quocirca

Business and IT Analysis

The new virtualised data centre

This report has been created by drawing together a series of 12 articles Quocirca wrote for SearchVirtualDataCentre.co.UK. The articles look at the various aspects of creating a strategy for a flexible and effective data centre aimed at providing a platform to carry the business forwards.

January 2011

Old-style, one application per physical server data centres are not only nearing the end of their useful lives, but are also becoming barriers to a business' future success. Low resource utilisation rates, high energy costs, lack of flexibility and information siloes are forcing businesses to look again at what the data centre function is. Virtualisation has come to the fore – yet it also creates headaches that data centre and facilities managers must deal with.

Clive Longbottom
Quocirca Ltd
Tel : +44 118 948 3360 ext 200
Email: Clive.Longbottom@Quocirca.com

Rob Bamforth
Quocirca Ltd
Tel: +44 7802 175796
Email: Rob.Bamforth@Quocirca.com

The new virtualised data centre

This report has been created by drawing together a series of 12 articles Quocirca wrote for SearchVirtualDataCentre.co.UK. The articles look at the various aspects of creating a strategy for a flexible and effective data centre aimed at providing a platform to carry the business forwards.

Virtualisation is here to stay	The need for virtualisation is not particularly a technological one - it is far more based on the needs of the business. Flexibility, greater control over costs and greater responsiveness to the business' needs means that virtualisation is increasingly being used across greater proportions of an organisation's IT estate.
Virtualisation is not a direct replacement	A virtualisation strategy has to be carefully thought out in order to maximise the benefits that can be gained. Just attempting to take existing IT and move it to a virtual platform will not work - far more planning and work is required to ensure that virtualisation works.
Vendors have not been massively helpful	Many virtualisation messages have been hijacked by vendors who do not fully understand virtualisation themselves. Many software vendors have failed to move their architectures over to fully utilise virtualisation, and resellers and service providers have also been seen to be guilty of replacing silos of functionality with silos of virtualised functionality for many of their customers.
It's not just the technology that is wrong	Software and operating system licensing models have to change to make virtualisation work, and yet Quocirca still sees many instances of per cpu/per core licensing, or other approaches that make full virtualisation difficult (or impossible) to achieve and manage.
The data centre fabric also has to be addressed	Virtualisation brings with it higher densities of hardware, with increased heat generation. Existing approaches to power distribution and cooling will not meet virtualisation's needs. A complete mindset change is required to move away from looking at cooling down the whole data centre to just ensuring that critical components are kept within their design limits. Better structured cabling and data centre zoning will be needed to keep up with best practices.
IT and facilities must work together	A modern data centre cannot be built with IT and facilities working in a vacuum. Each group has to work closely with the other to ensure that power distribution matches the needs of the denisty of IT equipment, and that cooling is adequate, yet is fully optimised to ensure that energy costs are kept under control.
Don't forget the future	Today's internal data centres - whether virtualised or not - will have to interoperate with external functions and data centres as cloud computing and other architectures gain a stronger foothold. Failure to plan for such hybridisation of the IT world will lead to the need for yet another data centre re-architecture - with a solid financial hit on the organisation.

Conclusions

For organisations looking to embrace virtualisation, much planning has to be undertaken. Virtualisation can offer a great deal in providing a flexible and cost-effective platform for supporting the business – but can also create a new level of chaos and cost for those who get it wrong.

The virtual data centre – it's a jungle out there

If 2009 was the year organisations started taking virtualisation seriously, 2010 is going to be the year when pan-data centre virtualisation becomes mainstream. Driven by the need to save energy, space and human resources, physical IT platforms that are rapidly becoming non-competitive at the business level due to inflexibility, non-resilience and high capital and operating costs will need to adopt virtualisation to provide the needed competitiveness in the market.

However, although virtualisation at the server, storage and network level is pretty well proven, Quocirca has found that many virtualisation projects are out of control and are seen by the business to be failing. This series of articles will try to help in identifying those areas where Quocirca has seen users struggle with virtualisation, and will aim to provide pointers for data centre managers to consider, helping to avoid the major pitfalls.

Ten areas that Quocirca will expand on over the coming months are:

1) Getting the architecture right. Virtualisation is not just about making everything into a single resource pool: it's about virtualising where it makes sense, using existing and new assets in the best possible way. Heterogeneity is possible - different workloads need different hardware.

2) Keeping the data centre dynamic. Next generation data centres need to be able to grow and shrink with the needs of the organisation. Building a 100,000 square foot data centre where only 40,000 square foot is to be used for the first few years does not make economic sense - nor does building one of 40,000 square foot, when 45,000 square foot will be required in a year's time - unless you can look at renting out the excess space as platform as a service (PaaS) or colocation data centre space.

3) The big enemy: the application. Hypervisors make server virtualisation easy. TCP/IP makes the network inherently virtual. Storage virtualisation is proven. However, many enterprise applications are still ill-architected for true virtualisation: the flexibility of having multiple small instances of an application, of being able to roll it out across an estate of unknown physical hardware and the capability to only use what functionality you need is still some way off being supported by most vendors.

4) Keeping control of virtual images. One of the biggest issues with virtualisation is that it is so easy. A developer can create a new image of operating system, application server, application and so on, spin it up in seconds, carry out some work and then spin the image back down again. However, the image still needs to be maintained - patches, upgrades, security and so on all need to be applied, and if you have thousands of images, then this can become a major ongoing task.

5) Maintaining control of software licensing. The same virtual images that may need continuous patching and upgrading are also eating up licences. An image that has been used once, but has not been deconstructed, is still using up an operating system licence, as well as an application server and application licence. Dynamic licence management is required, using libraries and check in/check out capabilities.

6) Optimising energy usage. Virtualisation should lower immediate energy needs just through server consolidation. However, there is more that can be done to drive energy usage much lower. Higher data centre temperatures are viable, targeted cooling and technologies such as hot aisle/cold aisle and heat pumping can all help in driving energy costs down.

7) Separating data and power pathways. Unstructured cabling in any data centre is bad practice, and in the virtualised data centre becomes a real issue due to the increased density of mission critical assets. By structuring

the cabling, not only are air pathways maintained and cleaning made easier, but data transmission issues caused by running power too close to poorly shielded or unshielded data cabling can be avoided.

