C3B Consulting's

INSIGHT-SPECTRA

Applying common sense to using technology intelligently

www.insight-spectra.com

Volume 23 Report 6; November 2010

VMworld 2010: what needs to be done to take virtual roads to the cloud

David Hill, Mesabi Group

Management introduction

VMworld 2010 in San Francisco was by any measure a great success. The 17,000+ total attendees was up significantly compared to last year's 12,500. A crowd that size does not descend upon a conference unless there is something really compelling is going on. If so, what was it?

In this analysis, David Hill looks at what was announced. In particular he reflects on what needs to happen if virtual roads are indeed to end in a cloud.

A virtualization vision

Let me start with VMware's vision of virtualization and then see why customers likely feel compelled to follow the company's path. Arguably there are three stages of virtualization — which may explain why the conference was titled "Virtual Roads. Actual Clouds" and why VMware (along with others) is pushing the concept that IT is on a three-phase (also called stages or waves) journey.

The first (stage) is the IT production phase where server virtualization is being used for the first time to consolidate physical servers with one operating and application system onto virtual servers or virtual hosts which then run many virtual machine (VM) guest operating systems along with their applications. The obvious end result is improved physical resource utilization (such as CPU utilization). The benefit of server consolidation is cost savings and these have proven to be advantageous to IT organizations, particularly those under cost pressures.

The second phase of the virtualization journey involves business production. This is about quality of service (QoS) where the ability to be able to set and then meet service level agreements (SLAs) comes into focus. A primary benefit of achieving this stage is nominally quality and that is not bad of itself. However, for IT organizations, it is really about being able to make sure that their initial ventures into virtualization continue to be successful and predictable, enabling them further to extend their use of virtualization.

IT organizations first started to virtualize those applications where the downside (or risks) that might occur if poor service occurred were unlikely to be work or job threatening. As confidence in virtualization has grown, so has the scope of what might be virtualized. Even so, IT does not want any more exposure to risk than necessary, so improving quality of service is a logical as well desirable goal. Moreover, guaranteed responsiveness and reliability are crucial as more sensitive and business critical applications undergo the process of virtualization.

To that point, VMware reported a survey stating that found only 28% of applications are virtualized today — while customers say that 75% is their goal. So a lot more remains to be done. Moreover, VMware's survey found that many customers now consider themselves to be in this second stage of the journey, meaning that QoS considerations are a necessary ingredient for their continued adoption of virtualization.

But improving quality alone is not enough — which brings us to the third phase: delivering IT as a service. IT has always been classified as a service organization. But this concept — of IT as a service — goes far beyond. Arguably it dramatically changes the game. For example, terms like 'service catalogs' and 'zero touch infrastructure' are bandied about with the end result becoming 'cloud computing'. VMware seems to consider the cloud as an inflection point that will fundamentally transform the delivery and management of IT services where the goal is to replace:

- static and brittle management frameworks, rigid processes and rudimentary automation
- with streamlined management enabling agile, self-service IT.

The intention is that not only are such services more reliable but they are also more flexible. You can do more things faster, easier and at less cost to satisfy business requirements.

Sounds great, does it not? It is, but it requires a signif-

icant transformation in the ways people work with and deploy IT.

A few years ago, Paul Maritz raised the question (roughly stated): 'is virtualization the last big product of the old IT era or the first new big product of the next IT era?' His answer then was that it was both and that seems acceptable. Nevertheless, it also presents challenges.

The challenge of moving to the cloud

By definition, moving from one era to another produces a transformation resulting in a significant change process. Progressing through the three virtualization stages affects not only the organization but individuals, as well. For example, a reasonable proportion of employees will wonder:

- will I lose my job?
- will I have to learn new skills (and who will pay for this)?
- will my status and pay be impacted positively or negatively?

One good way of looking at such a change process is through the long-recognized Lewin-Schein model, which also defines three stages of change:

- unfreezing
- change
- refreezing.

In the unfreezing stage, key issues include: understanding how present conditions lead to dissatisfaction and how to prevent ignoring information that is necessary to bridge the gap between what is believed now and what needs to be believed for change to occur. Lewin-Schein also recommends addressing any anxiety employees might feel about having to unlearn what had been previously accepted.

In the change phase, the Lewin-Schein model suggests that the most important task is to identify clearly the gaps between the current and proposed state (see also page 8 of this **INSIGHT-SPECTRA**). In the refreezing stage, the new behavior (in this case, the cloud) becomes the norm — and the changes becomes institutionalized.

In helping IT organizations determine what it will take to move to the cloud, the IT vendors that will be most successful will have to identify (and directly or subtly) address behavioral issues and habits that have to be altered in order to achieve success. They cannot just hope to emphasize the technical and RoI benefits of products. At bottom, addressing behavioral issues will help organizations deal with the complexity at the heart of what needs to change.

Cutting Gordian knots of complexity

The IT issue that creates the greatest dissatisfaction and, thus, the greatest need for change is complexity:

- old-style complexity is inherent in the way that IT organizations have evolved as custom-based shops
- new-style complexity comes about from the introduction of increasingly dispersed technologies, beginning with scale-out servers but also including mobile devices, such as the iPad and Android and other devices.

According to VMware, virtualization is key to how these devices will be supported and managed — via technologies that include its 'View' solution for virtual desktop infrastructures (VDI). But, after a certain level of virtualization, complexity increases and then sometimes impedes performance. For example, when running many VMs on a single physical server, I/O bandwidth becomes a limiting factor in particular processes (such as backup).

Other forms of complexity arise as vendors meet customer requirements. For example, to address IT organizations' desire to move VMs freely (in order to take best advantage of hardware and business process requirements), VMware developed vMotion. And the list goes on …

Speaking of 'whack-a-mole'…

Note well that increased complexity is a natural consequence of wanting to have more functionality or to do more (such as improving manageability and improving security). It is not due to the deliberate actions of vendors. Complexity is the inevitable result of the convergence of two immutable 'rules' in modern IT — you cannot:

- do just one thing (in isolation)
- ignore the law of unintended consequences.

