Vcap CALC

Virtualization technology has been a benefit to IT organizations and the businesses they support; reducing cost and administration overhead, improving system resiliency, and increasing the availability of critical applications. However, these benefits can be so great that IT departments might start to think of the capacity of these new virtualized environments as "free and infinite" – which is hardly the case.

Like any physical system, virtual environments have resource limitations, and running out of capacity can have significant consequences. For example, if a virtual environment runs out of memory, then all the dependent virtual servers can go offline or at least suffer severely degraded performance.

Once a virtual environment has been established, it is important to immediately start monitoring and managing its capacity. Without knowledge of an organizations capacity, rush purchase decisions or downtime can result.

Many physical servers (each referred to as Virtual Hosts or Host Servers) make up a virtual environment (referred to as a cluster), utilizes technology to effectively combine the individual resources of each server, and presents it as a total pool of available capacity. This capacity can be arbitrarily assigned to Virtual Machines (referred to as VMs, Virtual Servers, or Virtual Guests). In most clusters, there are three main limiting resources provided by the physical server: CPU, Memory, and Network. In high-availability (HA) clusters, Disk capacity is often provided by an external solution, such as a SAN or storage array.

CPU CAPACITY can be measured in several ways, such as total GHz vs consumption, or vCPU per physical core. Because most servers spend their time idling and waiting for requests, best practices allow for the oversubscription of CPU resources. Even on a fully loaded virtual host, it is typical to see CPU GHz consumption at ~10% of overall capacity. Other factors, such as CPU scheduling, can play a more significant role in performance in a loaded virtual environment, so the tool uses a ratio of vCPUs to physical cores to measure capacity.

MEMORY CAPACITY is usually the most limiting resource in a virtual environment, because unlike CPU, it is *not* a resource that should be oversubscribed. If memory resources become constrained, all virtual systems can suffer severe performance degradation, as memory management can push the data in memory to the significantly slower disk-based resources.

NETWORK CAPACITY limitations are typically bound to the supporting network infrastructure, the specific and fluctuating demand on each system, and many other complex factors. As such, it is not the focus of this document and tool. For virtual hosts in today's market, 10 Gbit network is enough to provide network to downstream VMs.

STORAGE CAPACITY can be more complex to manage than it would seem. Hard drives, at the base of each type of storage, have both a "bit" capacity and a "performance" capacity. Bit capacity is typically measured in Gigabytes (GB) or Terabytes (TB), whereas "performance" capacity is typically measured in the speed at which data can be written into or read out of the disk (referred to as Ins/Outs per second, or IOPs).

Additionally, storage arrays have benefited from many technological advancements to help manage capacity and reduce waste, such as thin provisioning, data deduplication, and flash/hybrid or virtualized storage. The Oakland County capacity tool provides a very high-level capacity report on the bit-capacity of a virtual environment, but it is recommended to rely on the actual storage array for the most accurate performance and capacity reporting.

Understanding the current capacity of the virtual environment allows the business to make well informed decisions about how – or if – the remaining capacity should be allocated or more capacity should be acquired. Detailed future planning can be performed by the business and its departments: new business projects will frequently need new virtual servers, or as Software as a Service (SaaS) solutions are adopted, this can result in existing virtual servers being decommissioned.

There is also a hidden capacity that should be accounted for in any planned maintenance activities. This "bubble" of resource consumption can exist for the duration of a project where a new version of an application is deployed alongside its current, production counterpart. Only when the business has cut-over to and accepted the new version can the original systems be decommissioned. There should always be enough capacity to deploy your largest application, or the business runs the risk of accruing significant technical debt or impacting system performance.

To ensure that Oakland County's systems remain healthy and available for the foreseeable future, we began a practice of Capacity Reporting and Management, coupled with detailed future planning, and developed a tool to ensure consistent and frequent monitoring. Tracking capacity over time helps us understand how environmental changes affect us, look for trends, and make better informed decisions. Additionally, planning out resource utilization in the future lets us plan more effectively for major projects, and minimize the size of our planned maintenance bubble.

Finally, though outside the scope of this tool, Oakland County highly recommends the practice of right-sizing virtual servers. There are other tools available (such as VMWare's vRealize Operations) that can monitor resource consumption over time and make recommendations about proper VM sizing. This can significantly improve the performance of all virtual servers and ensures that the available resources are used most efficiently.

Capacity Best Practices & Recommendations

- ✓ Always have a minimum 1 virtual host's worth of resources available so that neither hardware failure nor performing regular maintenance can on the virtual infrastructure impacts the business (considered "full" in the capacity report).
- ✓ Maintaining two hosts of available capacity is preferred, to provide enough planning time for hardware procurement, "bubble" management, and a lower likelihood of impacting the business (considered "warning" in the capacity report).
- ✓ Minimum of 3 physical hosts per virtual cluster.
- ✓ Oversubscribing vCPUs at no more than 3.5 vCPUs per physical core to reduce performance being impacted by CPU scheduling. This is built into the capacity calculations of the capacity report.
- ✓ No overprovisioning of memory. This is built into the capacity calculations of the capacity report.
- ✓ Rely on the disk array to report on storage capacity, and vendor threshold recommendations if available. Broadly, 80% is the defined warning threshold for storage, and 90% is defined as the full threshold. This is because typical storage arrays tend to experience degraded performance at above 90% used.
- ✓ Right-size virtual servers, to most efficiently utilize the available resources.