Enterprise Computing Insights

What is an Enterprise Server?

#5 IN A SERIES

An "enterprise server" is a very robust, large scale computing system. But, like many other complex products, trying to define a product based solely on one characteristic, like "scale", can be difficult or even misleading. This article explores several of the characteristics that make an enterprise server "robust" and define the qualities it is expected to exhibit. An enterprise server is sometimes referred to as a "mainframe".

As its name implies, the target market for an enterprise server is a business or enterprise. An enterprise server is most often found in large enterprises such as banks, insurance companies and manufacturers, but an enterprise server can be equally at home in a relatively small enterprise such as running a county government. So, while a large enterprise with hundreds or thousands of people is not a requirement for using an Enterprise server, it is true that they are most often found in large organizations that need to service the computing needs of a large number of customers and/or as employees. This requires an enterprise server to be able to deliver considerable compute capacity continuously.

Flexible Capacity and Scale

Another aspect of compute capacity provided by an enterprise server is the ability to provide flexible compute capacity to meet the changing needs of a business. As a business grows by acquiring new customers or acquiring other businesses, the enterprise server needs to grow (be upgraded) right along with it. Ideally, the enterprise server can even grow (and shrink) its capacity to match the hourly or daily needs of the business. For example, a surge of activity on a "black Friday" shopping day may need additional compute capacity for a day or two, but after that, the additional capacity may become idle and wasted as business returns to normal. Adding capacity "on demand" and then returning it when not needed is the ultimate in server capacity flexibility and an exclusive function of today's enterprise servers. Further, the enterprise server capacity must cover a wide range so as to be able to service the needs of both large and small enterprises. A range of over 300 times capacity from the smallest (single CPU) server to the largest (currently 120 CPU) server in an enterprise server family of servers is not unusual with today's zEC12 family of enterprise servers. And as an enterprise server grows in capacity, the majority of a new increment of capacity must be available to do the work of the enterprise, not wasted as "overhead". This means that the server must scale up efficiently and be able to run at high utilization to make best use of the compute resources. CPU utilization of 90-95% is not unusual on an enterprise server whereas distributed servers typically run at 50% or less utilization. So, while compute capacity is one of the key characteristics of an enterprise server, it is not the only characteristic that defines an enterprise server. Let's look at some of the many other characteristics that are equally important when defining an enterprise server.

Mixed Workloads

Because of the large number of end users and the variety of work found in a business organization, an enterprise server needs to be able to run a diverse set of workloads effectively.

Transaction processing[1], batch[2], and decision support workloads are all commonly run on an enterprise server. In an unconstrained environment where there are plenty of resources (CPU, memory, I/O, etc.) to satisfy all the workloads, there is no contention. But more often than not, there are resource constraints and competition for resources. Under these conditions, there needs to be a mechanism to control access to the scarce resource(s). Most servers use a static resource prioritization scheme to accomplish this. However, as workloads increase and decrease over time, this can result in resources being reserved for work that doesn't need them and therefore kept from workload(s) that could use them. An enterprise server uses a sophisticated workload manager component of its operating system in conjunction with enterprise server hardware (such as channel subsystem priority queuing) to allow dynamic resource allocation based on ever changing workload needs. This ensures that workloads are managed toward their business objective (e.g. one second response time for a mission critical transaction processing workload) over time and that resources are applied to the most important work running on the server.

Robust Input/Output

One of the hallmarks of commercial data processing is that the programs used typically have a high level of I/O activity (more so than CPU activity). Think, for example, about the data processing activity associated with credit card authorizations. Consequently, an enterprise server needs to have a robust I/O subsystem designed to service a large number of I/O requests simultaneously and transfer massive amounts of data to and from memory and I/O devices (current enterprise servers can have an aggregate data transfer rate of about 384GB/sec). On an enterprise server, the CPUs only initiate the I/O processing; they are not directly involved in the relatively long (millisecond) I/O operations that actually transfer data. This is unlike the I/O operations on a personal computer, for example, where the CPU is directly involved in the actual data transfer between an I/O device and memory. The enterprise server I/O subsystem therefore frees a CPU after initiating an I/O operation so that it can execute the instructions for one program while I/O is underway for another. While an enterprise server can run compute intensive engineering/scientific programs, its real forte is commercial data processing with heavy I/O activity.

