IBM's "On Demand" Campaign:

A New Paradigm Demands Our Attention

By Miles Jenkins

Introduction

Consider for a moment the light by which you're reading this magazine. Does the electricity that makes it shine come from a coal-burning power plant? Or from the water tumbling over Niagara Falls? Or is it imported from a nuclear plant in Pennsylvania? Unless you have way too much time on your hands, you don't know, and you don't care. Your home or office is connected to the electrical grid – a network of transmission lines connecting hundreds of power generation plants, whose combined output is merged and distributed to millions of consumers who get billed based on usage. Each user expects to be able to simply flip a switch and have the light come on – on demand.

That phrase "on demand" has become ubiquitous in IBM's recent marketing material. It seems to be a condition of employment at IBM that you use this phrase at least once per paragraph. Their ads imply that we are entering a new era, when IT services and capacity will similarly be accessible "on demand". IBM is investing billions in this initiative, which says something for their vision of where information technology is headed.

Outside the IBM world, Oracle's new version of their DBMS is called Oracle 10g – the "g" standing for "grid", implying that Oracle too is preparing for the coming of "Grid Computing".

In this article, I'll look behind the buzzword, at some of the concepts – virtualisation, grid computing, autonomic computing – that are coming out of the academic and scientific world, and poking their noses into business IT. And, at some actual products, current and future, that are designed to make this vision a reality.

"Ten years from now, we'll view the IT infrastructure more like a utility." – Frank Soltis (IBM)

The "on demand" vision is in contrast to our normal way of thinking in IT. Typically, if a new application is being implemented, or an existing one expanded, it means buying some hardware, and some software, and possibly some telecom bandwidth. And probably most expensive of all, technical specialists to implement, manage, and support this infrastructure.

"On demand" implies that whatever resources are required are already in place somewhere, waiting to be allocated. The key concept is virtualisation. Users and applications aren't hard-wired to actual servers, processors, and storage, but rather request these resources, and have them allocated from a pool of distributed, possibly heterogeneous devices.

At a high level, the on demand paradigm recognises businesses' need to reduce risk, and improve capital efficiency and financial predictability. It also recognises that businesses more and more are forming tightly integrated groups of strategic partners, each contributing its specialised piece of a larger function or supply chain.

And, integrated businesses are often 24/7 global operations, that require more robust, realtime systems, able to withstand spikes in usage, component failures, and attacks from outside.

While there is enormous inertia to be overcome in the IT world, I agree with the many who feel that in the coming decade, this concept has the revolutionary potential to permanently change the IT landscape, on the scale of the Internet explosion in the last decade.

Grid Computing

A Grid can be defined as a collection of distributed, networked, heterogeneous computing resources, that appear as a single, virtual computing system. Not only storage, but processing cycles, application software, and telecom bandwidth are virtualised – i.e., the requesting user is not bound to specific devices to process a request. Compare this to the "Need an application? Buy a server (and remember another password)" reality of today.

It's been estimated that less than 20% of server capacity worldwide is in use at any given time. But much more than simply utilising excess capacity, grid computing has a number of benefits:

- integration of heterogeneous systems and devices.
- resolving over-provisioned, excessively cost-bearing IT infrastructure.
- ease of increasing resource capacity dynamically in response to fluctuations in demand.
- optimisation of workloads – for example, where time zone differences mean resources are underutilised in one location, while another location is at peak demand.

Undoubtedly the most compelling reason the world will move to "utility IT" is the possibility of a drastic reduction in the cost of entry – and exit – to IT functionality. Consider a small architectural design firm, that now can't compete with larger companies, because the latest, most sophisticated design software costs millions of dollars, and requires hardware and technical resources well beyond its means. But if it were able to connect to a grid, and pay on a subscription basis, it could compete and prosper.

Miles Jenkins

Some of the early foundations of utility IT are falling into place: Storage Area Networks (SAN) already have a considerable degree of virtualisation. Administrators can make changes to storage allocation – for example changing volume sizes, or moving volumes between devices – transparently to applications using the storage. The sophistication of billing systems has reached the point of being able to meter networks for usage of various resources, whether for chargeback billing within organisations, or for billing external customers.

The Globus Project

The Globus Project, supported by IBM and others, has been founded to define an open-source Grid reference architecture, to develop industry standards for grid computing, and to develop a set of tools to assist in the implementation of a grid. It has published a set of specifications and standards to establish a common technical base for grid computing, called OGSA (Open Grid Services Architecture). OGSA combines a number of Web Services standards – such as XML and the XML-based WSDL (Web Services Description Language), UDDI (Universal Description, Discovery, and Integration), and SOAP (Simple Object Access Protocol) – with the grid computing standards. Together, these standards and services will enable cooperation and access to applications dispersed over either public or private networks.