Think of the 'whack-a-mole' carnival game where

popping down one mole leads to another popping up somewhere else. That other mole was an unintended consequence, which may be desirable or undesirable. A good friend of mine uses whack-a-mole to describe the situation in which solving one bottleneck merely transfers the 'opportunity' for another bottleneck to take its place (such as in a server-storage-network environment).

Embarking on a journey almost certainly leads to often unforeseen challenges. In which case, you might ask — why should IT organizations want to go through all the three stages of IT virtualization? Why not stop after the first stage? Well, whacking the first-stage mole delivered cost savings, but another mole popped up — creating issues such as manageability and the need for new key functionality. Once that mole is whacked, the next one will appear — and IT must deal with the consequences.

Does it ever stop? Probably not. But achieving the final goal of delivering 'IT as a service' hopefully will produce less complex, more productive IT environments and user experiences overall.

Management conclusion

As presented at VMworld 2010, VMware's vision is appealing. But there are a number of behavioral issues, as well as complexity issues, that will have to be addressed if organizations are successfully to take the 'Virtual Roads to Actual Clouds'.

That may seem daunting, but VMware and its customers have a lot of help from friends. The Solutions Exchange (a.k.a. the trade show portion of the event) contained a large number of vendor solutions, and the mass of attendees crowding the aisles attests to the level of interest. VMware stated that for each dollar of license revenues it receives, company partners in the ecosystem receive $15.

Thus the virtualization gold rush is on. That means that, while VMware does not have to have the revenues of some of the largest IT vendors (although it has and likely will continue to do very well), it will still have a major impact on the IT industry over the next decade. (The threats and opportunities that it poses to other vendors is a subject for another day.)

The bottom line is that the journey of the three phases of virtualization leading to IT as a service instantiated in a cloud will not be either short or easy. But if VMware's Maritz is correct in thinking that we are on the cusp of a new IT era, organizations that have started the journey to the cloud cannot retreat nor can they stay the same.

David Hill
Mesabi Group
www.mesabigroup.com

IBM's sustainability initiative: outstanding and already out of date

Wayne Kernochan, Infostructure Associates

Management introduction

IBM's new sustainability initiative launched in October prompted the following thoughts. To Wayne Kernochan this is one of IBM's best-targeted, best-thought-out initiatives and deals with most of the recent reservations that he has had previously about IBM's green IT strategy. It some ways it is all that he could have reasonably asked IBM to do.

But, as he describes below, it is not enough. IBM must go much further.

Key details of the initiative

One can readily skip IBM's assertion that the world is more instrumented and interconnected, and that systems are more intelligent so that we can make smarter decisions; that is IBM's way of relating its initiative to its Smarter Planet efforts. It is the effect of IBM's specific solutions on carbon emissions that really matters.

What is new — at least compared to a couple of years ago — is a focus on end-to-end solutions, and on solutions that are driven by extensive measurement. Also new is a particular focus on building efficiency, although the ways that IBM applies sustainability technology extend far beyond that.

Clearly IBM has thought through carefully what it means to instrument an organization (for example, the need to instrument and to to set up software 'sensors' that measure energy use, company-wide) and then use that information to drive reductions in consumption. This is the major initial thrust of any emission-reduction strategy. Without going too much into particular elements of the initiative, it is notable that IBM:

- considers the role of asset management
- ensures visibility of energy management at the local/department level
- includes trend analysis
- aims to improve space utilization
- seeks to switch to renewable energy sources where available
- optimizes HVAC for current weather predictions.

Moreover, the company partners with others in the Green Sigma coalition. This is an industry organization dedicated to delivering sustainable building, smart grid and monitoring solutions across a wide range of industries and the government sector. IBM also considers the political aspects of this effort. I can say, this sustainability initiative is well targeted and well thought out.

Finally, we should note that IBM has 'walked the walk', or 'eaten its own dog food' — if you prefer, in sustainability. Its citation of "having avoided carbon emissions by an amount equal to 50% of our 1990 emissions" is impressive in its own right.

The effects

Fairly, or unfairly, carbon emission reductions focus on two somewhat separate issues:

- reducing carbon emissions within an enterprise itself
- lowering emissions from the products that a company creates.

This includes just about everything that generates emissions that are used, administered, or produced by a company. There are buildings (factories and offices), direct energy consumption (heating, cooling and lighting), transportation (cars and trucking), entertainment and, of course, all the computing assets.

Buildings, as IBM notes, are a large part of emissions generation. Unlike cars and airplanes, buildings can relatively easily achieve much greater energy efficiency with a much shorter payback period. That means that a full implementation of building energy improvement across the world would typically lead to at least a 10% decrease in the *rate* of *human* emissions (please note the italics; I will explain later). It is hard to imagine any approach delivering a greater immediate impact.

IBM's emphasis on measurement is, however, likely to have a far greater impact in the long run. The fact is that we are not yet completely sure how to break down human-caused carbon emissions by business process or by use. Attempts to reduce them are, therefore blunt instruments, often hitting unintended targets or are akin to squashing flies with a sledge-hammer. That said, full company instrumentation, as well as full product instrumentation, would support major improvements in carbon-emission-reduction effectiveness, not just in buildings or data centers but across the board.

Overall, IBM's sustainability announcements paint a picture of major improvements in energy efficiency leading — very optimistically — to 30% improvements in energy efficiency and increases in renewable energy over the next 10 years. Sadly this is well beyond the targets of most of today's nations when seeking to achieve a 'moderate-cost' ultimate global warming of 2 degrees centigrade — in the best-case scenarios.

In effect, initiatives like IBM's plus global government efforts could reduce the rate of human emissions beyond existing targets. Meanwhile, Lester Brown from the Earth Policy Institute has noted that from 2008 to 2009, measurable US human carbon emissions from fossil fuels went down just 9%. This should be good news, but I find that it is just slightly less bad news.