High Availability

With hundreds or perhaps thousands of users dependent upon it, an enterprise server cannot fail. The loss of an enterprise server is very expensive in simple costs like lost productivity but can be even more expensive in terms of lost business (e.g. for a stock brokerage) or intangible costs like reputation (e.g. for ATM or home banking). Enterprise servers have considerable technology incorporated in their design to mask failures and keep the server running. There are basic availability technologies like redundant components (e.g. for power supplies and cooling units) that are designed to avoid single points of hardware failures. At the chip level, the CPU and instruction processing are at the heart of an enterprise server. These are protected from transient and permanent errors using error correcting codes and a check pointing mechanism. For transient errors, instructions can be retried on the same CPU. For permanent errors, the instruction check point state is restored on a spare CPU and instruction

[1] See ECI No. 3
[2] See ECI No. 4

processing continues there without disrupting any program execution or the program's users. (Note: an enterprise server always has at least one spare CPU available to take over for a failed CPU). The newest enterprise server added a new memory error protection technique called RAIM (Redundant Array of Independent Memory) to the existing parity checking and other protections. RAIM is conceptually similar to the RAID technology used on disk drives and helps ensure that memory errors do not impact applications and end users. Using these techniques, and many others, results in an enterprise server mean time before failure (MTBF) measured in decades (on the order of 30-40 years).

Virtualization

An enterprise server excels at server "virtualization" – the ability to make the physical resources (CPUs, memory, I/O channels, disks, network adapters, etc.) available on a real server act like multiple servers. On an enterprise server, the partitioning and mapping of virtual resources to real resources is done in microcode (as opposed to software for example) and consequently is very efficient and low in overhead. The ability to virtualize servers is becoming more popular even on personal computers since the benefits of virtualization are compelling. By consolidating real servers onto virtual servers on an enterprise server, a business can reduce data center resources such as floor space, electricity, cooling, etc. Virtualization of servers also provides a convenient way to build complex test environments (e.g. a clustered group of servers). Virtualization can also be used to satisfy the need for new servers by dynamically provisioning new virtual servers instead of new real servers. And finally, virtualization can be used to isolate a particular workload or application, run different types of operating systems, or even different levels or releases of the same operating system.

Security

The security characteristics of an enterprise server are legendary. A number of technologies contribute to the security of an enterprise server. Memory accesses are protected using a security key. Memory is divided into 4K blocks and each block has a storage key associated with it. When a program accesses the data in a block, the program's storage key is compared to the memory block key to ensure that they match. If they do not, the program cannot access the storage. This key checking occurs on every storage access and can prevent "random stores" and buffer overflows from corrupting or overlaying programs and data as well as unauthorized access to operating system control data. The latest enterprise servers also include a cryptographic co-processor for each of six cores on each processor chip. These cryptographic co-processors can be thought of as hardware accelerators for the encryption / decryption of data using industry standard algorithms (DES, TDES, etc.). These hardware capabilities are used in conjunction with other software provided security functions such as an Authorized Program Facility (APF), a security manager such as the Resource Access Control Facility (RACF) and extensive cryptographic key management by the Integrated Cryptographic Services Facility (ICSF).

Enterprise servers are designed with compatibility in mind so that investments in application software, for example, are protected so that software does not have to be re-purchased or re-written when new hardware or a new operating system is introduced. This is especially important to businesses that use

enterprise servers because of the magnitude of the investment in software which can be in the millions of dollars. Compatibility is less of a consideration for personal computers and is often the reason you need to occasionally upgrade a game or application to a newer version when an older version is no longer supported by new hardware or a new version of an operating system.

Serviceability

An enterprise server needs excellent serviceability so that any problem that is encountered can quickly be identified and fixed so as to not impact the users of the server. If a laptop fails, a user can usually find another personal computer to use while the failing laptop is being serviced. That is not the case with an enterprise server. So the enterprise server uses serviceability tools and techniques to make sure that when an error is encountered, the right data is captured to avoid having to recreate the problem for diagnostic purposes (this strategy is known as first failure data capture (FFDC)). Even if an error is masked from a program or end user by recovery actions (for example fail over to a redundant component) some action still needs to occur to eventually repair or replace the component that failed. Modern enterprise servers can automatically "call home" to the manufacturer to initiate a service call that will repair or replace the failing component.

Summary

To summarize, an enterprise server is much like any other server in some respects – it has the usual components you would expect to find in a computing system like a power supply, CPUs, memory, I/O connections and a cooling mechanism. But even some of these "common" components are unique and help provide capabilities like no other compute server. Its capacity, range of capacity, and ability to effectively use capacity are unmatched. It manages multiple competing workloads effortlessly as the workloads vary with the needs of the business it serves. It is able to handle both I/O intensive workloads as well as compute intensive workloads but excels at the former. It security and availability characteristics are legendary. The virtualization of resources available on an enterprise server can be done using multiple technologies and has evolved and been optimized over several decades. And finally, the enterprise server evolves in a way that allows prior programs to run unchanged – extreme compatibility protects the customer investment in hardware and software.

We will explore other aspects of Enterprise Computing in subsequent articles.