Dynamic memory management in cloud computing.
Contents..

- Introduction to cloud memory.
- Importance of memory management.
- Memory management techniques.
- DRAMS as the future
- Conclusion
- References
Introduction to cloud memory

-Where is it stored?
  - Everywhere (mostly of the time based on a policy, with no latency)

- How is it Shared?
  - In case of a public cloud the same cluster may be used as a swap for multiple end users or applications.

- How is it accessed?
  - Via a volatile memory which provides less latency than a read write operation.

Figure: Golondrinas Architecture
Importance of memory management

- **SaaS, IaaS** etc are in need of smart memory management protocols to be integrated in Cloud in order to get rid of the latency and load balancing issues.

- On demand resource allocation is the key in optimizing the data efficiency of the Cloud.

- There is a huge drainage of resources across the Cloud platform if the resources are allocated and left idle.

- Continuous checks and monitoring is necessary to get hold of the idle resources.

---

Fig. 1. Memory provisioning approaches under ideal and real conditions.
Importance of memory management

- The best example in this domain is the Amazon’s Elastic Compute Cloud (EC2).

- EC2 Cloud only allocates the resources to the virtual or real entities on demand.

- As Cloud environment is dynamic and volatile, there is a strong need to inculcate the dynamic memory allocation trends in the Cloud based systems.

- The memory allocation is done with a technique known as swapping.

- Virtual swap management mechanism (VSMM) is the most appropriate solution up till now for the dynamic environment where there is no static pre- allocation of the memory.
Memory management techniques

- **VSMM consists of four levels**
  - Memory watchdog (alerts the host about the memory leakage and out of space problems)
  - Swap manager (Swap manager keeps record of all the resources using the memory)
  - Data exporter (Communication channel to export data) also maintain logs
  - Data Importer (Communication channel to import data) also maintain logs

- Other techniques like low level paging, generation of strong pointers and database optimization are also useful in the Cloud memory management.

- The memory efficiency can be achieved by choosing the best resource plan and state of the art technologies.
ESX by VMware

- In ESX, the memory management techniques allow the virtual machines to use more memory than the physical machines available memory.

  - For example, you can have a host with 2GB memory and run four virtual machines with 1GB memory each. In that case, the memory is overcommitted. To improve memory utilization, ESX transfers memory from idle virtual machine to virtual machines that need more memory.

Transparent Page Sharing (TPS)

- When multiple VMs are running, some of them may have identical sets of memory content (several VMs may be running the same OS, applications and the same user data). With page sharing the hypervisor can reclaim the redundant copies and only keep one copy, which is shared by multiple virtual machines in the host physical memory. Example 4shared, Github.

Ballooning

- A driver used to communicate with applications to understand their memory needs to make room for other applications.
- **RAMCloud**
  
  - The core idea in RAMCloud is to keep everything in DRAM with disks used only as backups.
  - Challenge is to make sure that the storage system can be recovered quickly upon failure.
  - In steady state, there is a single copy of the data present in DRAM. Recovery is performed using a massively parallel read of data from disks.

- Facebook used 150 TB dram for session management in 2009.

- **PACMan**
  
  - PACMan is a caching mechanism and corresponding system for HDFS and similar distributed file systems. The key idea is that current clusters have a large amount of unused memory that can be used to cache frequently-used data blocks, and traditional caching strategies like LRU or LFU do not work well on cluster jobs.
DRAMs as the future

-RAMCloud is a more general system than PACMan, but clearly, it is more expensive as well. RAMCloud trades off price for speed, but it is likely to be used in many future systems if prices of DRAM and high-speed network equipments keep going down. PACMan, from the high level, may seem to be a more short-term fix for the existing clusters. However, the insight of all-or-nothing is important and will be useful even in the future.
The use of efficient and dynamic techniques decreases the use of load balancers and increases the throughput in terms of memory management. To avoid bottleneck scenarios proper management policies must be adopted in terms of a cloud environment. The use of Hyper managers in case of services provided on cloud is necessary.
References:


URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.149.7163\&rep=rep1\&type=pdf


URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5233608