Everybody suffers

Everyone who is trying to do something about global warming has been operating under a set of conservative scientific projections that, for the most part, correspond to the state of the science in 2007. As far as I can tell, here is what has happened since then…

Sea rise projections have doubled to a 5 feet of rise in 80 years. In fact, more rapid than expected, land ice loss means that 15 feet of rise may be more likely, with even more after that. Scientists have determined that 'feedback loops' — such as loss of the ability of ice to reflect back light and therefore decrease ocean heat — in turn increase global temperature and are in fact 'augmenting feedbacks' (meaning that they will contribute to additional global warming even if we decrease emissions to near zero right now).

Carbon in the atmosphere is apparently still headed towards the 'worst case' scenario of 1100 parts per million (ppm). That, in turn, apparently means that the 'moderate effect' scenario underlying all present global plans for mitigation of climate change with moderate cost (450 ppm) will in all likelihood not be achieved (each doubling of ppm leads to 3.5 degrees centigrade or 6 degrees Fahrenheit average rise in temperature, and in many cases, more like 10 degrees Fahrenheit in summer). The start level was about 280 ppm. So we are talking about a 12 degrees Fahrenheit rise if we reach 1100 ppm, with follow-on effects and costs that are linear up to 700-800 ppm and difficult to calculate but almost certainly accelerating beyond that.

There is growing consensus that technologies to sequester atmospheric carbon or carbon emissions in the ground, if feasible, will not contribute for at least 5-10 years and not reach full effectiveness until 5-10 years after that. Even then these will not able to take us back to 450 ppm for many years thereafter — and not able to end the continuing effects of global warming for many years beyond that, if ever.

What about that 9 % reduction in emissions in the US?

The 9% reduction throws up three problems:

- first, that this 9% reduction was achieved under conditions in which GNP was mostly going down; as we reach conditions of moderate or fast growth, that reduction goes to zero
- second, most of the reductions achieved up to now come from low-cost-to-implement technologies; that means that achieving the next 9%, and the next 9% after that, becomes more costly and difficult, both financially and politically, to implement
- third, at least some of the reductions come from outsourcing jobs and, therefore, plant and equipment to faster-growing economies with lower costs but greater 'energy inefficiency'.

As a result, even where IBM is applying energy efficiencies to its own sites, the follow-on jobs outside of the company's bailiwick are typically less energy-efficient, thereby reducing the net worldwide effect of US emission cuts. As noted above, the pace of worldwide atmospheric carbon dioxide rise continued unabated through 2008 and 2009. Reducing the rate of human emissions is not good enough. We have to

reduce the absolute amount of human, human-caused (like reduced reflection of sunlight by ice) and follow-on (like melting permafrost, which in the Arctic holds massive amounts of carbon and methane) emissions.

Can people adapt to adaptation?

That leaves adaptation to what some scientists call 'climate disruption'. What does this mean? How will we live with a rise in sea level of 15 feet in the next 60 years and an even larger rise in the 60 years after that? How will we survive natural, weather-related disasters that are 3-8 times more damaging and costly than they are now, on average? (A very rough calculation, based on the scientific estimate that a 3% C temperature rise doubles the frequency of category 4-5 hurricanes.)

Adaptation means:

- adjusting to the loss of food and water related to ecosystems that cannot move north or south, blocked by human paved cities and towns
- relocating to lower-cost areas, or constantly revising heating and cooling systems in the same area, because the amount of cooling and heating needed in an area will change drastically
- moving food sources from where they are (in response to climatic changes where some areas become better for growing food and some existing ones worse).

Adaptation may mean moving 1/6th of the world's population from 1/3rd of the world's arable land because of desertification. In other words, much of this adaptation will affect literally all of the global population, and the costs of carrying out this process will fall to some extent on everyone, no matter how — rich or poor. And we are talking about an adaptation that, according to recent research, appears already to be 'baked into the system'. Moreover, if we continue to be ineffectual at reducing emissions, each decade will bring additional adaptation costs on top of what we are already obliging ourselves to pay.

Adaptation will mean significant additional costs — because climate disruption affects everyone in their personal lives. It is hard to find a place on the globe that will not be affected by one or more of floods, hurricanes, sea-level rises, wildfires, desertification, heat increases (which will make some places effectively unlivable), drought, permafrost collapse or loss of food supplies. Spending to avoid these for one's own personal home will rise sharply — well beyond the costs of 'mitigating' further climate disruption by low-cost or even expensive carbon emission reductions.

What does IBM need to be doing?

Obviously, IBM cannot do much about this by itself. I would, however, suggest two further steps for the company (and other vendors and organizations) to consider. First, it is time to make physical infrastructure agile. As the climate continually changes, the feasible or optimum places for head offices, data centers, and residences will endlessly change, as well. It is time to design workplaces and homes that can be inexpensively transferred from physical location to physical location. Moving continually is not a pleasant process to contemplate; but virtual infrastructure is probably the least-cost solution.

Second, it is time to accept limits. The effort to pretend that we do not need to reduce emissions in absolute, overall terms, because technology, economics, or sheer willpower will save us (as we have practised since the first warnings of global warming arrived in the 1970s) is failing badly. Instead of talking in terms of improving energy efficiency, IBM needs to start talking in terms of absolute carbon emissions reduction every year, for itself, for its customers and in its products — no matter what the business' growth rate is.

One more (relatively minor) point: because climate will be changing continually, adjusting HVAC systems for upcoming weather forecasts, which only go five days out, is not enough. When a place that has seen four days of 100 degree weather every summer suddenly sees almost 3 months of it, no short-term HVAC adjustment will handle continual brownouts adequately. IBM needs to add longer term climate forecasts to the mix.

Alas, and I mention this only reluctantly and in the certain knowledge that for some this will devalue everything I have argued up to this point, there is every indication that without effective cooperation from governments, the sustainability goal that IBM seeks, along with the avoidance of catastrophes beyond what I have described here, will not be achievable.

IBM's membership, therefore of an organization like the US Chamber of Commerce which actively and preferentially funnels money to candidates and legislators that deny the scientific consensus about global warming (and its serious effects) undercuts its sustainability initiative. It causes serious damage to IBM's credibility and brand.