8) The dark data centre. Data centres are for machines and for bits and bytes. They should not be designed for humans, and the majority of data centres should be running in an essentially lights out environment, with automated monitoring and eventing occurring to administrators outside of the data centre itself.

9) The death of the application. Massive applications should be seen as the dinosaurs they are rapidly becoming, and organisations should be looking at far more dynamic composite applications built from aggregated functionality. Web services and SOA help this - and cloud computing will move things even further forward.

10) The hybrid cloud. Much has been said about cloud, and much has been misunderstood. Cloud is an important part of the future, but it is not a replacement for the general data centre - it is an adjunct. Understanding this will help to create a scalable, responsive and highly competitive IT platform for the future.

The next article will look at how to approach the general architecture of a virtual data centre - the use of tiering, the role of virtualisation in business continuity, and how best to drive optimum utilisation.

Getting the virtual architecture right

The approach to managing physical data centres has had to continually change, as the IT industry has moved from mainframe through midi computers, tower computers and on to rack mounted servers and blades. However, even then the old way of doing things - one or two applications-per-server - meant relatively low stress levels in data centres when it came to power requirements and heat production, even as overall equipment densities increased.

Virtualisation changes all of this, however. As virtualisation projects are used as a review point for rationalisation and consolidation of the software as well as hardware assets, many organisations find themselves not just with higher hardware densities, but also individual servers with far higher utilisation, which can really stress power and cooling capabilities. Indeed, in many early implementations of virtualisation, Quocirca saw how simple approaches of trying to stack similar equipment in specific areas led to disastrous results. For example, placing all power converters in one stack, all cpu capability in another and all storage in another led to severe problems at the power converter level: removing heat effectively from such a high dissipative source would require highly specialised cooling systems to ensure that overheating would not occur.

The majority of vendors and data centre implementation partners will have built up their own “best practice” capability in how to architect a platform for a highly virtualised environment. However, even these may not take everything into account when it comes to building an overall best architecture that looks not only at the main assets, but also at the interconnects, the incoming power distribution and the need for a dynamic capability to grow and shrink resources to meet the workloads.

Firstly, power distribution has to be looked at. In any data centre, power access has to allow for continuity, so a single source and single distribution board should be avoided. Multiple sources distributed around the data centre will provide not only greater capabilities to deal with any single point of failure, but will also help in another important area: structured cabling.

For a high density environment, cable management can be a major issue. Quocirca recommends that underfloor cabling in a virtual environment be avoided, and that power and data cables be kept away from each other. By doing this, cabling can be more easily maintained, and crosstalk between power and data cables is minimised.

Rack engineering also has to be taken in into account. A mix of power, cpu, storage and networking units in one rack is not a problem, and enables high-heat, dense units such as power converters to be mixed with less dense and

less hot units such as a router or switch. This then enables cooling systems to be better engineered: approaches such as cold/hot aisles and forced, ducted cooling will minimise costs while maximising cooling efficiency. Indeed, a well-engineered cooling system can provide outlet heat that can be used elsewhere in a building for space heating or, via a heat pump, hot water.

Although a well-architected virtual environment can lead to increased system availability, a basic system can reduce it. A move from a one-application-per-server to a many-virtual-applications-per-physical-server model can mean that the failure of a single physical server can bring down many applications or services, essentially bringing an organisation's IT services to its knees. However, using virtualisation to create "n+fractional" resilience should be considered.

In a physical set up, availability is generally provided through clustering and/or mirroring. At the simplest level, this will involve an "n+1" approach - the servers (and other assets) that are in use will require at least one extra asset to be in place to ensure a degree of availability. In fact, many organisations use an "n+m" approach, with multiple backup systems in place to provide higher levels of availability. Bearing in mind that in pre-virtualisation environments the majority of Windows-based servers are running at less than 10% utilisation, "n+1" will make it 5% (if n=1), and "n+m" will drive it down even further.

In a virtual environment, active or passive hot standby images can be used. An active image means a fully functional, fully working image ready to take over, complete with mirrored data and other resources. This is similar to an "n+1" approach - but as the image is virtual, it only uses a fraction of a (separate) physical server. For a passive image, this can be a very "thin" image: the application is available, but no other resource is attached. On the failure of the main image, resources are rapidly provisioned to the backup image, and it takes over in a short period of time – typically a few seconds.

Either will optimise utilisation and lower the costs of having a highly available single data centre. For ultimate availability an "n+fractional" multi-data centre approach may also be required - and this can be provisioned through using an on-demand IT infrastructure or a third party co-location provider's data centre facility as a primary failover site.

Virtualisation gives an opportunity to review a data centre's architecture, and to ensure that a highly effective, dynamic and responsive infrastructure is put in place. As in architecting a building, it's function that has to win over form. For a data centre, bear in mind that the function is variable: the form must be enabled to allow for this.

In a virtual space, the physical matters

Quocirca has seen many design plans drawn up by organisations around their new virtual platform. These designs show nice rows of physical racks for the servers, storage, networking systems and so on, and more often than not the design is just sat in the middle of the page as if the new platform is to hang resolutely in splendid isolation in space – the physical data centre shell is left out of the design.

Unfortunately, the physical data centre itself has to be part of the overall plans. If the new virtual platform involves the building of a new data centre, then there will be one set of rules to follow. However, for retro-fitting, a different set will be needed.

At a high level, the following are the main areas that Quocirca recommends looking at when planning a new physical data centre.

- Don't use raised floors. Equipment densities are leading to higher stresses on floor tiles and supports, and a failure of a support or tile will harm equipment. Also, the use of the under floor void as a route for cooling air is not an optimum solution: under floor wiring can easily cause obstructions that will impair cooling efficiency.

- Use targeted cooling. Hot and cold aisle cooling with ducted flows will save energy across the board. If the equipment is going to be full-volume air cooled, look at running the data centre at a higher temperature - say, 26°C/79°F. The equipment won't mind! Don't put windows into the data centre - not only are these a security hazard, but the lower thermal efficiency of glass will cause greater heating problems on sunny days.
- Look to cheaper, natural air cooling systems, such as the Kyoto Wheel system. Even offsetting a small amount of cooling costs will add up to big savings over the life of a data centre.
- Plan for the future. Make the data centre something that can grow and shrink as needs change. Look at the use of secure mobile walls and cages to ensure that only authorised people can access certain virtualised environments.
- Plan for lights out. Data centres and people don't mix: not only are people the major cause of problems in a data centre, but they are also the biggest security issue. Design the data centre with remote operations in mind, with people only going in to deal with exceptions such as failed equipment.

