercialise the open-source Globus, the way Red Hat is to Linux.

4.7 FUTURE TRENDS AND DEVELOPMENT

4.7.1 Fractal Growth – From smaller specialised communities to an open worldwide community.

The interest is in a Grid that can offer utility-like access to its resources to the user. Currently, Grids are being built for specialised and closed communities such as Biotechnology in Singapore, High Energy Physics in Europe, and research community in UK. Intel deployed a Grid to pool processing power for chip design and IBM is constructing Blue Grid for their internal research centres. The projects that pool spare cycles from the public Internet mainly serve a specific user such as the SETI Project and the Search for Cancer Cure Project.

General open Grids are expected to emerge in 2004. Models for this general Grid to watch are the TeraGrid DTF, the UK National Grid and IBM's Grid Computing initiative. The TeraGrid DTF's resources are open to the US research community, how they handle access to the Grid, perform resource accounting and ensure robustness, security and consistency. The UK National plans to offer its resources firstly to the a closed research community, then radiate out to selected industry partners and finally to the open public.

The users of current Grid projects are mainly domain specific. Grid Computing will not reach a global pervasive scale till an application area that crosses vertical boundaries or is more generic emerges.

4.7.2 Access to Grid

In the use of Grid, there is a simple division between the resource providers and the resource users. How the resource user gets access to the resource provider is through three ways. Portals will appear first, and OS last. A key factor for mass-market adoption requires the seamless transparency that OS level access offers.

Access through a Portal

- First is through a portal, this method requires the user to manually submit jobs and access resources through a browser based portal. There are several research projects in this area, and SUN has already released a commercial version, the Technical Compute Portal from iPlanet.

Access through Grid-enabled Applications

- Second is through the application. In this access mode, the submission of processing tasks is handled by the application and the user is shielded to a certain extent. This model of access is already implemented on applications ported to the Grid provider's platform. For example, United Devices and Avaki have biotech application ported over to their Grid platform.

Transparent Access through OS plug-ins

- Third is through OS layer. In this access mode, the OS handles the access of Grid resources transparently to the user. It is expected the OS will map these open global resources to the local system, hiding unnecessary details from the user. Today's dominant OS provider, Microsoft is investing in Globus, to develop a version that works on Windows. A press release at the SC2001 Conference, they announced that "Microsoft sees the potential to bring Grid Computing to millions as Windows XP becomes a portal into Globus networks."

- In that same press release "12 Companies Adopt Argonne Lab/USC Globus Toolkit as Standard Grid Technology Platform", eleven other firms, including IBM, Compaq, Cray, SGI, Sun Microsystems, Veridian, Fujitsu, Hitachi, NEC announced that they would adopt the Globus Toolkit and develop an optimised form of it for their platforms, as their preferred path toward secure, distributed, multi-vendor, Grid Computing.

4.7.3 Increasing Usability: Grid Computing for the Masses

Grid Computing is currently in the domain of the research community and the toolkits and software are in beta release. The tools have not reached the stage of being easy to use and user-friendly as to broach the mass market. However, we see several developments that point to the beginning of ease-of-use for Grid Computing tools and implementation packages.

Globus Toolkit getting more user friendly

The Globus project team released the Globus Toolkit 2.0, which features a new "packaging technology". This is important as the Globus Toolkit is used by different levels of the Grid Computing ecosystem, the same Globus toolkit aggregates a mass of features and technology for:

- Application framework developers who are using Globus services to build software frameworks that facilitate the development and execution of specific types of applications
- Application developers who use Globus services to construct innovative Grid-based applications, either directly or via Grid-enabled tools.
- Grid builders who are using Globus services to create production Grid Computing environments

Among the new features that this packaging technology provides are:

- "A la carte" installation (only the components that you want)
- Binary distributions for popular platforms
- More efficient configuration/compilation
- Install anywhere, move anytime

Platform & Globus = RedHat & Linux?

A second development that points to the Globus toolkits being commoditised is Platform's announcement that they are working with the Globus Project to commercialise the Globus Toolkit for governing the use of computers and storage systems joined into a large computing Grid. Platform drew an analogy, stating that they would position themselves to be to Globus, what Redhat is to Linux.

Easy to Use Grid Packages

A third development highlighted is the Alliance-In-A-Box initiative. The In-a-Box initiative created a layered, standard set of software packages. It consists of Cluster-in-a-Box, Grid-in-a-Box, Display Wall-in-a-Box and Access Grid-in-a-Box. These 4 packages are interoperable.

- Cluster-in-a-Box allows users to create clusters and by providing tools for monitoring jobs, measuring performance, and accounting.
- Grid-in-a-Box includes middleware for tasks such as authentication, job management, and information discovery to link Grid resources into a powerful, flexible problem-solving environment.
- The Alliance Display Wall-in-a-Box provides guidelines for building a tiled display wall, and software for displaying imagery on the wall such as movie playback, and visualisation tools that can be displayed on the tiled wall

- The Access Grid is an integrated environment that supports distributed meetings, remote visualisation, and distance education.