Management conclusion

IBM deserves substantial credit for creating what is one of the most comprehensive and effective efforts to tackle the climate disruption crisis as it was understood in 2007. But that effort, and its attendant goals, is three years out of date. Plus its support of organizations that fund climate change denial candidates is misplaced.

From where the world stands today, IBM needs to use its previous efforts as a starting point for creating new solutions within the next year, solutions that are aimed at a far bigger task: tackling the climate disruption crisis not as it was but as it is today (in 2010/2011). That effort, at least, might result in a sustainability initiative adequate to meet IBM's customers' (and the world's) long-term needs.

Wayne Kernochan
Infostructure Associates
www.home.comcast.net/~kernochan
w70

Defining IT state: a simple approach

Peter Bye, Bye Associates

Management introduction

In order for an organization to improve its overall IT performance, it is essential to understand the current state of the IT environment and its desired future state. These two states will rarely be the same, as most organizations can always see room for improvement in cost/performance. In fact, in today's business climate, CIOs are likely to be compelled to find improvements, whatever the current state.

By knowing the current and desired states, steps can be proposed to close the gap between the two. The desired state is likely to move over time, so the process has to be repeated with a new desired state. Ideally, there should be a process of continuous evaluation.

To understand state, it is necessary to be able to describe it. Unfortunately, many ways are complex, based on detailed evaluations of the current environment, business requirements of IT and many other factors. In this analysis, Peter Bye presents a simple approach for describing state, based on the infrastructure (primarily the software ecosystem) within which applications exist. This approach does not require special tools. Instead it aims to stimulate discussion between the relevant stakeholders, both business and IT management.

The state model

A major challenge facing CIOs for 2011 is how to deliver new IT services as and when the business needs them. IT agility is the term often used which raises questions about how organizations measure agility.

Systems increasingly have to co-operate with other systems, both within the data center and outside, to partner organizations. Applications become components in wider distributed applications, with existing applications and databases being re-used in new

ways, thereby increasing agility and responsiveness. For many the best way to achieve the desired levels of re-use (and hence agility) is for systems co-operate within a common infrastructure, most commonly by:

- building on Service Oriented Architecture (SOA) principles
- exploiting standardized technologies, such as Web services.

Infrastructure, however, is not just about connecting systems together; the intended result needs to be resilient. Resilience is concerned with infrastructure stability, availability, security and the ability to withstand shocks (whether as unexpected surges in demand or equipment failure).

Providing resilience has, in turn, become another major challenge for CIOs. High levels of automation — to handle both routine and abnormal conditions — are necessary. This is especially true when applied to failures, when IT service continuity must address everything up to and including disaster recovery.

Defining a 2 dimensional space

Integration as a measure of agility and automation as a measure of resilience together form a two-dimensional space within which one can position any IT environment. This approach can also be confined to parts of an IT environment — for instance a group of systems delivering retail banking — or to the whole IT infrastructure.

Figure 1 shows such an integration and automation model:

- the horizontal axis represents the level of integration, ranging from no integration at one end to a fully developed Service Oriented Architecture at the other
- the vertical axis shows the level of automation, extending from no automation at all to a fully-automated real time infrastructure (RTI).

IT professionals should expect integration and automation to advance approximately in step with each other. This is as shown by the diagonal arrow 'A' (in Figure 1)

Few, if any, organizations are at the extreme ends of either axis. Most are somewhere between the two. In general the CIO's drive is to move the environment in the direction of arrow A. But if, for example, the environment has arrived at B, there is then a clear indication of a lack of automation: corrective action is then required to move closer to arrow A.

Calibrating the model

Although Figure 1 is helpful in showing relative position, some form of calibration of the two axes is desirable (and necessary) in order to gain a clearer understanding of the characteristics of a given state. This is as per Figure 2 (in which ignore the 'CIO view' and other points for the time being).

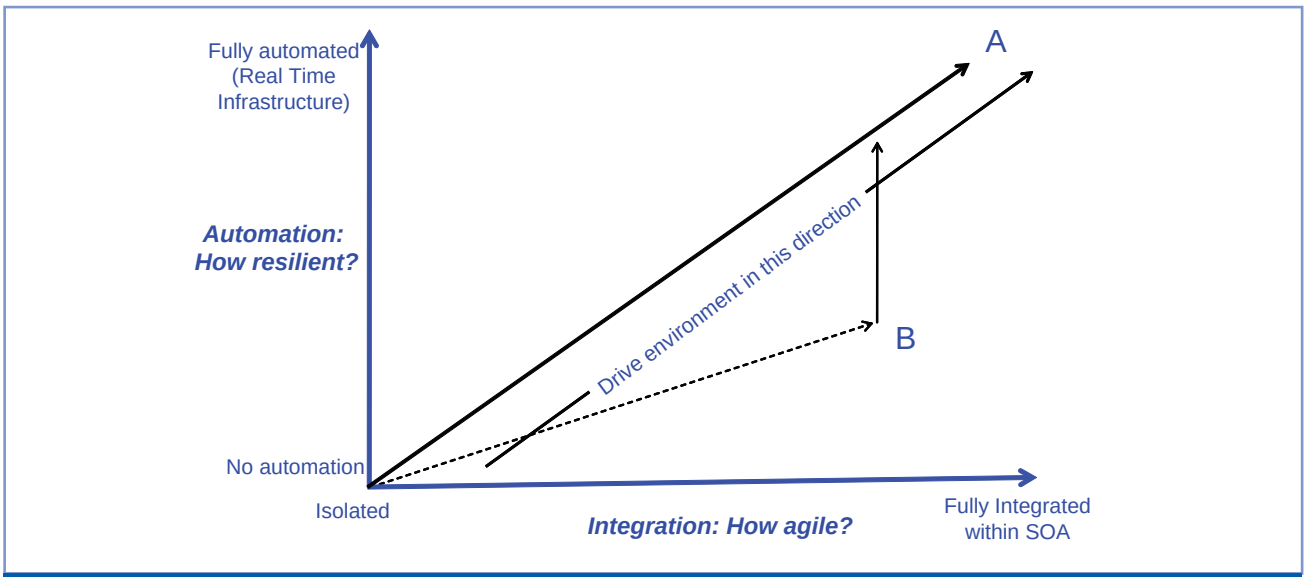

Figure 1: A model for assessing the current and the future

There are five integration stages, shown as bands covering a subset of the whole spectrum (these do not define a fixed state, because integration levels can increase within a band). The three bands on the right show the major stages in the generally-accepted SOA maturity model.

