Inside Track
Executive Brief

FREEFORM DYNAMICS

IN ASSOCIATION WITH

FUJITSU

Dispelling the myth of the industry-standard server

What you thought you knew has changed

Freeform Dynamics, 2020

Introduction

In this era of cloud computing, servers might seem desperately old-fashioned – yet when you go to a cloud provider, their main offering will most likely be virtual **servers**, which must inevitably run on real hardware in a data center somewhere. Similarly, whether you are building hybrid cloud or its more inclusive sibling hybrid IT, you will need physical servers to run the private cloud and other local elements, whether that's older applications, hypervisors, container frameworks, or whatever.

So while 'the cloud' gets the attention, servers still matter. At the same time though, there is the myth of the 'industry-standard server', typically visualized as one of dozens of identical flat boxes, all mounted into racks and with their lights flashing.

Yet servers – whether physical or virtual – come in a wide variety of configurations, with lists of optional extras. What cloud providers and server vendors alike know is that there is no such thing as a standard server: it can vary massively depending on service level requirements – and especially on the workload involved.

In this paper, we will discuss some of the many options available to the server buyer today and how these address different workload needs. We will also compare on-site, cloud and hybrid approaches, and ask what servers of the future might look like.

What matters is the workload

So why are there so many options on server price-lists? The answer is simple: workload requirements vary immensely. And today it is not just a question of how much memory and processor power your general-purpose application mix will need – workloads and their platform requirements are becoming more diverse and demanding than ever, as a few examples will demonstrate.

'Standard' servers

Long gone are the days when the server simply stored your files on its local hard disks, from where they could be backed up or shared with colleagues. Today's storage server might also provide block and object services, and replicate data to cloud storage or to a mirrored second system in another location.

And while the need for print servers has largely disappeared, with most printers connecting directly to the network, many other 'industry-standard' applications and tools need to run on a server. They include network infrastructure and security workloads such as DNS and firewalls; email, conferencing, messaging and other collaboration tools; and of course the many business process workloads such as ERP, CRM, databases and so on.

Mission-critical

Even a 'general purpose' server is essential if the applications it runs are core to your operations. These servers therefore require non-standard features for fault tolerance, such as RAID storage and reserved (spare) system boards to replace failed boards without downtime. Alternatively, they can be built as dual mirrored servers for high availability. You may also want to segment or separate your servers physically, instead of running multiple workloads as individual VMs all on one machine, so a single hardware failure doesn't affect every application.

AI and machine learning

Almost every area of technology is trying to exploit the hugely diverse field of artificial intelligence in some way, shape or form. The most popular AI-derived tools are machine learning (ML) and its subset deep learning (DL), but others with commercial relevance today include natural language processing and computer vision.

Even within just ML/DL, the two main segments – training, where the system learns or is taught, and inference, the operational part where it makes assessments or predictions – have different needs. Many training workloads are highly parallel, and run dramatically better if the main processor can offload the ML/DL algorithms to one or more GPUs, or graphics processing units. Depending on the individual application, the inference workload may require just an x86 processor. However, if it involves continuous retraining, it too might need GPUs.

Because it was originally designed to calculate thousands of pixels, each of them relatively simple, a GPU is at heart a massively parallel processor. To begin with, AI researchers used actual graphics boards, but increasingly, the same chip technology is now designed into modules purpose-built for parallel processing work.

Supercomputing and HPC

Other researchers and software developers have also picked up on the widespread use of GPUs as computing elements, and this has broadened and accelerated the use of high-performance computing (HPC) more generally. For example, a 'deskside HPC' might comprise an x86 host server with a number of GPU servant modules.

Supercomputing too has benefited from the increasing power of the x86, with fewer examples now using other processor types. However, while the x86 processors in today's supercomputers might be standard, many of the system architectures are not. Massively parallel systems combining multiple nodes are common, each node containing multiple GPUs and multiple x86 processors, and each processor having multiple compute cores. The nodes typically interconnect via specialist clustering schemes, network topologies and file systems, and may also have unusual features such as cluster-shared memory.