These packages will lower the cost and the expertise needed to utilise new technologies to support the needs of the national research community. By offering a simplified way to get and deploy these functionalities, use of these In-A-Box initiatives aim to help to expand the use of the Grid more rapidly.

4.7.4 Grid Computing Killer App

For the 3-5 years time frame, the killer application for Grid will be data Grids. The signals are clear:

Flood of Data. Firstly, the amount of data is growing. Biological databases needed by the biotechnology research community are growing at an exponential rate. The scientific community is predicted to generate petabytes of data for earth observation and High Energy Physics. In the e-sector, databases filled with Click Stream and locator data is being gathered. The rise of inter-organisation collaborations such as supply chain management, collaborative product design and selling chains will also contribute to this deluge of data.

Gap between Data Collection and Data Analysis. Secondly, there is currently a gap between our ability to collect and generate data and the ability to analyse, manipulate and sieve through these collected data. Knowledge discovery and data mining is a relatively new field and is currently mostly vendor specific.

Intelligence Wanted. One factor why data Grids and its analytical layer to synthesise new information is highlighted as a killer app is because it cuts across the verticals in the business computing and scientific research world.

- From the biotech community to the selling of books, intelligent analysis and aid in decision making is called for. Deciphering the masses of biological data leads to medical breakthroughs, which spells billion-dollar market opportunity for biotech and pharmaceuticals players.
- Weathermen in the US serving as private forecasters use the government data collected at 8,000 sites, then employ high performance computers and software to focus on a single weather challenge, to help companies such as Ford Motor, Sears and Wal-Mart predict swings in demand due to the economy and other factors.[10]
- In the B2C book selling business, Amazon's popular "recommender system" that features "Customers who bought this book also bought", utilises data mining algorithms to mine through a mass of data for cross-referencing.
- In a Zdnet article "Web data goes under the microscope", Click Stream data, locator information and transaction details are pulled from different data warehouses with over 50 terabytes of data, to a pooled source for data mining and analysis to be performed.

4.8 CHALLENGES

4.8.1 Standards to enable the world-wide Grid

An open infrastructure is required to enable Grid users to share and access resources openly and to enable application interoperability. Standards are required to allow users access across this future landscape of multi-vendor and varying resources pooled by different Grids. Currently, there is no formal standard adopted yet although Globus is being regarded as the de-facto standard. Among the Grid community, Distributed Resource Management Application API (DRMAA) is working on forming the first working group in the GGF that will address application portability to Distributed Resource Management software. However, there is still much to be done in the standards area.

4.8.2 Security

There are security risks involved in the openness of Grid Computing. Related areas such as payment and accounting mechanisms are also vital to ensure the economic viability of Grid Computing. Recent news have identified parasitic computing (the unauthorised stealing of spare cycles, and unauthorised donating of corporate spare cycles) and tampering with data processed on Grids as negative points to Grid Computing.

The open and networked environment that the Grid hopes to deliver brings about new sources of security risk. With users and customers depending heavily on the provision of services from Grid providers, a denial of service could have serious consequences. This may lead to survivability and vulnerability assessments gaining in importance. Furthermore, the sheers size and scale of these Grid deployments makes it even more difficult for security surveillance to be conducted. Intrusion detection would also evolve to support Grids. Another security related issue concerns intellectual rights. The need to enforce digital rights will become more imperative with the utility-like service model. The Grid also has the additional challenge of providing a robust platform to gain the trust of the users.

4.8.3 Risks of an Early Technology

Grid Computing is in early adopter phase, with little customer testimonials to warrant the analysts' bullish forecasts.

Unknown applications are also another source of complexity. The current Grid activities are working at providing existing applications with more computational resources. When the Grid environment reach a level of maturity, we expect to see a slew of dramatically different applications. While this presents opportunity, it is also a source of complexity, as the unknown nature of these expected applications throw a source of confusion and uncertainty into the Grid arena.

Despite the many promising developments in the Grid community and commercial interests, the challenge lies in determining how Grid computing will cross over to the commercial mass-market sectors.

4.8.4 Complexity of Control of a dynamic and unbounded environment

System Management - the monitoring, control and management of a pool of large-scale resources is complex. Adding to the difficulty are the administrative policies, the regulatory and political aspects, and the potential difficulty in harmonising these as inter-Grid resource gains prominence.

The Grid environment involves a paradigm shift on the mindsets of systems administrators. IT managers used to closed internal systems have to adapt to this unbounded open network and dependencies with parties external to their environment. Resistance in this area would affect adoption.