There is then the need to calibrate automation. Figure 2 shows five bands of increasing automation. Again, each band represents a subset of the total spectrum, so automation levels can improve while staying within a band (and Figure 3 provides more information on the characteristics of the calibration bands).

Using the model

This model was originally designed for IT service and product suppliers engaging with clients, and today is being used for that purpose. Understanding current and futures states is a prerequisite for proposing services and products to close the gap between identified states.

While the model is usually used within a larger assessment service, which consists of an interactive workshop to gather the required information and then to generate a report with recommendations based on the results of that workshop, it can also be used internally. Within an IT services supplier it might become a part of account planning for that client.

Practice shows that assessments reveal 'interesting' disagreements within organizations. The various points marked on Figure 2 show a company's CIO's view and the views of heads of department. Note the gap. (The ellipse marked X represents the external consultants view; the target state is also shown.)

IT service providers can learn much from a client's IT environment positioning (or a client could reverse apply this to understand on what a service provider is actually focusing). For example, a client at a low level of integration but high up the automation axis may indicate that the service provider is spending a disproportionate amount of time with a client's operations staff and not enough with the application developers. This may not be a problem — because a high level of automation is generally desirable. Yet it is generally better to be above arrow A (in the Figures) than below it.

Irrespective, such a positioning exercise should, at the minimum, trigger a review of where to spend time. Indeed, although this model was originally designed for the purposes described, it has mutated. In at least one instance it is being used internally by a service provider's client — where the CIO uses it to describe the current IT state when reporting progress to the board of directors.

Management conclusion

The model has already been proved effective for both clients and IT providers in improving the understanding of their IT environments. This simple approach, based on its two-dimensional model, is easy to under-

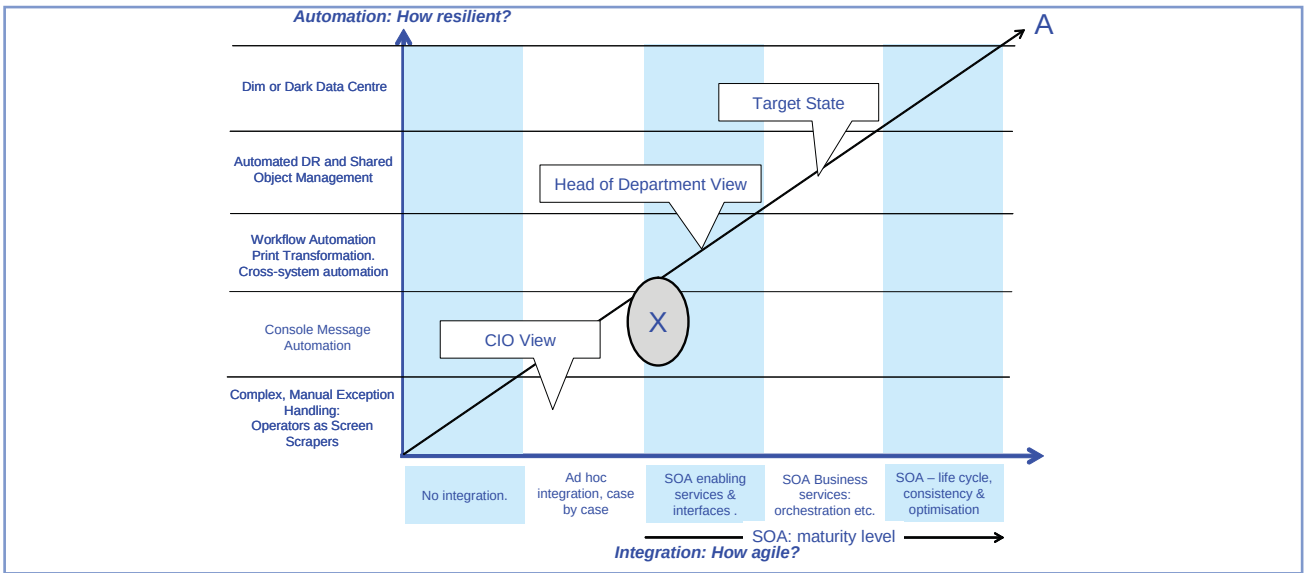

Figure 2: Calibration of integration and automation

stand and becomes valuable as a discussion aid (for service providers) for engaging with clients and internally to help with discussion and reporting. Its additional value is like the examples discussed — it can reveal 'interesting' differences of understanding.

That said, like all models, it is a simplification. It does not address, or try to, all aspects of the IT environment. In particular, the model's primary concern is with software infrastructure. Its focus is clearly on the way applications integrate and are managed — and it is not concerned with what applications do for the business, just how well or otherwise they do it. As Mr. Bye describes, it is more likely to throw up requirements for additional investigation which might not otherwise have occurred.