Diversified virtualization

After two decades of evolution, the virtual machine (VM) is now mainstream. For example, cloud computing and hyper-converged infrastructure (HCI) are largely built around virtualization. Data centers are now routinely virtualized, and in some cases entirely software-defined, and software vendors avoid dependencies by distributing applications as complete packaged VMs.

VM hosts typically require plenty of memory, while modern multi-core processors make it possible to allocate a VM its own core or cores. Other server components may need to be VM-aware too, for example the network adapters, and the I/O pattern generated by a VM host can vary dramatically.

The VM is no longer alone, however. We now have numerous other technologies that employ software abstraction in some way, such as streaming virtual applications or entire virtual desktops, and then we have containerization, exemplified by Kubernetes. Here, only the elements unique to that workload, application or application set – the 'user space' – are packaged into a virtualized container that can share a single host operating system with other containers, or indeed with VMs.

Containers can run within a VM or directly on the operating system, and are now accompanied by ever more granular forms of virtualization. Individual software modules or processes can be packaged as microservices or as a Function-as-a-Service (FaaS) infrastructure. Because services only run on demand and are not tied to a particular server, this approach is often referred to as 'serverless' – although of course every service must ultimately execute on physical server hardware somewhere!

Complex optimization problems

Many industries feature processes that could benefit from real-time optimization, but which are so complex that they cannot be optimized fast enough using standard computing techniques. Think for example of optimizing the layout of a busy warehouse, the traffic flow in a city, or the investment portfolios held by a bank's customers.

These are some of the problems that quantum computing may one day address, but for now quantum technology is too complex and expensive. However, there are so many problems of this kind that we have seen the evolution of quantum-inspired computing, where concepts such as quantum superposition are instead implemented either in software or in digital hardware. Some of these quantum-inspired technologies are already solving important business problems today.

Scale-up or scale-out?

Some servers are designed to grow or expand as a single 'unit' (the scale-up approach), while others grow by adding more units (the scale-out approach). Again, which is right for your needs will depend on the workload. For instance, modern containerized or webscale applications can be a good fit for scale-out technology, which is often cheaper and simpler to implement, whereas monolithic applications might prefer scale-up.

On-site, cloud and hybrid

While some organizations are able to migrate entirely to the public cloud, the options for most are hybrid. For example, we may have existing mission-critical applications running on traditional IT infrastructure, or regulatory issues that restrict where certain workloads can run, or highly predictable workloads where a private cloud is more cost-effective than using public cloud. In each of these cases, we need servers, which we might choose to house on-site or remotely in a co-location center.

Increasingly, we may also have a need for edge computing too, whether that's small servers in remote locations, or the smart machines and other devices that we think of as the Internet of Things, or IoT.

Each of these cases has a different workload profile, and therefore requires different hardware. For example, a cellular base station, a factory controller, and the HCI system hosting our private cloud all contain servers, but they are very different devices.

At the same time, we must also consider the increasingly dynamic and hybrid nature of IT, where a workload's profile can change during its life. A Kubernetes container might need to be pushed to the edge to improve responsiveness, say, or an application might be migrated from the public cloud to on-site systems in order to reduce costs. This means we also need to consider workload portability when making server decisions.

Becoming data-driven

Any organization serious about becoming fully data-driven in the future needs to plan its future server infrastructure. At the very least, it needs an integrated hybrid multi-cloud approach, combining appropriate cloud and on-site server platforms on a foundation of integrated and abstracted storage. For more details on this, and on how to assess and plan your data-readiness, see our recent report: *The road to becoming a data-driven business*.

As part of this planning process, you also need to consider how and where you will acquire the server infrastructure, and how you will get it all working together. Most likely you will have different servers to meet diverse needs, but that should not mean multiple administrative silos – a single management framework is the aim. Conversely, your diverse workloads may also imply diverse financial models, so you may need access to a range of procurement options.

So what's in a server today?

Not only are there multiple suppliers of x86 processors, with Intel and AMD the best-known, there are several other server platforms in use. In particular, servers based on energy-efficient ARM processors – better known for their use in mobile phones and other smart devices, but also available in much more powerful variants – are growing in popularity and are gaining more software support.

