

March 2008

Capacity Utilisation – Asking the Questions

By Adrian Johnson, HyPerformix & UKCMG Committee Member

“We won’t really know how the system performs until it is out there in production” – once upon a time, this was considered an acceptable reason to not take performance and capacity seriously prior to the release of a new application. Servers were often regarded as commodity items, at least in comparison to mainframe technologies, and it was simply easier if each new project brought in its own set of machines. The team that had to build the system were rarely the same people who had to manage it in production, and the explosive growth in the number of machines often went unchecked for far too long.

Analyst reports quote average capacity utilisation statistics of the server farms across the industry as being between 15-20% during working hours. No other corporate asset, such as real estate or manufacturing plant, would be allowed to continue at that level of utilisation – serious questions would be asked at board level. In fact, with space, cooling and power in the data centre becoming real issues, and the notion of a corporate carbon footprint starting to tug at the consciences of senior management, these questions *are* starting to be asked.

When we question why the IT infrastructure has been allowed to reach this state of under-utilisation, we hear about the business risk attached to allowing customers direct access via the web (exposure); we hear about beating competitors to the market opportunity window (time pressure); and we hear about the rate of change of new technology (risk to existing operations). But as the IT industry matures, surely some of these ‘reasons’ begin to look more and more like excuses?

Many of the current trends, such as Web Services / SOA and virtualisation, are being driven by a variety of perceived needs including maintainability and manageability. Whatever the driver, key to the success of all these trends will be effective use of common components and shared resources; improvements in the capacity utilisation of the server estate should start to be achieved as a result. However, there is a potential cost – whilst capacity utilisation may improve, the impact on performance and response times also needs to be considered. Web service re-use may save development time, but if only a small portion of the current functionality is actually required in the new application, how much unnecessary computing effort will be expended? Almost by definition, virtualisation is replacing real stuff with virtual stuff – if emulation is more time-consuming than direct operation, what does that do to performance?

Does this following scenario seem familiar? *“The operations team is handed a new application to manage. It comes with a whole new set of servers and SLAs; it is held together by some new middleware technology buzzword; development suffered slippages but the release date was held fixed by the business requirements so QA and test was compressed; and now the team that built the application has dispersed on the four winds to the next project/contract, so access to anyone who knows how it works is limited.”*

We’ve probably all been there, but there are signs of hope. IT Service Management has become a hot topic globally in recent years, so comprehensive and cohesive vendor and tool support is starting to arrive at last. The same Systems Management tools used in production are starting to appear in test environments; corporate standards and reference architectures are being brought in, although not yet always adhered to.

If a career in the IT industry is to be considered seriously as a 'profession' not just a job, engineering discipline that is beginning to emerge within IT development and management practices must continue, and become standard practice not just 'best practice'. Applying engineering discipline to IT development means designing and building-in appropriate functional and non-functional requirements from project conception and seeing them through the system lifecycle. Time-to-market is commonly cited as a reason for project time-scales that render such laudable goals as unrealistic, but this misses the point. A well-engineered system should be extensible such that additional functionality can be added with minimal change to existing services. Not only can time-to-market still be met, but the future maintainability and manageability of the system is also improved. Distributed responsibility for distributed systems can work, but only if 'end-to-end' is seen as being applied to the system lifecycle as well as the infrastructure itself.

To summarise, in order to derive best value from new system implementations or major enhancements, it is imperative to manage the capacity and performance throughout the project lifecycle from inception to implementation and on through the lifetime of the system. Establishing a set of robust processes in support of this objective will provide the full end-to-end Capacity Management solution that is needed.

UKCMG are an independent User Forum where End Users, Industry Experts and Vendors meet to share ideas and experience regarding Performance Engineering, Performance Testing, Capacity Modelling, Capacity Planning, Performance Monitoring, Performance Tuning and Capacity and Performance Reporting, and by sharing this knowledge provide its membership with a Centre of Excellence for Capacity Management.

If you would like to learn more through our regular events, please visit: www.ukcmg.org.uk