4.8.5 New computing model

Similar to the above challenge, the new computing model of utility-like delivery of large-scale (teraflop and petabyte) resources requires new programming models, new business model, new operational model etc. These models must be tried, tested, improved and adapted before it can become mature, and it will take a long while.

Businesses in the Grid community are currently working on pricing strategies and economic models for the metered usage of Grids. This is in early stage and is not proven, but is extremely crucial to users adopting the Grid as a computing platform for access to geographically dispersed resources.

5 Singapore Landscape

In this section, we outline a survey of the local landscape and the activities.

Web Services and Peer-To-Peer

Web services as a strategic technology has not gone unnoticed by the local industry and we are already seeing activities stirring up here. IDA is taking the role of a catalyst to help local enterprises acquire the relevant Web services capabilites through the Infocomm Local Industry Upgrading Programme (iLIUP).

ILIUP aims to accelerate the development of infocomm capability of Singapore Enterprises for international market competition; and enhance Infocomm market share and revenue for the mutual benefits of Multi-National Corporations and Singapore. Through the iLIUP partnership, Singapore enterprises can gain exposure to cutting-edge technologies and tap on an established global network of technical expertise that will reduce development costs and time-to-market for their products. Singapore enterprises are also able to leverage on the MNCs' international marketing and distribution expertise, thereby enhancing their capabilities and penetrate global infocomm markets. On the other hand, MNCs can increase their market share by adapting their technologies for locally developed products and services. They are also able to benefit from the Singapore enterprises' vertical domain expertise, make use of Singapore's IT-savvy talents, and IT-ready infrastructure as a test bed and launch pad to Asia. At the time of publication, we have 13 MNCs under Web services iLIUP, including Microsoft, IBM, Sun, Oracle, Hewlett-Packard and BEA.

On a separate front, Microsoft and Hewlett-Packard has set up an XML Web Services Centre and an XML Garage respectively. These centres help to develop and deliver industry training and certification for Web services and other XML-based technologies, thereby, building up a critical mass of manpower for Web services implementation.

In the area of P2P, there are early adopters of this computing paradigm by local software vendors. One such example is Gridnode, which uses P2P architecture in supply chain management. Its flagship software, GridTalk, allows for back-to-back integration across different trading partners in the supply chain, even when they have diverse information systems. It supports many of the industry and technical standards, such as RosettaNet and UDDI. GridNode has been selected by RosettaNet as a certified mid-tier independent software vendor to provide RosettaNet-enabled solutions for the masses.

IDA has in place the Innovative Development Scheme (IDS) to encourage innovative development of P2P products and services.

Grid Computing

In September 2001, the Agency for Science, Technology And Research (A*STAR, formerly NSTB) launched the BioMedical Grid (BMG) to allow sharing, collaboration and cooperation among all tertiary institutes, research institutes and centres involved in biomedical activities in Singapore. This Grid for the Biotech community in Singapore is envisioned to be a high-power, high-bandwidth computing, storage and communication facility, which will allow scientists to exchange and process huge amounts of biomedical information. The BMG will be operational by 2003. There will be collaborations with other Grid computing partners in

US, UK, Japan and Korea. There is also a collective group of intellectual capital drawn from universities and government agencies such as NUS, NTU, KRDL, IDA & A*STAR to provide direction ranging from strategic vision to technical expertise on the construction of a grid in One North.

Beside biomedical field, there are also other research activities deploying Grid computing for manufacturing product design, problem-solving environment for aerospace, and access grids for collaboration among several ASEAN (Association of Southeast Asian Nations) universities using Advanced Collaborative Environments.

In the pipeline are also activities to increase the education and training of Grid Computing resources in Singapore. The BioInformatics Institute (BII) and IDA will jointly organise a Grid Computing Symposium in March 2002. IDA is also partnering BII in offering the Grid Computing Training/Attachment Programme that aims to accelerate the training and development of manpower in Grid Computing.

R&D in Grid Computing

Some Grid Computing R&D activities in the academia are outlined below:

NUS Project in a Grid Management System

AliCE (Adaptive and scaLable Internet-based Computing Engine) Project under Prof. Teo Yong Meng. The project was started 1999-2000, and is perhaps the first implementation of grid using Sun's Javaspaces technology.

The project's aim is to support the development and execution of generic grid applications, i.e., run applications over a geographically distributed, heterogeneous collection of resources (computers, networks, etc.) It is a middleware that provides a programming abstraction as well as masking the heterogeneity of the underlying networks, hardware and operating systems.

NTU School of Computer Engineering Grid component projects

Several projects in the infrastructure layer of grid computing is being undertaken in NTU, namely under Dr Francis Lee, Dr Cai Wentong, Dr Stephen Turner and Dr Bertil Schmidt. Two highlighted projects are "Grid Resource Management" and "Grid Bandwidth Broker".

