Evolution of Hosted Server Computing

The emerging 'cloud' alternative

Dale Vile
Freeform Dynamics Ltd

June 2010

The research upon which this report is based was independently designed, analysed and reported by Freeform Dynamics in association with *The Register* news and information site. The work was sponsored by Microsoft.

Introduction

There has been much talk and activity in the industry recently around 'Infrastructure as a Service' (IaaS), part of which is about meeting your server computing needs through the use of hosted infrastructure rather than by installing and running equipment in your own data centre or computer room.

While there is obviously nothing new about the concept of hosted services *per se*, which have been around in the IT industry for the past 40 years, developments in the area of virtualisation, along with the emergence of more flexible service models, are raising the profile of commodity or utility style infrastructure services under the 'cloud computing' umbrella.

In this short paper, we look at the current state of play with regard to hosted server offerings and activity, and consider where the market is going as we look forward.

Inputs into the discussion

During the course of our discussion, we will draw upon the results of an online survey conducted by Freeform Dynamics via *The Register* (www.theregister.com), a popular IT news and information site which provides access to a large pool of IT professionals. In terms of sample composition, the respondent base reflects the natural distribution of IT professional roles in the mainstream, at both a management and practitioner level, with representation from the UK (33%), other countries in Europe (18%), the USA (28%), and various other geographies around the world (21%). The research was commissioned by Microsoft, and other information regarding methodology is mentioned in context at the appropriate point in the discussion.

Starting at the beginning

Before looking at hosted infrastructure services and their use, it is worth taking time out to recap on the range of activities, or 'workloads', that server estates support in an on-premise (on site) context. This provides us with a baseline for comparison.

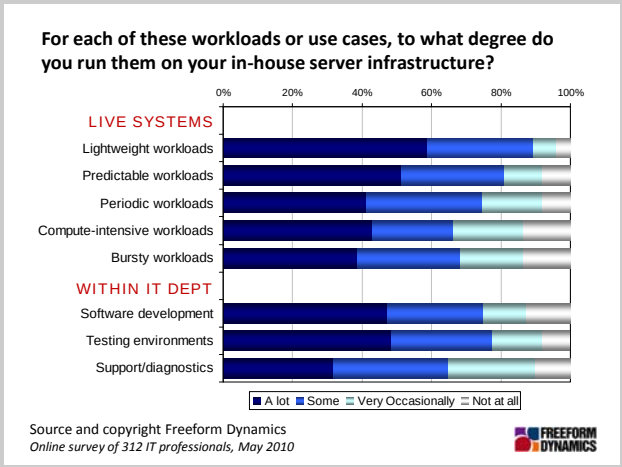
In previous studies[1], we have categorised workloads in a functional manner around things like application serving, web serving, database serving and so on, and for many purposes this is a great way of looking at activity.

When investigating hosted services, however, one of the most important attributes of a workload to consider is its resource consumption profile, as this can have a bearing on not just which workloads can successfully be run in a hosted environment, but also which type of hosted service is more appropriate for each.

For the purposes of our investigation, we therefore classified workloads in the following manner:

WORKLOAD CLASSIFICATION BASED ON RESOURCE CONSUMPTION PROFILE	
Lightweight workloads	Production applications with occasional and/or lightweight resource needs, e.g. tactical workgroup or departmental solutions
Predictable workloads	Production applications with significant ongoing and predictable resource requirements, e.g. accounting, ERP, and other core systems
Periodic workloads	Production applications that kick in periodically with significant resource requirements but otherwise just tick over, e.g. payroll and billing
Compute-intensive workloads	Production applications with extremely high, sometimes periodic, resource requirements, e.g. high transaction systems, HPC, etc
Bursty workloads	Production applications with highly fluctuating peaks and troughs of demand, e.g. online sales and service, some call centre scenarios, etc

Added to these, we then included the use of server resources within the IT department itself to deal with requirements in the areas of software development, testing and support. When we look across all of this activity, the following picture emerges in terms of overall workload mix for the 'average' on-premise server estate, which is pretty much as we would expect (Figure 1).