Peter Bye
Bye Associates

Integration band	Characteristics
No integration	Only connections are terminal/workstation networks No system to system connections, apart possibly from data extraction for MIS purposes
Ad hoc integration	Some internal system to system connections in place Some external connections are in place Each new connection is handled as a one-off, implemented as required No overall architectural approach
SOA: enabling services	There are three SOA stages in the SOA maturity model; enabling services is the first: • There is an architectural vision, based on SOA • Application functions are exposed as services, using standardised interfaces • Applications components invoke each other through service requests • The organization is starting to consider or experiment with Enterprise Service Bus (ESB) and other technologies
SOA: business services	The SOA implementation is now progressing to expose composite business services: • Business modelling • Aggregation/orchestration • Workflows • Complex event processing • Compliance
SOA: life cycle, consistency and optimization	In this stage, the SOA is moving towards completion: • Life cycle enabling: define road from design to production, version control, rules (e.g. mandate re-use, review, policies, metrics), roles (e.g. designers, developers, service champions) • Defined, measurable quality of service, service level agreements and key performance indicators • ITIL, ROI and TCO scenarios • Alerting, consolidated logging errors and events
Automation band	**Characteristics**
Complex, manual exception handling: operators are 'screen scrapers'	• Operators respond to most or all console output, even if routine • All exceptions are handled by operators, using only written procedures • Any disaster recovery (DR) is manually handled • Manual performance and security monitoring and management
Console message automation	• Responses to console output are now automated for routine cases • Exception handling at the console level is automated where possible • Some automation may be used for DR, as part of automated exception handling
Workflow automation, print transformation, cross system automation	• Workflow is now automated, for batch jobs and other processes • Output management and print transformation are implemented • Automation crosses systems for workflow and other processes such as exception handling
Automated DR and shared object management	• DR is now fully automated, although intervention may be required for approval • Shared object management, for example tapes, is implemented
Dim or dark data center. Fully automated, to provide real-time infrastructure (RTI)	This stage represents a real time infrastructure (RTI), with fully automated and policy-based management: • Data centers run unattended • Automated, policy-based responses to performance, security and other exceptions

Figure 3: Characterists of calibration stages for integration and automation

Value-based case study #9

International services provider cuts costs, creates new revenue streams while improving service delivery

Charles Brett, C3B Consulting

Management introduction

Imagine you are an IT service provider to a financial institution like a brokerage. It is 10 minutes to 4PM (when the markets close) and a $40M transaction has gone missing. If the trade is not completed, the brokerage will not only be liable for regulatory fines but also for interest foregone and, potentially, any price change that occurs the next day. It will hold its service provider liable if it proves to be an IT issue that has caused the problem.

This is a scenario that is all too common. The financial implications are unpleasant. The customer relationship implications are, if anything, worse.

This case study — which reaches across industries like aviation and finance in which sophisticated managed service providers operate — shows how it is possible to break out of the confining jail to which with much of operational IT seems to operate. The outcome of the (true) scenario above will be shared later but before that, consider what happened at a major international airline.

Traditional systems management can be insufficient

This airline; like many others, contracted out much of its IT operation to an international managed services provider. Amongst the 'assets' that became the responsibility of this service provider was an extensive network of IBM's WebSphere MQ (WMQ) running on anything from zOS mainframes to HP's nonStop servers to many flavors and instances of UNIX, Linux and Windows. All these WMQ instances provided mission critical support to, for example:

- reservations, to communicate rate and fare changes to travel agents
- aircraft operations, in the calculation of load factors (for this applications adjust almost in real time as passenger numbers, weather conditions, cargo and fuel loads all continue to change up right up to departure).

In these two situations availability and reliability are critical both to commercial operations (sales, via the travel agents) and to safety for the airline. The service provider knew this when it accepted contract responsibility.

In the past the airline had introduced BMC's Patrol for systems and network management. This had worked reasonably well on its larger systems. But, for management of many of the distributed ones, it was inconsistently applied, not applied or not available. Indeed, some of the WMQ server platforms were being monitored by non-Patrol solutions, or not being monitored at all.

There was, therefore, a broad mis-match, as far as the service provider was concerned, about how it was to be able to meet its service obligations. In addition, there were serious reservations about aspects ofthe way Patrol worked — for Patrol requires the installation and use of software agents on the platforms it manages. For the service provider using such agents possessed four significant downsides:

- rolling out agents takes time and skilled effort for their successful installation
- software agents necessarily alter each system on which they run, with attendant risks
- upgrading or changing systems where such agents are installed is cumbersome but, worse, using agents mandates extensive retesting to ensure that all is working as expected
- there were thousands of WMQ end points but each Patrol management instance could only handle 250 nodes at that time (a legacy from fewer, larger systems days).

The service provider decided that this situation was too expensive to continue if it was to fulfill its obligations to the airline. It decided to investigate alternatives, and not just related to WMQ but also for Tomcat's ActiveMQ (JMS messaging) which was increasingly being adopted.

Making a choice

The systems provider examined multiple alternatives, including IBM's Tivoli, MO71 (an Explorer-like WMQ SupportPac) and even Nastel. None really satisfied the requirements in the way the service provider felt it needed, especially:

- the scalability to handle thousands, not hundreds, of WMQ and Active MQ nodes
- operational flexibility
- no agents
- the combination of maximizing automation where possible but also improving operational collaboration where automation would not solve everything.

The service provider's objective was, therefore, to reduce costs — in two forms: reducing daily operating expenses while also minimizing the investment required when making changes.

As part of the investigation of alternatives the services provider talked with one of the large airline reservation companies. The latter had faced a similar situation and in its own researches it had come across a product called Infrared 360 (IR360) — developed by Avada Software. From the reservation company the service provider learned that IR360 was apparently capable of high degrees of automation for managing WMQ resources. This included being able to access an WMQ manager even if it had failed — using the IR360's SOAP listener module (which enables remote restarting, reinstalling, definition and log downloading — all occurring with security preserved).

The service provider then undertook its own research into IR360. The first obvious attraction was that IR360 is agentless. It could also be installed in parallel. One of the issues with adopting Tivoli, for instance, was it would mean de-installing all the Patrol agents, then installing Tivoli agents (for these could not be run in parallel) — and then retesting everything again. In contrast, IR360 could be installed, run in parallel with Patrol and its agents and, only if IR360 was proven, would Patrol then be removed.

A second immediate attraction, again because IR360 is agentless, was that the roll out to all platforms — from zOS to Linux or Windows — was uniform. There would be no discrepancies or differences in WMQ management tools or approach. In part this was delivered because IR360 is a J2EE application (Figure 1) acting as a portal which installs and runs on any App. Server (WebSphere, Apache, JBoss, etc.) as well as work with most databases (DB2, Oracle, SQL Server, Informix, etc.). This portal approach would then link to (and from) all the instances of WMQ. Nothing would need to be deployed out into the field instances of WMQ. Furthermore, updating or upgrading IR360 itself would be as simple as dropping a new WAR file onto the chosen IR360 application server: there would be none of the complexity and cost associated with Patrol and the previous WMQ management approach.