For an existing data centre, certain constraints will make it more difficult to carry out the items as above. However, Quocirca recommends the following:

- Plan for hot and cold aisle operation: this may be a relatively large investment, but the payback is rapid.
- Instrument for lights out. As above, make the data centre and environment for equipment only - keep humans out as much as possible!
- Either brick up or coat any existing windows with UV reflective film to reduce the heating effect of the sun.
- If virtualisation is being done as part of a rationalisation and consolidation programme, then the amount of data centre space required initially will be far lower than that already being used. Evaluate how to best contain this space, and remember that, if using raised floors or dropped ceilings, any new walls must go from sub-floor to super-ceiling to maintain cooling efficiency.
- If not already being done, move to fully structured cabling. This will ensure that any under floor cooling paths are kept clear, and will also enable better planning and maintenance. It also provides the opportunity to move data and power cables apart to minimise any crosstalk issues.

In both cases, it has to be borne in mind that a virtualised platform will work in a different manner to a physical one. Anything that requires high levels of security, in the form of named people only being able to access the hardware, will need its own virtualised environment that is kept physically separate from the rest of the systems. However, growth has to be allowed for. In a physical data centre where one-application-per-server is the norm, any large step change in requirement can lead to a "fork-lift" (complete replacement) upgrade. In a virtualised environment, it should lead to the introduction of incremental hardware capability. This needs to be planned in to ensure minimum data pathways, sufficient power distribution and cooling capabilities along with the capacity for data and power wiring to be laid to the systems. This has more impact on the physical design of the data centre than on the design of how the existing or planned racks of servers, storage and network equipment should be laid out.

Overall, the design of a data centre has to be seen as one where the function comes first and the ascetic design second. The primary aim has to be to provide an environment that will provide a workable data centre for the foreseeable future - which will be around 5 years. Through careful upfront planning, a well-designed physical data centre should be adaptable to make the most of new architectures beyond this period - but only if the data centre itself is designed with flexibility in mind.

Indeed, when people were building data centres 10 years ago, rack mounted systems were still in their infancy. The idea of blade systems and multi-petabyte storage based around massively virtualised hardware was unthinkable - and yet many modern IT platforms are based on this concept, shoe-horned in to a physical building designed to deal with low-density, low-heat output tower systems. The typical IT platform in 10 years is likely to be just as different to what we have now: the data centre building has to reflect the need for such change.

Keep control – image is everything

An area where Quocirca continually sees organisations struggle with when it comes to virtualisation is image management. Here' we're not looking at how well an organisation manages its brand in the market, but in how they look after the multiple different virtual assets under their ownership. As virtualisation spreads, organisations find that there are a few problems that weren't apparent when they were first looking at it. Although virtualisation is good at squeezing the most out of existing assets, it has an unfortunate flaw in its own simplicity.

Take an example. A developer needs to check that an application they have written works as expected. There's plenty of hardware resource available, so they create an image of the application and everything else that is needed as a stack underneath it, spin it up and carry out a few tests. Things aren't working as expected, so off they go, change a load of things, and do the same routine for a new test. This probably happens several times before a working version of the application is obtained, which is then shipped off for full pre-mainstream testing, where new images will be created, provisioned and run as required.

All seems nice and simple, so where's the problem? Certainly, the business has gained from faster development cycles, the developer is encouraged to carry out more testing cycles and everyone seems to have gained.

The problem comes down to how easy it is to set up a virtual image - and how easy it is to then forget to break it down again when finished with. Starting from a position of a virtualised hardware platform, a lot of other things are required before an application can run. An operating system will be needed, probably along with an application server platform. Any dependent software will need to be present as well. And it all needs to be licensed.

Herein lies the problem. If virtualisation makes it easy to create or copy images and spin them up as individuals deem necessary, then each one requires its own license stack. If the user does not deprovision the image correctly, the licenses remain there - unused, but current, accruing the need for renewal and maintenance payments. Even if the software is based on SaaS subscription licensing, any "live" spinning image will nominally attract the need to continue paying subscriptions unless the image is deprovisioned correctly. Therefore, it is an imperative that image management has to be in place to manage all the licenses that are in use across a fully audited asset register of the images that are in use within an organisation.

And it doesn't stop there.

Let's assume that an organisation has a collection of images that are all necessary - for example, 100 mission critical images that are all provisioned in the live environment. Each of these will need to be patched or upgraded any time that any part of the stack changes - and may also need to be fully retrotested to ensure that everything still runs as it should. If there are also a number of none-used, but live extra images in the development, test and live environments, each one of these will also need to be patched or upgraded, each with at least a resource cost to the organisation.

So, image management is a necessity within the virtual world, ensuring that only what needs to be licensed is licensed, minimising costs by using library functions for checking licenses in and out as required. Only images that are tagged as being live and in use need to be patched and upgraded - any images that are found which are not tagged as such should raise an event, notifying the owner of the image that it has been identified, and giving them a period of time before the image is deprovisioned and its licenses freed up for use elsewhere.

However, the future may be brighter. Many of the vendors have realised that image sprawl is a growing issue, and steps are being taken to make life easier. For example, in-line provisioning of composite images are coming through. Here, a description of an image is held - rather than an image. When it becomes necessary to provision the image, the description is found and the image is built on the fly. Therefore, patching and updating is minimised - the "golden" images of the operating system, app server and so on are stored just once and are only brought to life

as and when needed. Certainly, cloud computing adoption will also drive a need for a different way of dealing with composite applications, one where functional services can be aggregated together to facilitate business processes.

This requires micro-aggregation of small subscription payments against transactional usage. Images here may be spun up and down in a matter of minutes. The thought of trying to run such a system without full image management in place should be enough to make even the hardest IT manager shudder.

Combine this approach with automatic image management, based on ageing images around variables such as lack of use or defined project lengths, and the capabilities of the virtual environment to provide ongoing support to the business at the most effective cost can be found. Unused images taking up virtual and physical resource are minimised, unused licenses taking up costs and human resources are avoided. new images can be rapidly created for use in any scenario, in the knowledge that these will be gracefully aged and any licenses freed for reuse up in a defined period of time.

Without control, virtual images can bring down what could be a well-worked virtual infrastructure. So, polish up your image - gain control now and provide a much greater level of service to your organisation!