Needless to say, security is a major consideration in grid environments. A security protocol known as GSI (Grid Security Infrastructure) is a public-key-based protocol that uses X.509 certificates to provide single sign-on authentication, by which a user can establish a proxy credential that can authenticate with arbitrary remote services on the user's behalf.

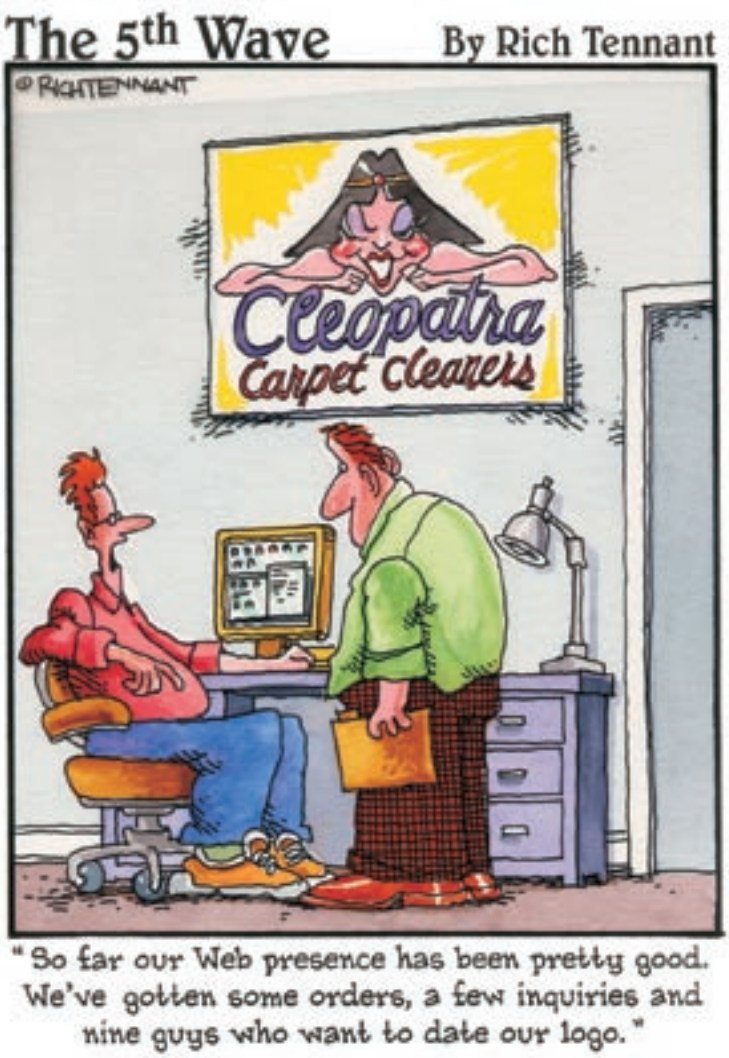
A reference implementation of the OGSA has been created: The Globus Toolkit is a set of software components used to establish security, resource management, data management, communication, fault detection, and cross-system portability. In addition to a number of grid implementations for major customers, IBM has its own internal "intraGrid" that links R&D resources around the world, and serves as a test bed for Grid solutions.

"Grid computing will allow companies to treat IT more like other utilities, such as gas or electricity."
– Larry Ellison (Oracle)

This scenario, of memory and CPU cycles dynamically allocated among jobs according to priorities, begins to look simply like a higher order operating system, working above all the subsidiary elements in the grid. (If you're experiencing déja vu, it could be because the futuristic world of grid computing looks suspiciously like a 1985-era System/38 shop!)

The CERN Grid

One of the first major implementations of grid computing is being undertaken by CERN, the European particle physics laboratory, which is constructing the world's biggest particle accelerator under the Swiss countryside. When operational in 2007, the data it generates in a single year would fill a stack of CDs the height of the Eiffel tower. The storage and analysis of all that data would outstrip the capabilities of any single computer.



© The 5th Wave, www.the5thwave.com

So a worldwide network of supercomputers, scattered in laboratories and universities around the world, is being linked to form a single, giant virtual supercomputer. When a scientist somewhere in the world wants to solve a problem, a software agent will sniff out where the software, processing power, memory, and data storage appropriate for the task is available. In a way, it builds on the Internet idea, with the web's anywhere-to-anywhere information display being extended to the other dimensions of information technology. (Trivia buffs may note an eerie parallel: CERN is where the World Wide Web was invented!)