Grid Resource Management – A Clearinghouse

A distinctive feature of the Grid is the sharing of various resources over the Internet. Globus has been used widely by researchers and provides a number of services, such as resource allocation and security, for enabling sharing of resources over Internet. However, it requires the user to explicitly express the resource requirements and to have explicit knowledge about using various services.

This project's aim is to study the resource management for grid computing and to develop middleware tools that will hide the difficulties of using the Grid and provide various resources automatically to meet the end-user's performance requirements. One of such tools is the Grid Resource Broker that will act as a "Clearinghouse" for the international Grid environment. It is a brokerage system, interfaced with the Globus, for resource accountability, access-control and automatic allocation.

Grid Bandwidth Broker

The key component of a grid infrastructure is the grid management system or the middleware that provides this grid-like behaviour. Grid Management Systems are defines as a middleware that enables flexible, secure, co-ordinated resource sharing among dynamic collections of institutions by Ian Foster. As the definition indicates, the network will play a major role in Grid as large amounts of data are transported between institution.

In this project, NTU will focus on building up a collection a Bandwidth Brokers (BBs) which inter-operate with each other through a defined Bandwidth Broker protocol, called Simple Inter-BB Signaling protocol. With the BB in place, Grid application would be able to not only reserve computational resources but also network bandwidth resources. This project aims to fill a gap not filled by Globus and complement it.

NTU School of Computer Engineering Grid Applications

Unlike the above 2 projects that work on capability in the grid infrastructure layer, the below 2 projects are aimed at building applications that are built on top of grid computing infrastructure.

Distributed Simulation on Grid

Distributed simulation is an important strategic technology for linking simulation components of various types at multiple locations to create a realistic and complex "virtual world" for the simulation of highly interactive activities.

The High Level Architecture (HLA) for Simulation was developed under the auspices of the US Defense Modelling and Simulation Office as a worldwide standard to facilitate interoperability among simulations and promote reuse of simulation components. Large-scale HLA-based distributed simulations may involve a huge number of geographically distributed computing nodes. To meet the real-time requirement usually demanded by the distributed simulations, resource management mechanisms that balance the load in the simulation and provide fault-tolerance capability are desired.

NTU is investigating how the HLA/ Real Time Infrastructure (RTI) and the Globus can be married together so that the HLA/RTI takes care of simulation related tasks and the Globus looks after all other issues (e.g., security, resource management and fault-tolerance).

Bioinformatics and Biomedical Grid Applications

Bioinformatics and biomedical applications usually require a huge amount of computing power (in the order of Teraflops), involve a large amount of data (in the order of Terabytes) and demand an ultra high bandwidth network. We are studying two applications using the Grid technology. One is on protein alignment and the other on Magnetoencephalography (MEG) image analysis and visualisation.

Discovered new protein sequence is usually analysed by comparison with genetic sequence databases which are growing exponentially. To speedup the analysis, a sequence database is partitioned and the tasks of sequence alignment are distributed over a computational grid.

MEG is an imaging system that enables the doctor to trace brain patterns of patients. This instrument is place around the head of the patient and the brain signals are recorded. Doctors can then analyse the brain patterns and narrow down the source of the disease. This task of analysis takes a long time due to large amount of data involved. The joint project between School of Computer Engineering and Cyber Media Centre, Osaka University, uses Grid technology to enable near real-time analysis of data.

IHPC's Science & Engineering Research Grid

IHPC is the leading research institution and source of expertise in high performance computing in Singapore. They are involved in several projects and inititiatves in high performance computing. One such project in Advanced Collaborative Environment was outlined in the Grid Computing Advanced Collaborative environment application section.

A notable project in the area of grid computing is their in-house project, the Science and Engineering Research Grid (SER-GRID). This grid infrastructure is built to enable shared, access to HPC resources housed at IHPC including their 1 teraflop supercomputer, a cluster of IBM eServer p690 compute servers (Regatta). They will be using Globus, linking up with NUS/NTU and potentially with overseas research institutes in US and UK and overseas industrial and academic R & D collaborators.

6 Conclusion

6.1 The Vision

Web services, P2P and Grid Computing will take roots in the NGIA landscape, bringing significant change to the Internet environment. These distributed computing technologies will be manifested in broad reaching concepts such as Utility-like computing, Virtualisation, Services on Demand, the Ubiquitous Internet, and the Internet as a Computing Platform. It will diverge off to niche areas such as P2P established as the underlying distributed architecture for dynamic collaboration, Grid for collective computing and distributed computing, and Web services for Velcro-like enterprise application integration and B2B process integration. With the current community acceptance of standards and the popularity of open source, businesses will no longer compete on proprietary technologies, but move towards competing on implementations of open specifications and services.