Maximise savings – the license conundrum

In a virtualised world, licensing becomes a major headache. Many vendors are still pricing their offerings based on per cpu, per core or per MIP levels – all of which can be dynamic in a new virtualised architecture. For someone responsible for operating a virtualised data centre, there are several things to bear in mind when negotiating with software vendors on new or renewing contracts.

For many incumbent software vendors, the problem is not an inherent incapability to change their licensing models, it's how to face up to external stakeholders that is often the overriding issue.

Imagine that you are a very large software vendor with a model based on the number of cores that the software is installed upon. The move to multi-core, multi-socket servers is good news to shareholders and to the financial analysts watching the company – license revenues should increase along with the numbers of cores being installed in the largest organisations. Now imagine that the software vendor says that it needs to move to a different model, one which is far harder to define, maybe based around per user licensing. Although an organisation has to have a minimum number of cpus and cores in order to meet the technical requirements of the software and its workloads, it can economise through having fewer users. Therefore, revenues for the software vendor may be hit – and hit hard. An alternative move to a subscription model means that there is less tie-in for the user, so long-term revenues are threatened. The financial stakeholders get worried, stock process fall, and doom is forecast for the vendor.

Also, a move towards "functional computing", where archetypal applications break down into component parts which are then brought back together as required to facilitate business processes, is creating further stresses that many existing software vendors are struggling to address adequately with their stakeholders, preferring to stick with more monolithic approaches to technology "solutions".

It is likely that the licensing landscape will change dramatically in the next 2 to 3 years. However, that is of no use to businesses trying to deal with the here and now; with how best to ensure that they get the best out of their investments in existing software, and how contracts are renegotiated to reflect the needs of organisations in a new virtualised world.

If you are big enough, the majority of software vendors will be flexible. The attitude tends to be "if you have enough money to spend on us, we'll be flexible enough to meet your needs on how to take it off you". Therefore, you can pretty much say that you want an enterprise license enabling you to install the software wherever you want

for as many people as you want. The majority of software vendors will also enable such licenses to be used for migration into the cloud, should the organisation so choose at some point in the future. Contracts can be front loaded, levelled or even financed as the customer requires. The vendor just wants to get that money off your balance sheet and on to their own. The high cost of negotiation and sale are dwarfed by the overall size of the deal, and a high touch approach is therefore seen as being worthwhile.

However, what to do if you are not big enough for the software vendor to see you as a large named account? High touch negotiation isn't on – the cost just can't be covered by the overall deal size for the vendor. Cheaper means of dealing with smaller customers have to be the order of the day – but this tends to constrain the options available to the customer.

Firstly, see if per-user licensing works for you, and see if the vendor is then happy for you to install the software on any hardware (physical or virtual) to meet the individual users' needs. This may not work, however – certain back office systems are difficult to define on a per user basis, with transactional volumes being far more important.

So, transactional charging may be a different approach to take. However, if you want predictable costs, pure transactional charging may not be what you really want. The best approach here is to see if the software vendor will work within "bands", where charging is based on usage between lower and upper limits. In this manner, your costs should not change from month to month, unless there has been an appreciable spike in transaction volumes. Also, ensure that even if you do move in to the next band of transaction volumes, you are then able to move back down within a reasonable period of time should transaction volumes decrease again.

A move to open source could also be considered. Quocirca does not believe that this should be done for reasons of cost only, but provided that the functionality that you require can be obtained in a fully supportable manner, the lack of a direct license cost makes open source a very neat fit with virtualisation. However, it remains necessary to understand any license implications arising from the use of GPL or other open source licensing models, such as any reuse of code in a commercial sense.

Maybe you just want to move away from traditional licensing models completely. As accounting rules change and there is a growing need to be able to assign an asset value to software held by the organisation, it becomes increasingly attractive to move to a subscription based model, where the software is never owned – and therefore requires far less complex management. Note that last comment, however – "far less complex management". Subscriptions still need managing, and knowledge of the underlying software license models may well still be required.

As part of this, cloud computing can also be an option. Combined with your own virtualised data centre, cloud can enable functions to be called on a per use, a subscription or a transactional volume basis to plug holes in the functionality within your own environment.

The new virtual model is severely stressing existing license models, and Quocirca is seeing new models come through from new companies, as well as seeing existing software companies begin to address major issues in older style licensing agreements. As cloud computing becomes more of a reality, licensing will have to become less complex and more flexible. Until then, being forewarned is definitely being forearmed. With knowledge of the basic approaches above, you will be better placed to choose a suitable software vendor or channel, and then to carry out negotiation wherever possible to ensure that your organisation gains both flexibility and cost optimisation against a software asset base.

Optimising energy usage – air con is for wimps

A commonly used measure of the effectiveness of a data centre's energy utilisation is PUE – the Power Usage Effectiveness score. The means of calculating this is to take the total energy used by a data centre and divide this by the amount of energy that is used by the IT equipment (servers, network equipment, storage and so on). The overall power usage has to include the rest of the data centre's power usage – areas such as lighting and, more substantively, cooling.

Overall, a "reasonable" data centre will use around 2W of energy overall for every 1W used by the IT equipment, leading to a PUE score of 2. Many data centres are far worse than this, with some being well over a level of 5 – essentially, 4W of energy being used (or essentially wasted) in running 1W of IT equipment.

As soon as virtualisation is brought in to the mix, things can get far worse. Taking an existing data centre, and virtualising from, say, 1 000 to 300 servers (and then rationalising network and storage equipment appropriately) will undoubtedly save a lot of energy at the IT equipment level. However, unless some heavy reengineering of the data centre itself is carried out, it is unlikely that much in the way of change will be made to the existing cooling arrangements.

A data centre for 1 000 rack mount servers will probably be around 2 500 square feet, which with a roof height of 10 feet (to include raised floor and dropped ceilings) gives 25 000 cubic feet of air volume. Existing approaches of volume cooling of the air means that even when space requirements have dropped, the systems will still be cooling down the full 25 000 cubic feet of air. However, a highly dense, virtualised environment will have different heat characteristics to a standard rack mount or tower based data centre, and hotspots will tend to run at higher temperatures than before.

For many, it can be surprising to find that less equipment just runs hotter even though an amount of cooling is beings used that used to manage 3 or 4 times the amount of equipment quite happily.