Server workloads vary in terms of their resource demand profiles.

Figure 1

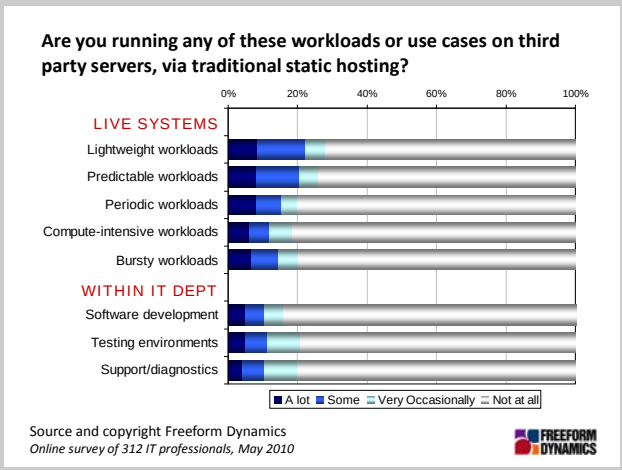
Of course in reality, there is no such thing as an average organisation, but this kind of view does confirm that workloads vary considerably in their resource demand profiles.

With this baseline in mind, let's now take a look at hosting related activity, beginning with the most common type of service in use within the mainstream today.

Traditional fixed capacity hosting

Turning to the role of hosting, around a third of respondents indicate regular/ongoing use of traditional fixed capacity services if we add up the activity across the various workload categories. The services we are talking about here are those in which either physical or virtual servers with fixed capacity are rented on a fee per month basis, often under a minimum term contract.

Breaking this activity out across the same range of workload types we were looking at before, we see the following spread of activity (Figure 2).



Around a third make regular ongoing use of static traditional hosting if we add up the activity across all workload categories, but with more emphasis on predictable and lightweight applications.

Figure 2

As we can see, traditional fixed capacity hosting has pretty broad applicability, but with an increased emphasis on its use for lightweight/tactical requirements and more predictable core workloads. One of the common behaviours behind this chart is the renting of an appropriate amount of fixed capacity to run a specific application that will utilise the resources available on a steady and continuous basis. The other common scenario is the renting of a fixed amount of hosted server space that can be used to accommodate a number of smaller lightweight applications.

It is probably worth mentioning at this point that the results presented here are derived from a study that was positioned as an investigation of generic server computing needs (with no mention of hosting up front), so we will have largely avoided the skew we often see in online surveys towards those with more knowledge of or interest in the topic (often referred to as the 'self-selection' effect). What we are seeing is probably therefore pretty close to the actual penetration of fixed capacity contract based infrastructure hosting at this moment in time (though we are not capturing more bespoke outsourcing arrangements that were outside the scope of our survey, but are obviously also quite common).

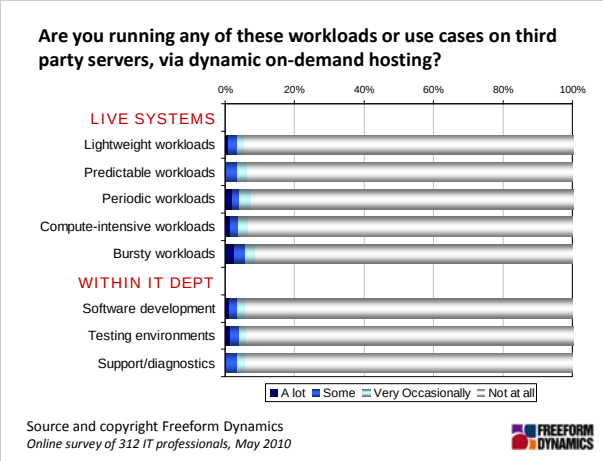
On-demand or 'elastic' hosting

While there is nothing to stop fixed capacity hosted services being used to deal with workloads that have variable resource requirements, and indeed some have gone down this route as we can see from the above, the problem is that you end up with the same challenges around utilisation that exist with on premise equipment. In the case of the latter, it has been common practice over the years to give each application its own server, and 'over provision' capacity (cpu, memory, etc) to deal with peak loads, living with low average utilisation rates, and even the machine siting there idling for much of the time. The alternative has been to lower the server spec/capacity and accept occasional performance issues as workloads hit a resource ceiling at peak times.