The Change is happening – Impact on Industry, Usage

These technologies contribute to a change in the market and the technology user landscape. Software vendors will evolve from a pure product and license fee structure to include services platform provisioning and subscription-based model. Some resource providers will evolve beyond selling hardware and services to be a utility provider of metered processing cycles. Data Centres may become Grid Centres, providing access to data, computational and application resources. Skill sets will also change in line with industry needs.

These technologies Converge and Diverge to shape the NGIA

Convergence - While these three areas evolve in maturity to fulfil their individual end goals, the development is not without interdependencies and synergies. We expect some level of convergence among them in the form of standards and shared services such as directories and registries. Web services standards will be adopted in P2P and Grid communities. UDDI registries will host not only pointers to Web services but grid computing resources in the future. P2P architecture will be used extensively in Web services and Grid computing applications.

The core of the NGIA will be formed by a convergence of all three technologies. Conceptually, they all embody distribution and decentralisation, services and metered computing. Resource provisioning of processing power in unused edge devices and distributed storage using P2P architecture is already an early manifestation of high throughput Grid Computing beyond academic world.

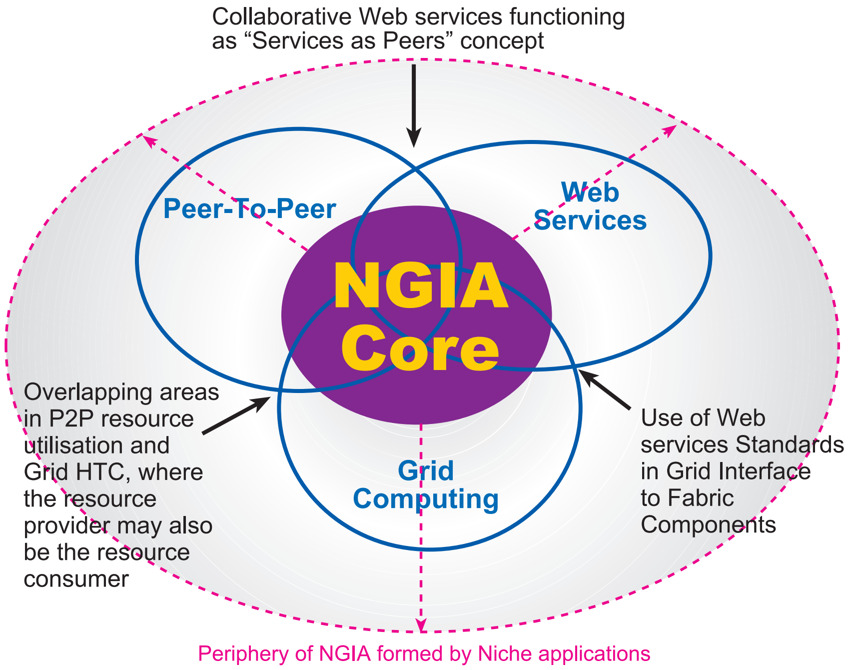


Figure 16. Convergence of the Web Services, P2P and Grid Computing

Niche Areas **-** From this common core of the NGIA, each technology diverges off into specialised areas to enable niche applications. P2P will have resource provisioning, distributed processing, distributed storage and edge computing. Web services will have process integration, code exposed on the open internet and reusable software modules. Grid will have ACEs, Data Grids, and transparent OS level access to computational Grids.

6.2 THE CHALLENGES AHEAD

In this report, we look at the vision of the technology, its potential, its shape, form and usage, collating the views of analysts, industry leaders and researchers from labs in Singapore, undeterred by this nascent technology's current look and feel.

All the three technologies are at different stages of development and deployment, and will continue to evolve. Each faces different challenges and still has numerous technical hurdles to overcome before materialising their vision statements and achieving mainstream adoption. Although some of the business models and usage pattern have yet to be established and proven, we see the merits of these technologies and their great potential in fulfilling many of the emerging market and business requirements, and the vision of NGIA.

In this section, we summarise and highlight some of the key challenges outlined in this report. They are broadly classified into two levels. The first level is at the application level and the second level is at the broader NGI environment level.

6.2.1 NGIA Challenges

Web services has been identified as a strategic technology. If the vision come true, the reach will be far-reaching and deep. Several higher level functions are still work-in-progress, such as end-to-end security, multi-party transaction model, context awareness, user-interfacing layer, dynamic discovery and management interfaces. Multiple competing approaches and standards will emerge in these higher level functions though OASIS is increasingly seen as a platform to drive industry adoption of these standards. Besides technical hurdles, Web services adoption will also face the issues of trust and resistance in the paradigm shift towards service-oriented model. Companies may be hesitant to publish their business processes and catalogues of services in the public domain, as there is no framework to build trust and privacy. There is also a need to adopt a different software development model with the service-oriented architecture, where external software modules are dynamically integrated.