There are a few things that should be done here. Firstly, it is necessary to reset perceptions. In the olden days (up to around 2 years ago), running IT equipment at high temperatures led to premature failure of items such as CPUs and disk drives. On a one application per server with dedicated disk, this was an obvious problem, and so the aim was to run the equipment well within the thermal envelope defined by the vendor. For example, when a disk drive's rated life or mean time between failure (MTBF) was around 7 000 hours (around 1 year), everything had to be done to make sure the drive would last as long as was possible.

Now, it's time to revisit this and look at what the vendors themselves are saying. The MTBF for a standard consumer disk drive is now greater than 600 000 hours or close to 100 years. Newer enterprise drives have MTBFs quoted in millions of hours. Research by Google also found that, surprisingly, disk drives had more failures when run at the colder end of the scale (80°F) rather than at a warmer level (97-120°F).

At the CPU level, it has been known that running at higher than recommended levels will lead to disastrous failure. However, by using computational fluid dynamics (CFD), the areas where cooling needs to be applied most can be identified, and targeted, ducted cooling can be applied to remove heat from the specific hot spot.

This then leaves the rest of the data centre. Time and time again, Quocirca enters a data centre and thinks "Wow! Cool!". Whereas the data centre manager tends to take this as a positive comment on how their data centre has been set up, what we really mean is that the ambient temperature in the data centre has been set too low.

Average data centre ambient temperatures tend to be around 65°F. This enables mass volume cooling across all the equipment in the data centre, but also sets a nice temperature for any humans. However, the move should be towards a lights out environment – humans should only be in the data centre when there is a problem to deal with. New guidelines from the American Society of Heating, Refrigeration and Air-conditioning Engineers (ASHRAE) say that the ambient temperature in a data centre can be around 80°F, and vendors rate much of their equipment to 90°F or above.

At 80°F mass volumetric cooling becomes far cheaper, as inlet air temperatures into the data centre can be far higher than when trying to main mass volumes at 65°F.

Further gains can be made if contained cooling is used. For this, the use of cold aisles or in-rack cooling systems means that the volumes of air being cooled are minimised to only that required to bring the temperatures of the various hotspots back into limits. Outlet air can be vented directly to the outside, vented into the data centre outside of the controlled zones, or preferably reused for space or water heating elsewhere in the building as required.

In many cases, such an approach will allow some of the existing cooling units to be switched off, saving energy and improving PUE. Where investment allows, older units can be replaced with variable flow and variable load units, which are far more energy efficient and can adapt rapidly to changes in needs in cooling volumes.

The one key, though, is to stop looking at the data centre in human terms: any human that enters the data centre should only be in there for a short period of time, and so should be able to deal with a raised temperature. Anything that has to be in the data centre on a continuous basis is not a human – and a temperature of 80°F (or higher) will not adversely affect it.

Structuring the cabling – avoiding the warm rat's nest

A virtual data centre tends to work towards very high density hardware – which in turn tends to lead to the need for highly engineered and effective cooling. With many data centres using more energy for cooling than they are using for powering the IT equipment, ensuring that cooling works at its maximum efficiency is of high concern.

However, it often seems that there is a conspiracy to counter this. The majority of data centres still work on a raised floor basis, with the space between the solid and raised floors used for carrying the cooling air to the IT equipment. In this "void" also run the various cables that are required for transferring data between the various pieces of IT equipment, as well as the electricity required to power them. Whereas this may all have been well thought out to start with, as time goes on it is likely that the cabling in this out-of-site environment becomes less structured – and introduces a few problems.

The first problem is that the sheer volume of cables can impeded cooling air flows – stressing the hardware and leading to premature failures. Secondly, the mixing of data and power cables can lead to degradation of the data signal through interference ("crosstalk"), leading to more errors and the need for the whole data centre to retransmit data packets on a regular basis, impacting the overall bandwidth available.

Another side effect that Quocirca has seen in some circumstances is where standard length power cables are used. For example, if the actual length of cable required is 7 feet, a 12 foot cable may be used, just because it is available – and the spare is curled up and clipped with a cable tie to keep it neat under the raised floor. Unfortunately, any coiled power cable becomes an effective heater – and if it is under the raised floor, it is there continuously heating the very air that the organisation is paying so much to cool to a level to keep the IT equipment within thermal limits.

Therefore, more cool air or a colder inlet temperature is required – and a downward spiral of cost effectiveness begins.

How should an organisation go about providing an efficient and effective cabling approach?

Firstly, Quocirca firmly advises against using raised floors. Increasing equipment densities are leading to higher floor loadings, and anything that then uses feet at the bottom of the racks will be applying massive effective loads (often in the tons per square inch) to the raised floor. The horizontal movement of a rack through less than an inch can lead to the failure of a floor tile, and to a full rack of equipment tipping over, possibly pulling other equipment over with it. Far better to use the solid floor as the real floor – but what can then be done with all the cabling?

A full structured cabling approach where the cables are carried in trays above the IT equipment gives much better accessibility to the cabling. It also tends to force better standards, as all the cabling is in permanent view. Power cables should be kept separate to data cables using different trunking and trays to prevent cross talk. Data and power cabling can be colour coded to designate, for example, high speed data cables, SAN fibre cables, WAN cabling, Ac and DC power and so on, so that everyone understands the purpose of each cable. Also, each cable end – whether it is for power or data – must be fully labelled so that there is no need for tracing cable runs between equipment and plug boards or power distribution boards. This should also be done at intervals along the cable, so that tracing a cable is easier.

Any changes or additions to cables must be carried out under a full change management system, ensuring that if an engineer moves a cable from one point to another, it is fully logged and all notations on the cable are consistently maintained. For example, taking an existing cable and moving one end to a new piece of equipment and labelling that is not enough – the two ends need relabeling and the new connections need to be fully logged within the data centre plans and in the system management software.

To give an industry best practice guide to data centre cabling the Telecommunications Industry Association (TIA) has issued a structured cabling standard (TIA-942) which provides data centre designers, builders and managers a set of guidelines that will ensure consistency of effective cabling across a data centre. The standard takes a zoned cabling model, with the data centre being broken down into and entrance room (ER), a main distribution area (MDA), horizontal distribution area (HDA), zone distribution areas (ZDAs) and equipment distribution areas (EDAs). The MDA is the main distribution bus for data cabling, and should be a one-time installation. The HDA provides the capability for horizontal cabling interconnects to other areas. Each EDA is where the main computing equipment is deployed, and feeds off the MDA via the HDA. ZDAs can be deployed if required to allow for frequent reconfiguration of cabling within a data centre – essentially consisting entirely of patch panels and cabling racks.