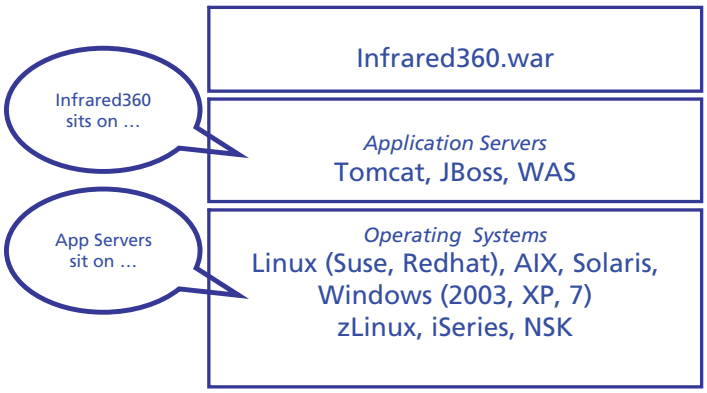

Figure 1: WAR files runs on an App. Server

A further attraction of this approach was that it could run on and exploit existing App. Servers in the airline. This meant that scalability (and other positives like failover and disaster recovery) could be 'inherited' from the airline's already robust App. Server environment without additional work.

The next attraction was on the manager side: everything is managed through a browser (IE, Mozilla). The practical implications of this were considered far reaching because it would:

- remove the need for fat clients to be deployed on designated management workstations as well as the need for specific security to be applied to these designated management workstations
- extend the principle of management consistency across the whole of the WMQ 'estate' (which had not been possible before)
- mean that access to the WMQ environment would available from anywhere specialists were located or where they needed to be
- improve flexibility both for employees and their tasks.

For this services provider these attractions were compelling. It had signed the management services deal with the airline. Any costs — like those described above — that could be taken out (of the cost of providing the services to the airline) would drop straight through to the service provider's bottom line. These were sufficient reasons to decide to introduce IR360 (and practice has since proven these to be accurate).

Practice proves to be beyond effective

IR360 installed smoothly and in parallel, as planned. It started to produce the anticipated benefits. At the same time, as its capabilities were better understood and as practice made the services provider more effective in its use, further abilities emerged that had not been sought or expected. These 'extras' included:

- improved management of alerts
- new service and revenue opportunities
- better QA and SLA delivery
- secure delegation of user authorities
- automated problem fixing.

Take, for example, the first of these — management of alerts. Alerts are the bedrock of systems management. When something goes wrong or a threshold is crossed (say that a WMQ queue depth had exceeded a given percentage, an alert is issued to (say) Patrol, Unicenter or a Tivoli Console — so that someone investigates and remedies the outstanding issue. In principle this works well. It does, however, possess a downside. Because the alert can go a large number of people it can cause one of two effects:

- the first is that nobody does anything (thinking that someone else is dealing with it)
- the second is where too many people become involved — and too many are 'contributing' to the resolution (in the experience of the service provider this was extremely wasteful of precious skills and resources).

Using IR360 the service provider discovered it could generate cascading alerts that would only go to the people that mattered. For example, if the WMQ depth exceeded 20% and appeared unthreatening, then this would be dealt with at a routine low level of responsibility. If, however, the queue depth continued to rise to (say) over 30%, then the new alert could be escalated to management within the service provider with the authority to ensure resolution. If it then exceeded (say) 50%, a broader warning might be sent out, including to the airline so that the latter might knew in advance that there might be a problem. The net effect of exploiting this approach was that SLA observance improved, as well as relations with the airline: the latter could prepare rather than react to a problem.

Unexpectedly, IR360 threw up the possibility of new revenues for the services provider. For example, in non-production areas — like development, testing and QA — there was a constant need for clearing out of queues and resetting queue managers as part of daily activity. Previously there had been no automated tools to accomplish this. Using IR360 the service provider created a new income stream by selling access to the control service to do what had not been possible before. Interestingly this was delivered from the same portal that ran the WMQ production systems, because IR360 possesses the security segmentation to enable the various different environments (development, testing, QA and production) to coexist but still remain distinctly and securely separate.

A further benefit came with test harnesses and message generation (IR360 is script-less). These can be

prepared in advance, for example for initiating a new WMQ instance with the correct queue manager and queue values. These same harnesses/message formats and content can, however, be re-used. The services provider is now using this capability along with the IR360 scheduler to do regular automated re-certification. For example, at the end of the month at the airline there was a procedure to empty queues and restart them. Before, this was at best semi-automated. Now it is fully automated.

It goes further, however, by including a retest of the communications and infrastructure as well as bringing back up the 'refreshed' WMQ resources — all without user intervention. This has not only become part of ongoing QA; it now provides data points and reporting to the airline about how SLA commitments are being met. Indeed, so extensive has this function become it is now run on a backup portal (but still using the common database) so as not to impact the operational portal plus the airline has access so as to be able to see in real time what is happening within its own WMQ estate (thereby removing another area of people involvement).

Related to this was the issue of 'secure virtual delegated authority' (Figure 2) and the opportunity to improve collaboration across the service provider's own business. Most IT shops have multiple logical separations of activities (the most obvious are those related to development, testing, QA, operations, etc.). These tend to be distinct and to exist as their own microcosms, with minimal 'cross-border' communications between the people responsible in each. This service provider discovered that IR360's security approach facilitates collaboration between logical groups, yet without prejudicing security. The result has caused barriers to reduce. For example, at the airline this has even been added to production environments themselves where certain applications — such as Cargo, Crew, Reservations, Fares, etc. — involve distinct separation of duties but also require that both support and business units be able to collaborate in order to resolve problems.

