

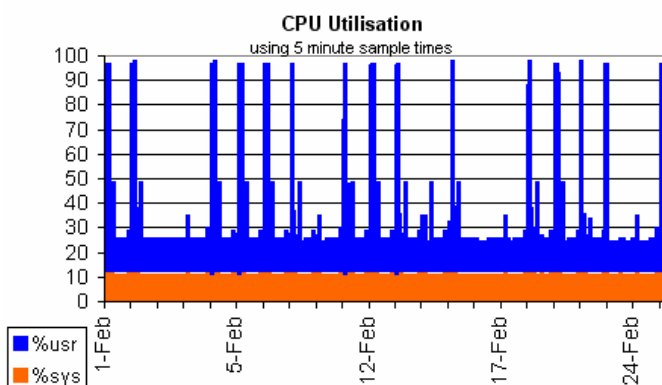
SERVER PERFORMANCE MANAGEMENT

Server consolidation

Here's a question for you: total up all the CPU time used by your servers in a given day. Take the percent utilisation for each box in each sample, multiply that by the number of samples. That gives you the total amount of CPU time used for a server in one day. Do this for each server and then add the totals up. Divide the total by the number of samples in a day and then by the number of servers you have. The answer is the average CPU utilisation for your organisation - pretty low isn't it?

Most servers are only running at a small fraction of their theoretical capacity. Some servers are designed to cope with a peak load that only occurs for a short time and some are only used during business hours, or overnight. If you have a functional Disaster Recovery architecture, it's entirely possible that half your servers, disks and networking assets are doing no work at all. A survey of a large utility company revealed that their average server utilisation (excluding the D.R. servers) was 19%. This same company had previously commissioned a business efficiency audit from a leading consultancy and came away with an "average" rating for their industrial sector.

As an example, look at this performance chart of a 4 processor server

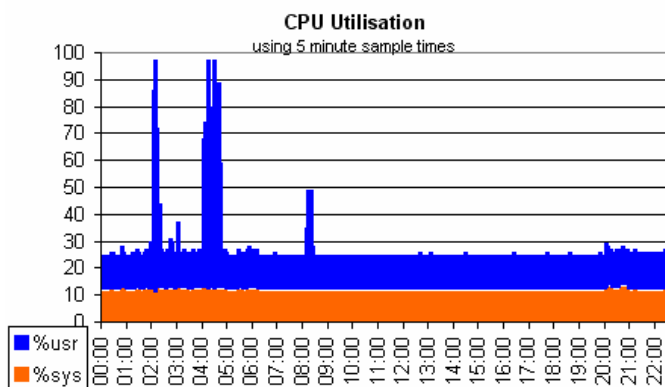


It's only job was to collect customer letters generated from a CRM system during business hours and transfer them to a centrally

SERVER PERFORMANCE MANAGEMENT

located print facility during a predefined timeslot. The spikes show that it had some very intense peaks - as we know to expect from batch operations. It also has a process running on one CPU that's using up 25% of the server's capacity. We reported this looping process to the owners, who were reluctant to do anything as they were unsure what it did. At the time we got involved and produced this analysis, it had been looping for 93 days. The ironic thing about this kind of activity is that for the utility company, this looping process would actually improve their average server utilisation.

Focusing on one single day shows the detail of what's going on.



There is a period of one hour, from 4:00 - 5:00 when this server runs near to full capacity. The earlier spike is a nightly backup and the spike at around 8:00 is a single auditing process. You can see that this runs on another single CPU and uses another 25% of the server's capacity. During the day customer letters trickle in, but take almost no resource to process: they simply have to be stored, waiting to be sent out. Given that the server only has 1 hour to transmit it's work, it is actually quite well specified to meet it's targets, the disadvantage is that the rest of the time, it is very inefficient.