By following a well designed and implemented structured cabling architecture, not only will the data centre be better maintained and its appearance significantly enhanced, but data throughput will be optimised and cooling architectures can be better managed through the use of a structured cooling architecture, using hot and cold aisles along with highly targeted spot cooling, so massively reducing energy needs across the whole data centre.

The age of just connecting equipment A to distribution board B via an under floor void has to be brought to an end. Structured cabling has to be seen as a necessity in today's data centre – not a "nice to have". Even for older data centres, looking at moving to a structured cabling system can bring great benefits in traceability and managing adds changes and deletions of equipment in the data centre, as well as energy savings and better cooling.

Turn the lights out – people and data centres don't mix

As the history of the data centre has unfolded, we have seen the dynamics of how these work change. In the first place, the mainframe-centric data centre was a hallowed place, only to be entered by the neophytes of the new technology world. Air quality had to be maintained, temperatures were to be very closely monitored and vibrations minimised to protect the fragile components of this very expensive piece of kit. "Users" – including coders, testers and sysadmins, as well as those actually needing the capabilities of the mainframe, accessed the computer from afar through the 3270 "green screen" terminal.

Then came the era of the midi computer, with the growth of companies such as DEC and Bull. These new computer devices needed a higher degree of management, and parts of the data centre suddenly found themselves with more people coming in to apply operating system patches, to oversee reboots and to add additional pieces of hardware to the systems. The final part of this evolution was the introduction of the PC-based server – the explosion from companies such as Compaq, HP, IBM and the move to standard high volume (SHV) Intel-architected servers costing a few thousand dollars each. The data centre moved from hosting a few items of large kit to the need to provide an environment where hundreds and the thousands of smaller devices had to be housed. The explosion of compute devices also meant that extra kit had to be provided to deal with the power and data cabling that was required, and this also led to the need for distributed cooling, rather than spot cooling.

The data centre became more like Grand Central Station with facilities management, data centre, telecoms, network, development, testing and other staff wandering in and out as they saw fit. Trying to control this has become a major issue, as each person is certain that they have to be right there next to the hardware to carry out their particular task. The impact on corporate security is high – but there are also other impacts that need to be kept in mind.

The data centre for people needs to be fully lit and needs to be kept at a temperature that is conducive to work. The data centre for hardware can be in complete darkness and run at a temperature that is OK for hardware. These two basic points already point to the fact that people and data centres don't mix. Build on this the fact that the vast majority of problems in the compute world are caused by human error – and the more you can keep people away from the hardware, the better.

Can it be done? Quocirca believes that the vast majority of data centres could be moved to "lights out" operation for 95%+ of the time – and the other 5% or less of the time should have minimal impact on how the data centre is run.

Systems management should be carried out as it was back in the mainframe era – from afar. Today's systems management packages can manage both physical and logical platforms from the one pane of glass, and will either be agentless or will be able to push agents onto the required assets in a remote manner. Applications should be able to be provisioned and deprovisioned without the need to go into the data centre. The use of advanced sensors and monitoring systems should be able to pick up on issues as they start – and for automated systems to take actions to prevent these issues from becoming problems. For example, in a virtualised environment, the quick recognition of the appearance of a hot spot in the virtualised estate can lead to the workloads dependent on the physical asset being moved over to other hardware through virtual to virtual porting, followed by the affected physical resource being turned off, so negating the hot spot issue.

This will mean, however, that there is now a non-working asset in the data centre – and that a person will now need to go into the data centre to deal with the problem. However, this should be the exception rather than the rule, and the engineer should be in there just long enough to replace the affected system. These days, it's not worth trying to fix the item in situ – just pull it out, replace it and deal with the affected item outside of the data centre. Low

intensity, cold lights only need to be on while the engineer is in the data centre – indeed, the area where the affected item is located is the only area that needs lighting so that the engineer can carry out the task.

The same approach should be there for those responsible for patching and updates – the data centre is out of bounds unless there is a solid need to go in there – and the person going in will have to deal with the 30°C/86°F+ temperature that is in there. Networking in a virtual data centre should be virtualised – new links can be set up in the same manner as new applications or functions can be provisioned – safely, from outside using automated processes. Facilities people should ensure that a zoned system is used, so that they don't have to go into the main data centre areas for day to day work, but that cooling equipment (where used) and power distribution equipment are all housed in accessible areas outside of the main areas.

A well-managed, lights out virtualised data centre not only leads to a better managed data centre, but also saves money through not requiring lighting, being able to run the data centre at its optimum temperature, in how this drives the requirement for fully automated processes, and in the extra security that is enabled through minimising the number of people who are allowed anywhere near the hardware assets.

Go for function, not application

The virtualised data centre gives an opportunity to review where we are with IT as a platform that supports the business. Sure, virtualisation can be used to drive up resource utilisation while lowering energy requirements – but if that is all it is used for, are organisations missing out on some hidden opportunities?

The enterprise application has been the mainstay of computing for a long time – organisations firstly bought highly specific applications that were then packaged together as e.g. customer relationship management (CRM), enterprise resource planning (ERP) and supply chain management (SCM) "solutions". This has led to a marked redundancy of function throughout the data centre – siloes of not only data but of code that is there to find details on customers, suppliers, inventory or whatever. The knock on effect from this is that information may differ from system to system, leading to errors in how data is interrogated and making an organisation's decision making capabilities suspect. It also has the obvious impact on technology resources – the redundancy of function means that the effective utilisation of the resources is even lower than is measured through standard means.

Attempting to remove such redundancy through approaches as enterprise application integration (EAI) has proven complex and constraining – as well as expensive. Even then, it has proven difficult to isolate and turn off the redundant functionality, so efficiency gains in resource utilisation have been low. Further, the software vendors aren't interested in whether you are using a function or not – you still have to pay for it, so there is no option for saving at that level.

The emergence of web services and service oriented architectures (SOAs) has provided a basis for this to all change – and virtualisation along with cloud provides the last pieces of a jigsaw that should change the way that technology as a corporate resource is viewed.

If a standard business process is taken as an example, there will be a customer that needs to find an item, order it, pay for it and get it delivered. In today's world, this crosses the web ecommerce site, the CRM, ERP and SCM systems, and may well involve a few more applications as well. For the process to run smoothly, each system has to have its own data stores, and they all have to be linked through some means so that each one can share its knowledge as required.

