Memory Efficiency in VMware-Virtualized Data Centers

In this document, we will see how to take advantage of Transparent Page Sharing, Memory Limits, Memory Ballooning and Memory Rightsizing in the best way with Eco4Cloud solutions.

Field results show that, even in a small server farm, tens of TeraBytes of RAM can be saved without compromising performances, but actually increasing them.

This allows to reduce and/or delay CapEx to buy new RAM or even entire physical servers altogether, as memory is typically the bottleneck resource of a virtualized environment.

Eco4Cloud — www.eco4cloud.com | Phone +39 0984494276 | E-mail info@eco4cloud.com

Memory Reservation

Active VMs enjoy the memory reservation feature, while the less active ones do not pile-up and waste idle memory.

As described in VMware’s white paper Understanding Memory Resource Management in VMware® ESX™ Server, ESX offers several memory levels:

- **Host physical memory** refers to the memory that is visible to the hypervisor as available on the system.
- **Guest physical memory** refers to the memory that is visible to the guest operating system running in the virtual machine.
- **Guest virtual memory** refers to a continuous virtual address space presented by the guest operating system to applications. It is the memory that is visible to the applications running inside the virtual machine.

Memory reservation is essentially the process guaranteeing that the guest physical memory is always backed by the host physical memory, whether the ESX server is under memory pressure or not (memory pressure will be covered by the following articles of this series).

Memory reservation is a critical configuration that must be dealt with carefully, in fact this is what VMware states in its vSphere Resource Management guide:

Use Reservation to specify the minimum acceptable amount of CPU or memory, not the amount you want to have available.

When specifying the reservations for virtual machines, do not commit all resources (plan to leave at least 10% unreserved). As you move closer to fully reserving all capacity in the system, it becomes increasingly difficult to make changes to reservations and to the resource pool hierarchy without violating admission control.

Memory reservation strategy advised by Eco4Cloud

Memory reservation is then an essential feature to guarantee performance to virtual machines, but aggressive memory reservation can have BAD effects on other virtual machines, such as poor consolidation, awful performances, or inability to power on new virtual machines because of admission control rules.

Eco4Cloud advises to use a *de-facto* standard strategy, which consists of grouping virtual machines in priority-based resource pools, defined at each cluster level, and setting reservations at resource pools level.

By doing so, very active VMs enjoy the memory reservation feature, while the less active ones do not pile-up and waste idle memory.

The reserved memory right-sizing for each resource pool can then be set following several strategies, based on the active memory of the virtual machines in the resource pool.

Transparent Page Sharing

Eco4Cloud's workload consolidation is fully aware of the TPS memory reclamation capabilities.

VMware defines Transparent Page Sharing (TPS) as *a method by which redundant copies of pages are eliminated*.

TPS is an ESX/ESXi level process, scanning every page of the guest physical memory, searching for sharing opportunities. A full bit-by-bit comparison is performed between pages having same TPS hash values and, if two pages match, then the guest-physical to host-physical mapping of the candidate page is changed to the shared host-physical page, and the redundant host memory copy is reclaimed, thus reducing memory consumption and enabling a higher level of memory over-commitment.

TPS works with memory pages with “regular” small pages (i.e., 4KB contiguous memory regions), while with newer OSes and/or hosts with hardware MMU systems that make use of large pages (i.e., 2MB contiguous memory regions), the sharing is postponed until memory pressure happens, and ESX/ESXi breaks each large page into 2048 small pages to ease memory swapping (and activate TPS, of course).

Put that simple, one can challenge this approach as the probability that two memory pages composed of 4KB, that is 32768 bits, fully match is clearly infinitesimal: 1/2^32768, which is right.

Actually, it is an established fact that when different VMs run the same OS and/or applications, and have same data, they WILL have an amount of memory pages that fully match, by design.

In addition, there are several situations (e.g., OS boot) where guest OS zeroes-out many memory pages; that is, it deletes data of a each memory page by over-writing “0” on every byte of the page. Obviously every zeroed memory page matches with the other zeroed memory pages, and gets shared by TPS.

TPS can also lead to performance increase with reference to memory latency for large VMs in systems composed of Non-Uniform Memory Access (NUMA) nodes. More insights on TPS and NUMA nodes are available here and here.