Autonomic Computing

Last August, when southern Ontario was hit by a major power blackout, the most astonishing thing to me was not that the system broke down, but that no mortal human could explain what had happened – not just hours, but even days after the most critical utility in society had massively failed. We would learn that the "grid" is computer-controlled – indeed, its complexity is far beyond the capability of human management. It goes without saying that grid computing likewise entails a complexity of system and network management beyond the abilities of human operators.

So, often mentioned in the same breath with grid computing is "Autonomic" computing. The word "autonomic" sounds a bit cooler and less political than "autonomous", but they both mean the same thing – self-governing. The word "autonomic" was chosen for its association with the human body's autonomic nervous system, which frees the brain from the burden of managing lower-level functions.

An IBM Senior VP and Director of Research has stated that the single biggest challenge facing the IT industry, which if unsolved will prevent us moving to the next era of computing, is a problem of IT's own success – that obstacle is complexity.

We're on the verge of a complexity crisis, with exponential growth in the number and variety of systems and components. And increasingly complex systems require more and more expensive human resources to implement and maintain them. Consider the following estimates:

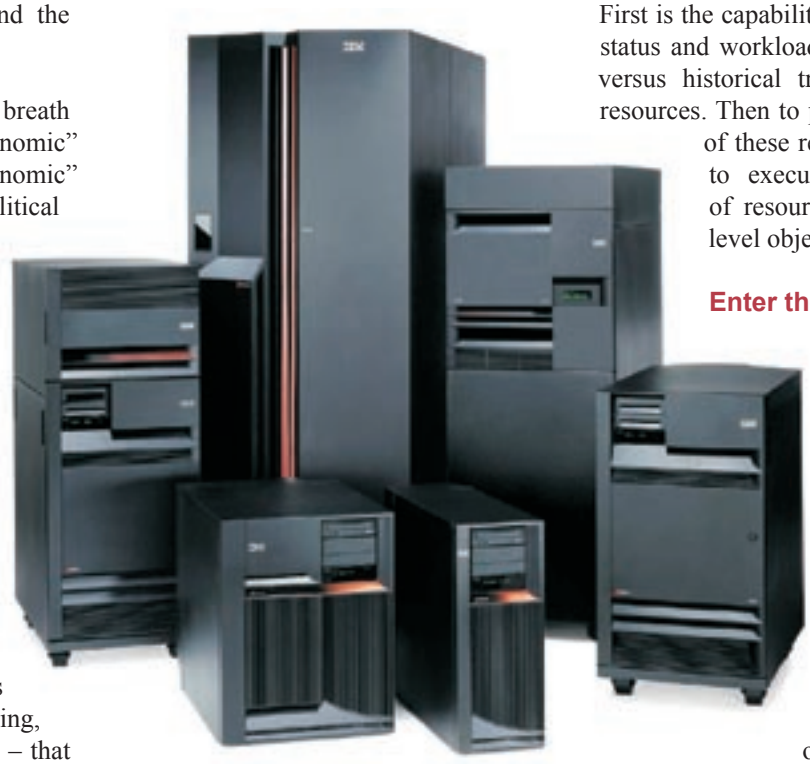
- For every dollar spent to purchase storage, $9 is spent to have someone manage it.
- In some shops, 40% of software development time is devoted to testing: as the complexity and interrelatedness of systems grows, the testing problem grows exponentially.
- More than 40% of companies' IT investments are used just trying to get technologies to work together, which doesn't directly drive business value.
- The root cause of about 40% of system outages is operator error – not due to negligence or inexperience, but because systems are too difficult to understand or manage.

A number of forces are leading the industry towards autonomic systems. Complex heterogeneous infrastructures comprising several applications, hundreds of system components and interfaces, and thousands of tuning parameters have became a reality.

And of course, human costs have begun to exceed the costs of technology. The intent of autonomic computing is to make systems:

- **self-configuring:** where new features, whether hardware or software, can be dynamically added to the infrastructure without a disruption to service.
- **self-healing:** where systems discover, diagnose, and react to disruptions.
- **self-optimising:** where systems monitor and allocate resources automatically. I.e., extending concepts like logical partitioning and dynamic workload management across multiple heterogeneous systems.
- **self-protecting:** where systems can anticipate, detect, identify, and protect themselves from intrusions such as unauthorised access and virus attacks.