Turn the problem around: each customer, supplier and product is unique. If these were created as discrete records, anything that needed details on them could call the records associated with that item. It now becomes a case of is the customer already known? If yes, get all payment, delivery and other details from the master record for that

customer. Is the product in stock? Look at the master record for the product – if it is in stock, great; if not, then get the supplier master record and see if and when they can deliver one. Each stage of the process becomes a requesting or a responding service. If we change the process, the information flows are changed – there is no "application" that has to change.

Further, take a more immediate example. A retail operation wants to open a new retail outlet. It can interrogate the details of its existing customers, looking for all those who have bought off the web and where they live. It can call a Google or Bing maps function – served externally – and map the data directly on this and identify the hot spots where the largest density of customers are. No need for a mapping application, no need to run it inside the corporate data centre. Want to see if it would be better to open up a concession inside someone else's store instead? Use the same data, compare with existing shopping habits from the target store's systems and see where the hotspots are now. This is not using an application, it's making the most of the data that has been growing across organisational value chains.

The virtual data centre makes the most out of this through being dynamic, through being able to apply the resource to the workloads that need it. From an application point of view, any resource calculation is aimed at ensuring that the application can meet at least average workloads plus a bit, while in many cases they are engineered to meet expected peak loads.

A functional approach means that resources can be allocated to deal with data as the main focus. Functions tend to be small pieces of code that require small amounts of resource. If a lot of a function is required, it is far easier to allocate more resources on the fly. The function can be delineated, the basic workload of the function is known, its performance limitations (such as whether it needs a lot of cpu, i/o or storage) should be more easily understood and as such be able to be dealt with in a flexible manner.

In essence, the days of the monolithic application are numbered. Sure, we're in a "long tail" situation: the applications will be around long after I'm dead and buried. However, the clever organisation will start to plug holes in existing application functionality through calling discrete functions provided either internally or externally. As more functions become available, the "composite" application takes over: the application that is built on the fly to meet the needs off the individual, group or organisation as it battles to deal with ever-changing market forces.

The virtual data centre is a necessity to enable the functional composite application: the composite application is a necessity to be ultra-competitive in tomorrow's markets. As IT, it is your responsibility to advise your organisation accordingly. Prepare for function!

Prepare for hybridisation – cloud is coming

Throughout this series of articles, we have looked at how best to move towards a virtualised data centre. However, there is one area where virtualisation promises to have far greater impact on the organisation, and that is through cloud computing. Cloud computing is, at the moment, suffering from mixed vendor messaging and a massive amount of market hype. However, Quocirca firmly believes that cloud is a game changer, and that organisations need to be able to prepare for its coming.

So, just what is cloud computing? In its simplest form, it is the capability to create a massively shared environment where IT functions can be sourced, provisioned, run and managed in a highly flexible manner. There will be some functions that will remain running in the internal private cloud (based on the virtualised data centre that has been previously discussed), some functions will be provided through an external private cloud (essentially, co-location or functions hosted privately through an external provider), while other functions (such as mapping or use of publically available data sets) will be provided through public clouds through companies such as Google, Amazon, Microsoft and others.

The issue here is to choose what fits where carefully. Many vendors are talking about storage as a service, where an organisation's storage needs are met by cloud based services. However, many organisations are spending time and money in optimising their storage internally to provide the best response times for on-line transaction processing (OLTP), for data searches and business intelligence. The introduction of solid-state disk (SSD) and in-memory capabilities along with optimised I/O and high speed core networks means that any latency in the network is minimised. Now, imagine taking all that and moving it to the cloud – the 10GB core network is now down to a 10MB wide area network – this is not the right way to look at core storage needs. However, for granular backup and file-level restore, cloud-based storage can make a lot of sense.

How about if the cloud is used purely as an access mechanism, with all the business logic and storage being carried out in the public cloud, and only the presentation layer being sent back to the access device? Certainly, advances in thin-client or server-based computing means that this is far more viable than it has been in the past. With the external cloud provider hopefully being able to invest in the latest technology, the basic compute equipment should be faster and more up to the job than an environment where organic growth has resulted in a heterogeneous environment. Similarly, networks, storage and other systems should be optimised for the task in hand, and systems management should be capable of ensuring high availability and proactive fixing of possible issues. A well architected shared external cloud platform will enable greater flexibility and possible economies of scale than an internal cloud.

But how about the public cloud? Quocirca's impressions of the market are that the majority are still wary of using cloud in this manner, with security being quoted as the top concern. Quocirca's advice is to create a list of the top ten or twenty security areas that you would expect an external to be able to demonstrate they can deal with – and then see how your internal environment can deal with these. Quocirca has found that many internal systems show massive security shortfalls when submitted to this sort of scrutiny – and that this should then be used as the benchmark for any external offerings.

It should also be remembered that it is likely that your organisation is already using the public cloud anyway – Google or Bing Maps are public cloud offerings, and it can well be argued that Facebook, Twitter and other social networking sites are increasingly moving from being a hosted application to being cloud services through the increased use of provided application programmer interfaces (APIs). Organisations are either already using the public web – or are fighting a Cnut-style rear guard action in trying to stop the tide of usage – and so missing out on the opportunities it can provide.

The key is to apply the resources where they make most business sense. As the archetypal data centre changes into a highly virtualised environment, the cloud has to become a core part of any plans. Commodity processes and functions, such as email and calendaring, as well as other communication and collaboration technologies are best outsourced due to the way that an external can provide economies of scale in hardware support and in licencing. Similarly, specific functions, such as expense claims management and payroll are best outsourced as the domain expertise is there to deal with the changing legal aspects of these tasks. Although these tasks may be viewed as business important, the underlying technologies that support the functions are now commodity – and your organisation shouldn't be bearing the support costs for the complete stack internally.

This then enables an organisation to concentrate its resources on what is really important to it – those areas where an external provider does not make sense, where corporate intellectual property has to be created and dealt with in a fully managed manner. By outsourcing commodity, financial resources and skills are freed up to work at optimising the existing data centre to better manage these workloads – and with a degree of existing equipment now being made redundant, space is cleared in the existing data centre for new virtualised kit that is optimised to deal with these workloads.