When we talked about *quantifying benefits* back in part 1 of the book, I said that most of your time should be spent on high-value work, not on turning out reports and analyses that could be easily automated. This is one of those high-value areas. You have access to the work patterns of all the organisation's servers. The value you can create is to come up with workloads on some servers that

SERVER PERFORMANCE MANAGEMENT

would be a good match to the spare capacity on others. Obviously there's more to it than merely fitting CPU utilisations together: you need to consider all the other resources of the servers, plus the software they are hosting, the interfaces they have to other systems, any security/confidentiality issues and not least of all the licensing situation. If you have software that is licensed on a per-CPU basis, you want to avoid consolidating that software from a 2 CPU server onto a 48-CPU server, just because it has sufficient free resources - the additional licensing costs could be extreme. There may be ways to do this, we'll see this when we discuss virtualisation.

Strategy.

The biggest objection I get from organisations is their resistance to change. With the customer letters machine above, the response to my proposal to move the work onto another platform that was only busy during the day was met with objections about change, risk, flexibility and other intangible aspects. It boiled down to the simple points that the server was running OK, the money had already been spent and the systems architect had designed it like that, so they didn't want to change it. Their reluctance to even remove the looping process is all part of the same mind-set.

The time when organisations are most receptive to proposals for doing this kind of consolidation is when they have to make changes anyway. For example, when a server comes to the end of its design life or when there is a need to upgrade it for other reasons (Y2K was fantastic for this). Therefore it's easier to get your ideas accepted if you focus on consolidation plans for the older servers in your estate. When hardware come up for renewal - every 3-5 years on average, depending on your company's accounting arrangements, that is the time to propose not buying a replacement server, but moving the workload onto another platform that you have identified would be a good match for it.

SERVER PERFORMANCE MANAGEMENT

Virtualisation

Objections, like the ones above, can come from people who simply don't like the idea of losing "their" server. Whether it is business managers who see their influence shrinking or a tech-support person who accuses you of putting them out of a job. While it's not your place to judge the virtue of their claims, sometimes you can ease their objections by retaining the identity of the old server, either as a domain on a partitionable platform or by using server virtualisation, so that the software image of the server is kept. This can also help with addressing issues: the name and I.P addresses of the old server can be retained by the new image, so changes to DNS, LDAP, management and monitoring packages (including yours) won't have to be made. If you do take the virtualisation path, check the licensing terms of the server's software before you make any commitments.

The other main benefit of virtualisation is that you can reconfigure the virtual servers to provide more or less capabilities as circumstances require. The customer letters virtual server can be configured to run on 4 processors for a short time, when needed. After it's work has finished it can be reset back down to a level appropriate for it's needs. The released capacity can then be transferred to, for example, an Exchange server or to wherever the business needs extra power.

It is a big piece of work to add server virtualisation to a platform that did not have it installed when it was being set up. For this reason I would suggest you propose that at least some new servers are specified with a low-level virtualisation product installed when they are commissioned. By having them already configured you can easily add virtual servers later, as the need arises. A further benefit is that you can assign images for systems that only need a small amount of resource, but for whatever reason used to run on individual small servers, now they can run in images on a single hardware platform.

There are drawbacks to this type of virtualisation by using a low-level software layer. You don't get to use all the power of the underlying hardware, there are some overheads. Also, you still have to circumvent any other bottlenecks such as disk or network

SERVER PERFORMANCE MANAGEMENT

performance to use the full potential of the applications and these channels may be shared by all the virtual servers. You may also find that some applications won't run on virtual servers or that their end-user license impose onerous or expensive conditions on it's use.

You also have to be careful about lowering the reliability of your systems. A failure on the hardware which supports 12 virtual servers will take them all out, rather than just affecting one if they weren't virtualised. For this reason it may be prudent not to run all your firewalls for example on the same platform, but to mix services: firewalls, file-servers, active-directory systems etc. across multiple sets of hardware. That way if there is a failure, you don;t lose all your capabilities in one area, but a proportion (hopefully a manageable proportion) from many.