With fixed capacity hosting based on a committed fee per month on contract, the challenges are very similar. You may not have taken the same unproductive hit when comes to capital expenditure, or absorbed the additional overhead of looking after yet another server from a management perspective, but you are still potentially paying for capacity that is not being used through the monthly rental fee. This can be eased in the same way as it is with on-premise equipment through the use of virtualisation to consolidate workloads onto fewer hosted servers, but an alternative type of hosting model has recently emerged that could provide a better option.

This is based on a much more flexible or 'elastic' approach which allows more of a 'pay as you go' billing model for consuming hosted server resources as you need them. In practical terms, such services revolve around the concept of the user being able to rapidly create and provision virtual servers in 'the cloud' that have no fixed specification and can grow or shrink in capacity as the load put upon them fluctuates. The customer pays for the resources actually consumed with no ongoing obligation, e.g. servers can be freely discarded with no penalty or other commercial implications when they are no longer needed. Going hand-in-hand with this model are online tools for users to self-provision and manage on a very granular basis, which is another way in which such services differ from fixed arrangements, where provisioning often needs to be carried out by the provider.

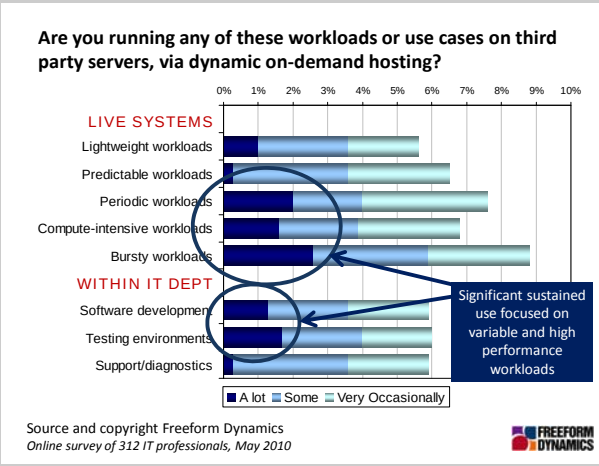
Right now, we are in the very early market for these kinds of 'cloud' based services, but while uptake is currently minimal, we can already see signs adoption with around one in twenty respondents to our survey indicating ongoing activity in this area (Figure 3).



While we are still in the early market and uptake of cloud based services is currently limited, we can already see signs of adoption in the mainstream.

Figure 3

With the relatively low level of activity, it is difficult to discern any obvious pattern from the above chart, but if we adjust the scale and magnify the more solid activity around dynamic hosting, we see some types of workload stand out, particularly those relating to more variable requirements (Figure 4).



Some types of workload stand out, particularly those relating to more variable and compute-intensive requirements.

Figure 4

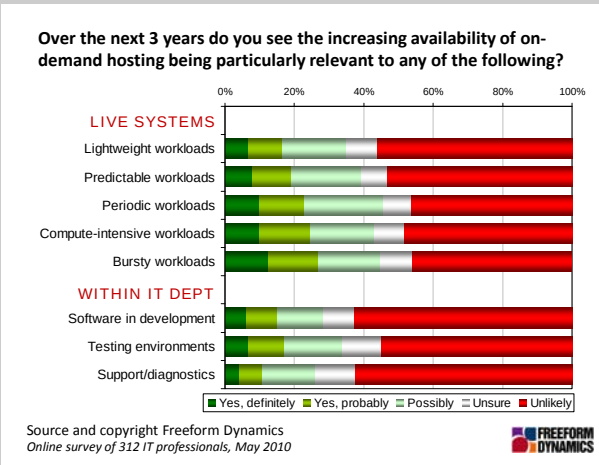
Despite the limited number of data points underlying this chart, it does illustrate that some IT professionals are already starting to see a place for on-demand infrastructure services, particularly to provide a more flexible and cost effective way of dealing with some variable server computing requirements. The prominence of compute-intensive workloads is also understandable as service providers enjoy certain advantages of scale when it comes to building high performance and high capacity infrastructures. Through the cloud computing approach, customers can benefit from this and potentially run exceptionally demanding workloads more cost effectively.