Autonomic computing implements a four-stage "Autonomic Control Loop". First is the capability to monitor current status and workload, then to analyze it versus historical trends and available resources. Then to plan the reallocation of these resources, and finally to execute this reallocation of resources to meet service level objectives.



Enter the iSeries

iSeries servers are well-suited to playing a role in "utility IT". Already, they have many features that point towards the "on demand" future: dynamic partitioning, on-off capacity upgrades, the ability to self-optimize and balance system performance.

Out of the box, it features a high level of integration, and its strength in hosting multiple operating systems – Linux, AIX, Windows – in addition to its native operating system, and its Java capabilities – and of course its industrial-strength security and availability, compared to the Unix and Intel/Windows competition – give it a great advantage in a grid scenario. The entire eServer product line, as well as the DB2 and WebSphere products and tools, support Grid computing.

In Summary…

Many have long thought that computing technology would follow the same historical path as the telephone. In the early days of telephony, each company bought and maintained its own private telephone system. Until finally one day, the advantages in function, and the reduction in cost, of simply connecting to a ubiquitous public network from standard phone sets changed the paradigm.

Today, when you go on a business trip to Pittsburgh, you have to take your own laptop computer, and hope you can get a connection back to your home company to access your email and company applications. You never consider unplugging your phone and your TV set and lugging them to Pittsburgh. You assume that your client's office will have a phone connected to the public network, and your hotel will have a TV connected to the global *Simpsons* Rerun network.

However, if you've spent more than three days working in IT, you can probably list several barriers to "on demand" computing. In Information Technology, just because something makes sense, and the technology exists to make it happen, doesn't mean it will happen any time soon. The industry is dominated by profit-motivated vendors, and a true "on demand" era won't arrive until there is a convergence of their interests with their customers' interests.

One of the biggest barriers I see to the concept of utility IT is that application software is not architected in a way that makes this simple. It has been in the interest of software vendors, and the convenience of developers, to have their software tightly coupled to hardware platforms, operating systems, and DBMSs. Whereas true grid computing requires a degree of abstraction of functionality from the infrastructure supporting it.

It would seem to me that "utility IT" initially plays to the advantage of the very large integrators and Application Service Providers, who already have large numbers of dispersed resources, and multiple customers to accommodate. In contrast to the Internet explosion, which was popularised by the masses in a bottom-up fashion, the "On demand" revolution is likely to be a top-down phenomenon.

Grids will most likely start as "intraGrids", within a single enterprise, and then migrate beyond corporate borders, entailing server consolidation, and capacity on demand, with more flexible purchasing and financing arrangements from IT vendors.

So that five years from now, when Acme Widgets decides it finally has to upgrade its systems, it may compare the cost of upgrading and maintaining its IT infrastructure, and find it much more cost-effective to simply subscribe to off-site services.

If you're sceptical, it's well worth remembering that Microsoft Windows, and the Internet, both percolated in obscurity for many years before their time finally came, and they changed the landscape permanently, and fortunes were won and lost.

For more information…

This article has only scratched the surface of some complex technical areas and issues. There is a some excellent information on Grid computing at the Globus Alliance Web site: www.globus.org. Details of the OGSA (Open Grid Services Architecture) are at www.globus.org/ogsa. IBM has sites consolidating On Demand material: www.ibm.com/ondemand, Grid computing: www.ibm.com/grid, and Autonomic computing: www.ibm.com/autonomic. TUG

Miles Jenkins *is a consultant with ADC Software Systems, in Thornhill, Ontario. He can be reached at (905) 695-4028, or via email at miles.jenkins@rogers.com.*

InternetBusinessSimplified

sofCast Inc. now offers the Decentrix Web Site Solution: a secure, centrally hosted service that allows you to create, modify, and manage a professional Web site, all from a standard Web browser.

sofCast powered by Decentrix

No longer is it necessary to hire or contract expensive technical and design specialists. There is no hardware or software to buy, no contract to sign: only a low initial expenditure, and fixed, affordable monthly billing. In addition to a full-function Web site, your subscription gives your organization its own private, secured Intranet – a full suite of collaboration and communication tools: Email, Shared File Folders, Calendars, Contacts, and more. And, if you have a product or service to sell, your site can optionally have an on-line store, giving your business 24/7 promotion and selling, around the world. Call us today, or visit our site:

www.sofcast.com

Eclipse

Eclipse Technologies Inc.
authorized representative 1-877-644-4482