P2P is starting to gain sales traction as an enterprise software. P2P collaborative applications, like IM, enterprise file-sharing and P2P groupware, are increasing being found in the PCs and laptops of enterprises. Industries that require vast computing resources in terms of computation, storage and bandwidth, are also experiencing significant cost savings through more optimal utilisation of computing resources by using P2P resource utilisation software. However, issues like the perceived lack of network resources control associated with a P2P networking environment needs to be addressed before greater enterprise adoption can be expected. More importantly, we felt that the adoption of general technology standards by the P2P community has to be established in order for P2P applications to proliferate.

Grid Computing is currently in research and pilot stage, with most applications being deployed in test beds and field trials. Technologies in many areas, such as resource scheduling, allocation and monitoring, security, accounting and fault detection still need to be developed and refined. There is much to learn and the flagship Grid implementation such as the TeraGrid DTF and EU DataGrid will serve as important grounds for concept proofing, pilot and trial, and further refinement of technologies. Global standards are required for interconnection of grids and its spread to the commercial Internet. These efforts are not trivial and will take a long while.

6.2.2 NGI Environment Challenges

Security

Compare a closed, bounded network such as an enterprise environment that is managed by in-house capability to an open unbounded network. The enterprise network is rigorously profiled, monitored and controlled. The open Internet is a distributed, unbounded network where there is no central administrative control and no common security policies. Web services, P2P and Grid Computing thrive in this open environment, there are security and performance risks. Each administrative domain operates in a black box, relying on information from its neighbouring domains, and implements some sort of control and policies outside its local domain. This could result in a chain of dependencies that links the parties involved. Failures result in a denial of service spreading along this chain. A virus is easily spread among connected nodes.

Survivability is an emerging discipline, which focuses on the capability to maintain integrity, performance and security in the face of attack, failure, or accident, which should be explored to counteract the risks pose by distributed computing.

Economic and Usage Models

Web services, P2P and Grid Computing facilitate sharing of resources distributed across the world, be it spare cycles, databases or exposed code. However, for economic viability, two new requirements must be met.

- First is "metering". Measurement of resources sold and pricing strategies is required.
- Second is the profit margin. Vendors used to installation and license fee revenues need to adapt to this new open environment and service-based model.

In this open network, the geographical dispersion of the resources and different ownership added to the complexity of the envisioned service model. The resource owners of each of these resources have different access policies and pricing strategies, and levels of QoS. A framework for resource distribution and balancing supply and demand in an environment where parties are inter-connected to provide services to each other is needed.

Various economic models are being explored, but lack customer testimonials, proven market track records and best practices. Examples of economic models include subscription, utility-based pricing, tenders and auctions. In addition to economic models, an infrastructure is needed to support access, scheduling mechanisms, cross-border currency and legal issues, e-payment, and regulatory policies.

Identity vs. Privacy

In this open unbounded environment, loosely coupled systems are vaulted to change market structures and allow global reach by allowing entities to dynamically join in a business transaction. Identity and reputation systems are being explored to allow unknown parties across the globe to connect to each other to form agile groups for business transactions.

On the same identity issue, as the depth of the Internet functionality increases from email and shopping online, to work and e-business, we see privacy issues increasingly. Programs that monitor web usage and collect profile data from users are increasing and a backlash from the public is beginning.

Challenges of an Emerging Technology

Web services is expected to reach mainstream after 2004. P2P is projected to achieve widespread adoption around the same period. Grid Computing is still in infancy phase and the aim of it being a reliable, consistent and secure platform is expected to emerge in around 2006 or beyond. NGIA will not happen overnight, but evolve over a long period of time. On this long road ahead, with many competing driving forces, stumbling blocks of varying strength and transient trends, the future is indeed difficult to predict.