The existing data centre is not dead, nor is it even at risk of beginning to die out. It has its part to play – as do external cloud services. By allocating workloads appropriately, a highly effective platform can be built that fully supports the dynamic needs of the business. The key is not to concentrate on whether cloud is right, or even which type of cloud is right for your business. The focus has to be on what best fits where – and how to move to this nirvana as soon as possible.

The dynamics of the data centre

In the last 11 articles, Quocirca has looked at what is required to move to a virtualised responsive data centre. This is increasingly necessary, as the market forces on the business means that it has to be more responsive to the competition, and begins to realise that the overriding need is to be able to facilitate the business through flexible business processes – not through implementing layer after layer of constraining applications and other technologies.

To get to a position of the best optimised data centre that does support the business, Quocirca therefore advises that the following steps are taken:

Application rationalisation
Bringing the number of instances of applications under control is a great starting point for moving to a virtualised data centre. If you have different applications running doing the same thing, choose a strategic vendor, and rationalise to their product wherever possible. If you have 5 different instances of SAP running, get it down to one or two.

Virtualisation
The aim here is to take as much of the physical estate as possible and make it appear as a single pool of resources that can then be sliced up into logical units as required. This does not just apply to servers, but also to storage and the network. However, certain workloads may still need to be placed on physical servers – either for distinct workload reasons, or because the business is not yet ready to move to a completely virtual approach.

Application consolidation
Once you have the minimum number of applications and application instances running, it is time to consolidate them onto the smallest footprint possible. This can be carried out as a pre-virtualisation task, with the aim of minimising the number of physical servers being used to support each application. In itself, this does have value, but Quocirca recommends that consolidation be carried out as a post-virtualisation task, so that resources can be applied to workloads on a far more granular basis.

Application granularisation
The lack of functional granularity within the majority of enterprise applications is a brake on many organisations. Through the use of web services and service oriented architectures, the various functions within an existing application can be surfaced in a manner that allows each function to be called and used outside of the application itself. This then enables business processes to be supported on the fly, as functions are brought together to meet the needs of the process as it changes. New functions can be introduced either as new internal functions, or can be sourced and implemented as external cloud services.

Data centre optimisation
Once a suitable IT platform has been reached, this will have impact on the data centre itself. Power distribution, cooling, space and other aspects will all need to be reviewed to make sure that everything can still meet the needs of a far more densely populated environment. More targeted cooling will be required to minimise energy usage, along with floor-to-ceiling baffling to create suitable baffling for what should now be a far smaller overall data centre.

However, the act of virtualisation and optimisation also provides the opportunity for the business to move towards far more ecological and long-term data centre efficiency approaches, such as free air or water cooling, energy reclamation and reuse and so on. The move can also be taken as an opportunity to review how functions are being provided to the business, and whether the external cloud is a suitable means of providing new, or even existing, functions in a more appropriate manner to the business. The move to a new data centre architecture is the best

time for such decisions to be made, and to create a long-term, flexible architecture that will support the business for the future.

The aim of a highly virtualised data centre is to move away from a constraining, under-utilised platform to a far more dynamic, responsive and cost effective approach. New management tooling may well be required that can manage the mix of physical and virtual systems, but the move to higher resource utilisations and densities, combined with lower energy costs and space requirements should help to move the IT budget from one where around 70-80% of a typical organisation's budget is spent in "keeping the lights on" to one where less that 50% is spent in this way. By freeing up 20-30% of the IT budget, the IT and facilities management departments can respond to the current "do more with less" calls from the business in a manner that still enables increased investment in what IT should be there for – in supporting the business in carrying out its core operations.

Sure, there is a distinct cost in making the move, but the business, along with IT and FM need to be aware of the cost of not carrying out such changes. The data centre increasingly becomes a brake on the business and business decisions cannot be made that reflect the needs of a dynamic market. Newer technologies and architectural approaches cannot be adopted in an effective manner, and the competition that does move in this direction becomes far more effective, attracting custom from your base and so lowering the overall effectiveness of the business in the market.

Neither will maintaining a mindset around applications sitting on physical servers help – as more applications are required, or as workloads increase, the end result will be the need to build a new data centre due to space, cooling or energy requirements. If the aim is to save money, suddenly having to find the money to build a new state-of-the-art facility will make retro-fitting an existing facility to meet the needs of virtualisation look like petty cash.

Make no mistake, virtualisation is happening – in many cases has already happened. As a cornerstone in ensuring that IT and FM departments support a business in the manner it demands, virtualisation has to be embraced, but with eyes fully open and with IT and FM working together in a strategic manner.

REPORT NOTE:
This report has been written independently by Quocirca Ltd to provide an overview of the issues facing organisations seeking to maximise the effectiveness of today's dynamic workforce.

The report draws on Quocirca's extensive knowledge of the technology and business arenas, and provides advice on the approach that organisations should take to create a more effective and efficient environment for future growth.

About Quocirca

Quocirca is a primary research and analysis company specialising in the business impact of information technology and communications (ITC). With world-wide, native language reach, Quocirca provides in-depth insights into the views of buyers and influencers in large, mid-sized and small organisations. Its analyst team is made up of real-world practitioners with first-hand experience of ITC delivery who continuously research and track the industry and its real usage in the markets.

Through researching perceptions, Quocirca uncovers the real hurdles to technology adoption – the personal and political aspects of an organisation's environment and the pressures of the need for demonstrable business value in any implementation. This capability to uncover and report back on the end-user perceptions in the market enables Quocirca to provide advice on the realities of technology adoption, not the promises.

Quocirca research is always pragmatic, business orientated and conducted in the context of the bigger picture. ITC has the ability to transform businesses and the processes that drive them, but often fails to do so. Quocirca's mission is to help organisations improve their success rate in process enablement through better levels of understanding and the adoption of the correct technologies at the correct time.

Quocirca has a pro-active primary research programme, regularly surveying users, purchasers and resellers of ITC products and services on emerging, evolving and maturing technologies. Over time, Quocirca has built a picture of long term investment trends, providing invaluable information for the whole of the ITC community.

Quocirca works with global and local providers of ITC products and services to help them deliver on the promise that ITC holds for business. Quocirca's clients include Oracle, Microsoft, IBM, O2, T-Mobile, HP, Xerox, EMC, Symantec and Cisco, along with other large and medium-sized vendors, service providers and more specialist firms.

Details of Quocirca's work and the services it offers can be found at http://www.quocirca.com

quocirca