The picture that has emerged, however, also suggests that those with experience are putting less emphasis on the use of elastic services for workloads with more static demand profiles. Indeed, this isn't where the long term opportunity resides as such workloads gain little benefit from elasticity, and the commercial reality is that fixed capacity and/or contract based services will probably always be more cost effective when requirements are more predictable. To put it another way, flexibility usually comes at a price premium, and there is no point paying for it if you don't really need it.

Looking to the future

As service provider offerings and portfolios continue to mature, providing options for dealing with the range of different workloads and usage scenarios we have touched on in this paper, infrastructure services will become more relevant, accessible and attractive over time. Dynamic/elastic services will be an important part of this, and while uptake is limited today, it is encouraging that their role in dealing with variable and high performance workloads is acknowledged among those with experience.

One of the challenges, however, is around educating the broader audience. The majority of IT professionals, for example, are not convinced of the relevance of on-demand/elastic cloud services at the moment, with the jury still being out in the minds of many others (Figure 5).



The majority of IT professionals are not convinced of the relevance of elastic cloud services at the moment, with the jury still being out in the minds of many others.

Figure 5

It is also noteworthy from this chart that while we see an emphasis on variable and high performance requirements in terms of positive sentiment (in line with early adopter experience), many are also indicating the possible use of on-demand services for routine and predictable workloads. This is not a natural fit and suggests a degree of guesswork is going on, as a result of many not having yet had the opportunity or motivation to think through the requirements and solutions precisely and objectively.

Ironically, the over-enthusiasm of some in the industry for all things cloud[2] has almost certainly contributed to the level of dismissal and uncertainty we see. By portraying the 'elastic cloud' as the answer to all IT delivery problems, a line which experienced IT professionals find difficult to swallow, there has been a tendency for many to reject the notion of cloud as simply the next marketing bandwagon, with even those remaining open minded often being confused about where it fits into the overall scheme of things.

The good news is that many of the existing service provider incumbents that have been operating in the hosted infrastructure arena for years and have strong existing customer bases are starting to get in on the act. While this community has hitherto focused on the delivery of traditional fixed capacity offerings, they are extending their service portfolios to embrace the elastic alternatives, and going to market with a much more grounded, measured and inclusive view. The message from these providers is one of 'horses for courses', based on the reality that most customers will be looking for suppliers who can deliver a blend of fixed and elastic services with the corresponding range of pricing models, contract options and terms.

Activity such as this will help to drive a more solid understanding and appreciation in the mainstream over time, and together with some inevitable 'chilling out' of the more evangelical elements within the industry, we can expect interest and adoption in this whole area to ramp up over the coming couple of years.

Meanwhile, the best advice we can give to IT professionals trying to work their way through the promise and practicalities of cloud, IaaS, elasticity, and so on, is to focus on the requirements that need to be met in relation to specific workloads and applications, and select the delivery and execution model that is most appropriate on a case by case basis.

References

The following reports referred to in this document, along with a range of other free research on the deployment and use of IT, are available for free download from www.freeformdynamics.com.

1. Evolution of x86 Server Estates, Modernisation drivers and practicalities, November 2009
2. But is that really cloud computing?, The problem of ill-defined terminology, May 2010

About Freeform Dynamics

Freeform Dynamics is a research and analysis firm. We track and report on the business impact of developments in the IT and communications sectors.

As part of this, we use an innovative research methodology to gather feedback directly from those involved in IT strategy, planning, procurement and implementation. Our output is therefore grounded in real-world practicality for use by mainstream business and IT professionals.

For further information or to subscribe to the Freeform Dynamics free research service, please visit www.freeformdynamics.com or contact us via info@freeformdynamics.com.

About Microsoft

Founded in 1975, Microsoft (NASDAQ "MSFT") is the worldwide leader in software, services and solutions that help people and businesses realise their full potential.