TECHNOLOGY ROADMAP CHART
Next Generation Internet Applications

Year	< 2002	2002	2003	2004	2005	2006
	•WS launched in 2000 •WS development tools emerged •Informational WS appears	• E-marketplace WS emerged • WS technology adopted in EAI • OASIS and W3C champion WS stds	• WS tools mature. • Consensus reached in transaction and security stds	• WS enters mainstream • Implementations of WS with complex transactions • Pricing and accounting models for WS mature		Towards automatic and dynamic Web Services collaboration
	•Biz interests in P2P apps emerged •JXTA launched in 2000	• Regulated brokerage P2P file-sharing sys emerged • Co-existence of decentralized P2P file-sharing sys	• Leaders in P2P collaborative groupware space emerged • System requiring distributed computing will incorporate P2P APIs	• Leaders in the P2P Internet file-sharing space emerged • P2P development tools mature		Utility-like computing on Grid: A robust, distributed, collaborative service architecture
	•Increased commercial interest in grid computing •Several National Grid projects announced. •Data Grid construction capability in Globus released	• Commercial activities in data and storage grids emerged • Commercial release of Globus expected • Web Services technology incorporated into Globus	• High-throughput Grids take-off • Flag-ship Grids (TeraGrid, UK National Grid, CERN) become operational.	• Grid applications begin commercial adoption - First wave of Commercial Grids provide utility-like access to processing power • Globus features are incorporated into OS		Cluster Grids provide major source of high performance computing power, capacity nears 1 Petaflop
	STANDARDS 2001 •SOAP v1.2 in working draft stage. •WSDL v1.1 draft submitted to W3C. •UDDI v2 released. •WSFL and XKMS submitted to W3C as draft. •WSUI initiative announced.	**STANDARDS 2002** • UDDI v3 released • WSEL and WSFL delivered	**STANDARDS 2003** • BTP expected to be finalised • WSUI/ WSCM expected to be finalised	**STANDARDS 2004** • XKMS Standards ratified		
	Industry Analysts Figures: <2002 •Total investments in P2P reached US$378m for Y2000(Aberdeen Group)		**Industry Analysts Figures: 2003** • Web Services Market market projected to worth US$1.7b (Gartner) • Utility Computing will be used by nearly 10% of the US$160b IT outsourcing market (Hewlett Packard)	**Industry Analysts Figures: 2004** • 33% of online population will uses P2P services for storing and sharing personal data. (Forrester) • Revenues from P2P platforms, applications, and professional services to hit US$925m (Aberdeen Group)	**Industry Analysts Figures: 2005** • 150m IM corporate users (META Group) • 500m total IM users (META Group)	**Industry Analysts Figures: 2006** • 5% of commercial organisations that routinely use supercomputing cycles will turn to cheaper grid computing for nonsecurity-sensitive resources (Gartner)

Glossary

A*STAR	Agency for Science, Technology And Research (Formally NSTB)
ACE	Advanced Collaborative Environment
AIM	America Online Instant Messenging
ALICE	Adaptive and scaLable Internet based Computing Engine
AOL	America Online
APGrid	Asia-Pacific Grid
API	Application Program Interface
Argonne	Argonne National Laboratory
ARPANET	Advanced Research Project Agency Network
ASEAN	Association of Southeast Asian Nations
ASP	Application Service Provider
B2B	Business-to-Business
B2C	Business-to-Consumer
BB	Bandwidth Broker
BII	BioInformatics Institute
BMG	Bio Medical Grid
BTP	Business Transaction Protocol
Caltech	California Institute of Technology
CAVE	CAVE Automatic Virtual Environment
CERN	European Organization for Nuclear Research
CNRI	Corporation for Networking Research Initiatives
CORBA	Common Object Request Broker Architecture
CPU	Computer Processor Unit
CRM	Customer Relationship Management
DARPA	Defense Advanced Research Projects Agency
DDBJ	DNA DataBank of Japan
DMSO	Defense Modelling and Simulation Office
DoD	United States Department of Defense
DRM	Digital Rights Management
DSSR	Debis Systemhaus Solutions for Research
EAI	Enterprise Application Integration
EDI	Electronic Data Interchange
ebXML	Electronic Business eXensible Markup Language
EJB	Enterprise Java Beans
EMBL	European Molecular Biology Laboratory
ERM	Enterprise Resource Management
ETL	Extract, Transform and Load
EU	European Union
Genbank	GenBank at the National Centre for Biotechnology Information
GGF	Global Grid Forum
GIMPS	Great Internet Mersenne Prime Search
GUSTO	Globus Ubiquitous Supercomputing Testbed Organization
HLA	High Level Architecture
HPC	High Performance Computing
HTC	High Throughput Computing
HTTP	Hyper-Text Transfer Protocol
IDC	International Data Corporation
IDL	Interface Definition Language

IETF	Internet Engineering Task Force
iLIUP	Infocomm Local Industry Upgrading Programme
IM	Instant Messaging
IMPP	Instant Messaging and Presence Protocol
IP	Internet Protocol
ISV	Independent Service Vendors
J2EE	Java 2 Platform Enterprise Edition
JDBC	Java Database Connectivity
JXTA	Juxtapose
KM	Knowledge Management
LAN	Local Area Network
LDAP	Lightweight Directory Access Protocol
LHC	Large Hadron Collider
MDTD	Molecular Dynamics Trajectory Database
MEG	Magnetoencephalography
MEMS	MicroElectro Mechanical System
MIME	Multipurpose Internet Mail Extensions
MSN	Microsoft Network
NASA	National Aeronautics and Space Administration
NASSL	Network Accessible Service Specification Language
NCSA	National Centre for Supercomputing Applications at the University of Illinois at Urbana-Champaign
NGI	Next Generation Internet
NGIA	Next Generation Internet Applications
NPACI	National Partnership for Advanced Computational Infrastructure
NPI	New Productivity Initiative
NSF	National Science Foundation
NTU	Nanyang Technological University
NUS	National University of Singapore
OASIS	Organization for the Advancement of Structured Information Standards
OPIE	Open Portal Interface Environment
OS	Operating System
P2P	Peer-to-Peer
P2PWG	Peer-to-Peer Working Group
PDA	Personal Digital Assistance
PDB	Protein Data Bank
PKI	Public Key Infrastructure
PSE	Problem Solving Environments
QoS	Quality of Service
RCSB	Research Collaboratory for Structural Bioinformatics
RDAWG	Remote Data Access Working Group
RDF	Resource Description Protocol
RMI	Remote Method Invocation
RPC	Remote Procedure Call
RSS	Resource Description Protocol Site Summary
RTI	Real Time Infrastructure
SAML	Security Assertion Markup Language
SciDAC	Scientific Discovery through Advanced Computing
SCL	Simple Object Access Protocol Contract Language
SCM	Supply Chain Management
SDK	Software Development Kit
SDL	Service Description Language