Then there are all the add-ons, such as GPUs, Digital Annealers, RAID storage controllers with Flash backup, and high-bandwidth network adapters for the likes of 100Gbit Ethernet and 128Gbit Fibre Channel. Also of interest here is the smart-NIC or DPU (Data Processing Unit), which is an off-load engine for processing network traffic, for example packet processing, encryption, network virtualization, RDMA acceleration and more.

To make use of options such as these, the server must not only have enough PCIe slots, but those slots must be fast enough – PCIe has evolved through several generations. In addition, the server must have redundant power supplies (for fault-tolerance) that are both efficient and capable of powering all this hardware.

Coming from the other direction, there are also systems where the x86 server is the add-on. For example, there are mainframes with x86 'application units', enabling their owners to consolidate Windows-type workloads with their mainframe applications. This serves to remind us that mainframes still run the core business systems of organizations around the globe, and will continue to do so for years to come.

Even within the x86 family, not all servers are the same – far from it! There are many different server architectures, some with multiple x86 processor packages, and each processor typically has multiple compute cores. Because it is becoming harder to build faster cores, adding multiple cores is one way that processor designers have continued to increase performance. A system with a single 64-core processor may appear similar to one with four 16-core processor packages, but will perform differently depending on the workload (see the scale-up/out discussion above).

And looking forward...

The increasing use of AI, and in particular of ML/DL techniques, is driving greater use of GPUs as application accelerators, both for on-site servers and in public clouds. At the same time, we are also seeing AI-specific instructions being built into otherwise standard x86 processors, making it even easier to add a degree of AI acceleration to a server by choosing the right processor package.

The use of technologies such as containerization, 'serverless' and private cloud is growing too, and these workloads have their own specific platform expectations. Again, some server designs are better suited to this than others. For example, we can expect to see more adaptation and extension of modular converged and hyper-converged infrastructures to address this area.

Although the mainstream use of Quantum computing is still several years in the future, Quantum research is already spinning off computing techniques that can offer significant advantages, even when running on digital technology. A server that incorporates this Quantum-inspired technology has the potential to dramatically accelerate certain workloads, such as combinatorial optimization. In some cases, it can execute tasks that would not be possible using standard x86 systems, such as real-time routing or prioritization.

Summary

The growing diversity of workloads and of their desired service levels has spawned a server landscape that is at least equally diverse, and perhaps even more so. Tailoring a server to its workload will generally yield better and more cost-effective results. That means picking the right processor(s), giving it enough memory and storage, ensuring a big enough power supply and sufficient internal expansion capacity, and if necessary using the latter to add hardware acceleration engines.

Almost all of this is feasible in the cloud as well as in your own data center, yet few organizations of any scale can operate without any servers or IT infrastructure of their own. The model for most will instead be a hybrid one, combining public cloud services, edge infrastructure, and servers that host private cloud and traditional (classic) IT applications and services.

The keys to all of this are openness, planning, and supplier selection: Openness to the fact that there is rarely a single way to solve workload and infrastructure challenges. Planning to match the server's capabilities and capacities to its target workload, and to ensure that a hybrid IT infrastructure still has a single management framework. And supplier selection to make sure that whoever you work with has both the necessary technical and business breadth, in terms of infrastructure technologies, financing, support and so on, plus the essential depth in areas such as hardware and software expertise.

About Freeform Dynamics

Freeform Dynamics is an IT industry analyst firm. Through our research and insights, we help busy IT and business professionals get up to speed on the latest technology developments and make better-informed investment decisions.

For more information and access to our library of free research, please visit www.freeformdynamics.com or follow us on Twitter @FreeformCentral.

About Fujitsu

Fujitsu is the leading Japanese information and communication technology (ICT) company offering a full range of technology products, solutions and services. Approximately 140,000 Fujitsu people support customers in more than 100 countries. We use our experience and the power of ICT to shape the future of society with our customers.

For more information, please visit www.fujitsu.com

Terms of Use