Perhaps, however, the most impressive extension for the airline is the automated fixing of problems. The service provider now uses 'compound alerts' to anticipate and resolve issues. For example, a queue may need to be restarted; it may not initially be apparent that the source application or service at the far end has also failed and needs restarting. The traditional solution sequence was to solve each issue as the operator encountered it. Using IR360 the service provider now defines in advance 'compound alerts' which can then complete the whole correctly sequenced restart process (queue manager, channel, queue, application, etc.). Using the services interface within IR360, there is now no user intervention if the criteria for a given compound alert are satisfied. The mean time to fix failures:

- has dropped from a minimum of 2 hours to minutes (or less)

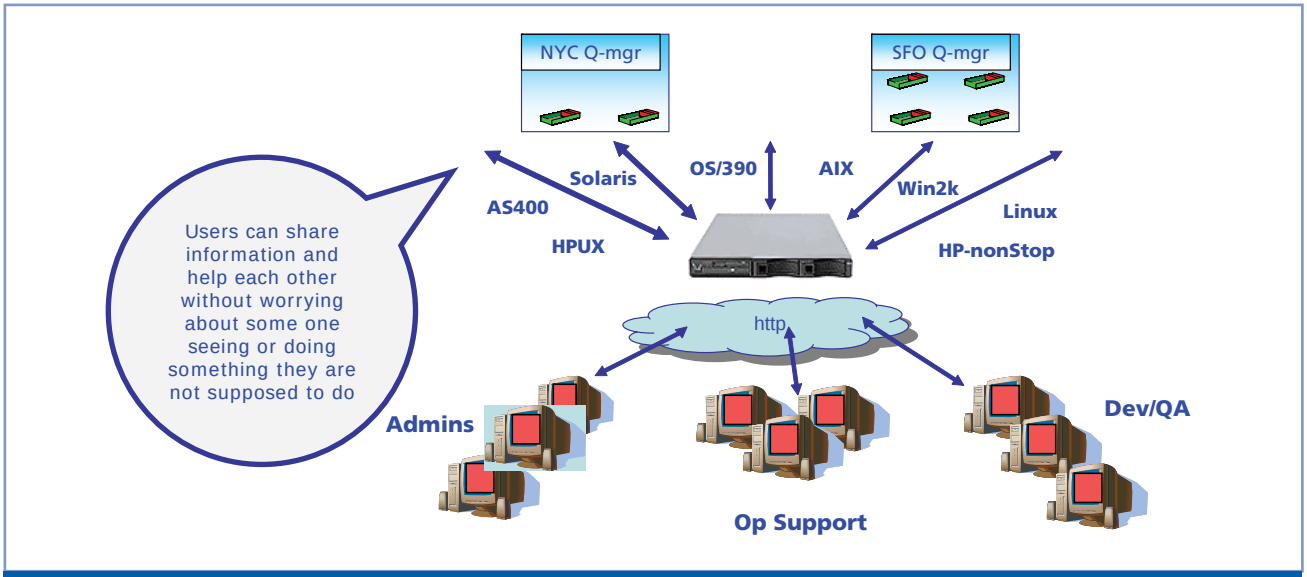

Figure 2: Collaboration which preserves security

- is accurate
- generates a notification of what has happened — for the reporting described above.

This continues to save a fortune in time. It also continues to improve the service to the airline.

Business impact

For this service provider the business impact of IR360 has been extensive, and goes well beyond the airline example above. The capabilities of IR360 have started to generate wholly new opportunities and revenue streams.

Sales teams are, today, able to be much more responsive to potential new customers. In the past the service provider would take weeks if not months of learning and preparing before it could start to introduce its management and techniques, and before it would accept responsibility. This delay is no longer necessary. Because IR360 is non-invasive, it can run in parallel. The effect, in the sales instance, is that the service provider can start to work earlier at prospects and in so doing convince potential customers that it can deliver. This has proven to be a powerful sales tool (and another unanticipated benefit).

While the airline support was focused on WMQ as its principle objective, this service provider is now looking to extend the same IR-360-based conceptual approach to App. Servers, whether WebSphere, Oracle, Tomcat, JBoss or others. There is no reason not to do this, and the limits may not stop there. Customers are already asking about the relevance of IR360 to managing TIBCO infrastructure and even PeopleSoft (and similar) applications.

Back to the missing $40M transaction

Finding a missing transaction with only minutes to go before the market closes is not simple in today's complex financial systems. If the transaction is not where it is expected, then it could be anywhere (just like your car keys, except that transactions are intangible and exist within an equally intangible infrastructure).

One attribute of IR360 is that it can search data within a message (beyond just the transaction ID, if this is even known) — for example looking for the name of the brokerage's customer, the destination account or even the transaction amount. Using this capability a search for the missing transaction was started. Within seconds it was found, its header was modified, it was resubmitted and it was processed before the market closed.

Where was that transaction? It was in the development environment, sitting on a development queue. Why? Someone had forgotten to change queue manager designations (from Dev. to Production) and so the $40M trade had gone to the wrong place. It happens. It happens all the time, if not so often for a combination of such a large amount just before the markets close. There was extreme relief all round.

Management conclusion

The airline that this service provider supports has obtained substantial operational benefits through the introduction of IR360 — from improved reliability through to improved SLA observance and vastly better insights into what is happening before it happens. For the service provider, deployment has not only enabled it to meet its obligations but it has been able to go further — reducing costs, opening up new revenue streams as well as embracing capabilities that can attract new customers.

The importance of the agentless approach continues to prove itself. Not only is it non-invasive (which matters to most large IT organizations that fear to 'alter' anything without extensive retesting) but it is flexible and can run in parallel. With an application that runs on common App. Servers, this service provider has simplified its operations while improving its quality. Its business model, and its customers (like the airline), now depend on IR360. To remove it has apparently become unthinkable (and could only be for something that was sigificantly better again).

Charles Brett
C3B Consulting
www.c3bcconsulting.com

INSIGHT-SPECTRA

is published and distributed worldwide by:

C3B Consulting Ltd.
19 St. Michael's Road
Winchester SO23 9JE
UK

Telephone:
+44 787 233 4000
+34 686 116 993

Email:
insight-spectra@
insight-spectra.com
or
charles.brett
@c3bconsulting.com

World Wide Web:
www.insight-spectra.com