SDSC	The San Diego Supercomputer Centre at the University of California at San Diego
SETI	Search for Extra-Terrestrial Intelligence
SF-Express	Synthetic Force-Express
SIP	Session Initiation Protocol
SLA	Service Level Agreement
SMS	Short Messaging Services
SNPs	Single Nucleotide Polymorphisms
SOA	Service-Oriented Architecture
SOAP	Simple Object Access Protocol
SODA	Service Oriented Development of Applications
TCP	Transmission Control Protocol
TeraGrid DTF	TeraGrid Distributed Terascale Facility
W3C	World Wide Web Consortium
WSCM	Web Services Component Model
WSDL	Web Services Description Language
WSEL	Web Services Endpoint Language
WSFL	Web Services Flow Language
WSUI	Web Services User Interface
UDDI	Universal Description, Discovery and Integration
UN/CEFACT	United Nations Centre for Trade Facilitation and Electronic Business
UNICORE	Uniform Interface to Computing Resource
URL	Uniform Resource Language
UW-Madison	University of Wisconsin-Madison
X-KISS	eXtensible Markup Language Key Information Service Specification
XKMS	eXtensible Markup Language Key Management Specification
X-KRSS	eXtensible Markup Language Key Registration Service Specification
XML	eXtensible Markup Language
XML-SIG	eXtensible Markup Language Signature

FEEDBACK

This Infocomm Technology Roadmap is a continuous effort, to be revised and updated as technologies evolve. For enquiry and proposals on technology development and initiatives, you can reach us at:

Infocomm Technology Roadmap
Technology Direction
Chief Technology Office
Infocomm Development Authority of Singapore
8 Temasek Boulevard
#14-00 Suntec Tower Three
Singapore 038988

Website: http://www.ida.gov.sg

(Click on "Technology Development", followed by "Infocomm Technology Roadmap")

Email: roadmap@ida.gov.sg

Survey Form

Survey Form

IDA Technology Roadmap 2002

Track 2: Next Generation Internet Applications

With active contribution from the industry and research community, IDA has launched the *Infocomm Technology Roadmap Release February 2002*. You have either attended the Roadmap Symposium or downloaded a copy of the Technology Roadmap document from our website. Your feedback is valuable to us to better our future services for you. We appreciate if you could spare a few minutes to fill up the following survey.

Please return the completed questionnaire to IDA:

via Fax: **+(65) 211 2211 (Attention to Ms Saliza Mohd)**

or *via Mail* to the address on the previous page.

Company Name : ______________________

Your Name : ______________________

Designation/ : ______________________

Area of Expertise : ______________________

Email Address : ______________________

Contact Number : ______________________

Q1. With regards to the Roadmap Report Release Feb 2002 Track 2, please rate the following on a scale of 1 to 5.

Factors	Excellent				Poor
Usefulness of the roadmap	5	4	3	2	1
Completeness of coverage and contents	5	4	3	2	1
Ease of understanding	5	4	3	2	1
Usefulness of the Roadmap Chart	5	4	3	2	1
Relevance to you or to your business strategy/planning	5	4	3	2	1

Comments (if any): ______________________

____

Q2. Please indicate the accuracy (in terms of trend & development) of each topic in the Technology Roadmap Report. Please rate them on a scale of 1 to 5.

Area/Topic	Accurate				Inaccurate
Web Services	5	4	3	2	1
Peer-To-Peer	5	4	3	2	1
Grid Computing	5	4	3	2	1
Singapore Landscape	5	4	3	2	1
Roadmap Chart	5	4	3	2	1

Comments (if any): ______________________________

____

Q3. Do you have any suggestions for improvement on the Technology Roadmap?

____

____

____

Q4. If you are an industry player in Web services, Peer-to-peer and Grid Computing, what are the supporting infrastructure that you think should be put in place in Singapore?

____

____

____

Q5. Would you like to be informed of our future Infocomm Technology Roadmap Symposiums/Reports?

Yes / No

Thank You