Eco4Cloud strategy for TPS maximization

Eco4Cloud’s workload consolidation optimizes the energy efficiency of physical hosts by maximizing the number of VMs running on each host, while increasing performances. In order to do that, it computes assignment scores of VMs to hosts. A high assignment score makes a vMotion more likely to be issued by Eco4Cloud, and the higher, the better.

Eco4Cloud’s workload consolidation is fully aware of the TPS memory reclamation capabilities. In fact, the homogeneity of guest OSes on each physical host plays an important role in the Eco4Cloud score assignment, increasing the odds of memory reclamation through TPS, and leading to higher consolidation levels.

Memory Ballooning

Smart Ballooning frees up unused guest memory, making it available to the host constantly.

The original difference between legacy and virtualized environment is the presence of a hypervisor, such as VMware ESX/ESXi. This kind of software has the main goal of decoupling operating systems and physical hosts, and inject intelligence in resource allocation to applications.

Ok, but why such an obvious introduction?

This OS/host decoupling, or VM isolation, introduced by several virtualization platforms (e.g. VMware ESX/ESXi and others) carries along a drawback, as the guest operating system is not aware that it is running inside a virtual machine and is not aware of the states of other virtual machines on the same host. When the hypervisor runs multiple virtual machines and the total amount of the free host memory becomes low, none of the virtual machines will free up guest physical memory because the guest operating system cannot detect the host's memory shortage altogether. Ballooning makes the guest operating system aware of the low memory status of the host.

VMware achieves that through the installation of a balloon driver, via VMware tools. The balloon driver will take charge of making the guest operating system aware of host memory shortage, by allocating memory, pinning it and making it available to the host, when it will communicate its memory pressure state to the balloon driver.

The problem is that, as soon as memory pressure is detected at the host level, all the balloon drivers in each VM will scan the entire VM memory, searching for free memory page, which will lead to two intrinsic drawbacks:

- high CPU utilization during host memory pressure situations;
- the memory allocated by the balloon driver will be available at the host level, but guest OS will consider it as not available, so memory ballooning is fundamentally a tool transferring memory pressure from the physical host to the virtual machines.

The final result is that memory ballooning deteriorates application performances.

Smart Ballooning

Smart Ballooning is a software developed by Eco4Cloud for virtual machines memory management for VMware's virtualization platform. Smart Ballooning is different from "simple" ballooning in its selectiveness.

In fact, Smart Ballooning constantly monitors virtual machines performances metrics, and when it spots a possibility of memory waste at guest level, it induces memory ballooning selectively into that virtual machine. This approach lead to several advantages:

- Smart Ballooning frees up unused guest memory, making it available to the host constantly, not only during host memory pressure;
- host memory pressure situations are less likely to happen as there is always more RAM memory available than without Smart Ballooning;
- less CPU usage due to memory ballooning exactly in those moments (memory pressure) when the hosts are under stress;
- memory is often the bottleneck resource in a virtualized environment, so a higher memory over-commitment lead to higher consolidation ratios.

VMs rightsizing

Troubleshooter reports oversized and undersized VMs, and allows resources reconfiguration

Assigning the right amount of resources to the virtual machines helps to ensure maximum performance of your workloads and the efficient use of your underlying hardware. Today adding resources to a virtual machine when required is just two clicks away, so we'd better get rid of our (bad) habit of over provisioning.

Often times the hardest part of right sizing VMs is convincing the application owners that they don't really need as many resources as they think they do. In the physical world, adding more CPUs and more memory could unlikely degrade performance. In the virtual world, on the other hand, this is certainly not the case.

In fact, many often wonder if there is overhead when creating a large virtual machine with excessive assignments of vCPUs and RAM that may or may not be used, letting the ESX/ESXi sort it out. Actually, adding more virtual memory could actually result in making matters worse if the host is contended, increasing memory ballooning and swapping.

Troubleshooter

Eco4Cloud Troubleshooter is a software which performs continuous monitoring of well-known virtualization options and sets immediate warning/alerts when wrong configurations are detected.

Among other misconfigurations, Troubleshooter reports oversized and undersized VMs, and allows resources reconfiguration. Troubleshooter achieves that by analyzing past workloads and suggesting the optimal resources assignment.

Useful Links

- Eco4Cloud – www.eco4cloud.com
- Architecture and Requirements - www.eco4cloud.com/download/E4C-Architecture-and-Requirements.pdf
- Saving energy in datacenters through workload consolidation - www.eco4cloud.com/download/e4c-white-paper.pdf