Designing a VM-level vertical scalability service in current cloud platforms: a new hope for wearable computers

Mustafa KAIIALI[1,*]

[1]Department of Computer Engineering, Mevlana University, Konya, Turkey

*Correspondence: mkaiiali@mevlana.edu.tr

Abstract: Public clouds are getting ripe for enterprise adoption. Many companies, including large enterprises, are increasingly relying on public clouds as a substitute for, or a supplement to, their own computing infrastructures. On the other hand, cloud storage service has attracted over 625 million users. However, apart from the storage service, other cloud services, such as the computing service, have not yet attracted the end users' interest for economical and technical reasons. Cloud service providers offers the horizontal scalability to make their services scalable and economical for enterprises while it is still not economical for the individual users to use their computing services due to the lack of vertical scalability. Moreover, current virtualization technologies and operating systems, specifically the guest operating systems installed on the virtual machines, do not support the concept of vertical scalability. In addition, the network remote access protocols are meant to administer remote machines but they are unable to run the non-administrative tasks such as playing heavy games and watching high quality videos remotely in a way that makes the users feel as if they are sitting locally on their personal machines. On the other hand, the industry is yet unable to bring efficient wearable computers to reality due to the limited size of the wearable devices where it is infeasible to place efficient processors and big enough hard disks. This paper aims to highlight the need for the vertical scalability service and design the appropriate cloud, virtualization layer and operating system services to incorporate vertical scalability in current cloud platforms in a way that will make it economically and

technically efficient for the end users to use cloud virtual machines as if they are using their personal laptops. Through these services, cloud takes wearable computing to the next stage and brings the wearable computers to reality.

Key Words: Cloud computing, wearable computing, virtualization technology, horizontal scalability, vertical scalability.

1. Introduction

Cloud computing is a model of computing through which services are commoditized and delivered in the same way as utilities such as water, electricity, gas, and telephony [1, 2]. Many companies have engaged in delivering cloud computing services to large enterprises as well as small ones. Providers such as Amazon, Google, Salesforce, IBM, Microsoft, and Sun Microsystems have established data centers for hosting cloud computing services in various locations around the world.

Enterprises currently aim to host their offered services on the cloud in order to improve their scalability to deal with rapid change in resource demands. Kingnet Technology is a company in Shanghai, China which develops games for worldwide social networks. With an estimated 30 million installations and 6 million daily active users, the company uses Amazon Elastic Compute Cloud (EC2) to get an infrastructure capable of handling such tremendous volume of requests (http://aws.amazon.com/solutions/case-studies/kingnet).

Small enterprises and startups can afford to translate their ideas into business results more quickly, without excessive upfront costs. Animoto is a company that creates videos out of images, music, and video fragments submitted by users. Animoto does not own any single server. It bases its computing infrastructure entirely on Amazon Web Services (AWS), which are sized on demand according to the overall workload [3].

Horizontal scalability is one of the key features of current cloud services that attracted enterprises to host their business services on the Cloud. It provides elasticity and an efficient Pay-Per-Use pricing model which help to avoid large capital expenditures and the “unable to serve customers” case that happens at peak times of the traditional computing systems as depicted in Figure 1.

Amazon provides this kind of scalability through the integration of three different services (Elastic Load Balancer, CloudWatch and AutoScaling) as depicted in Figure 2. Horizontal scalability allows scaling of the EC2 capacity up or down automatically according to the business load based on a predefined policy. Amazon CloudWatch lets you retrieve your monitoring data and set alarms to fire the AutoScaling service when needed. The Auto Scaling service ensures that the number of EC2 instances (sitting behind a load balancer) increases seamlessly during demand spikes to maintain performance, and decreases automatically during demand lulls to minimize costs. Elastic Load Balancing groups the running EC2 instances under one name and automatically distributes the incoming application traffic across them so they appear to the service consumers as a single coherent system.

However, despite its rising popularity, cloud computing has not yet reached its potentials. Its computing services are more helpful for enterprises or short-term usage and have not yet reached the individual users. This is due to economical and technical reasons [4]. The lack of virtual machine level vertical scalability in current cloud platforms creates an economical barrier against reaching the individual users. Users generally use their personal machines for heavy and light activities. Booking a powerful virtual machine (VM) which is capable of running their heavy kind of activities is going to increase their monthly/yearly bill even though users do not use their VMs always for such heavy activities. We need to adopt the VM-level vertical scalability by introducing the dynamic VM concept through which we can offer the services that allow auto scaling of the computing resources allocated for a single VM as per the user needs. So that, if the user is performing normal activities on his/her cloud VM (such as editing a word document) then the allocated RAM for this VM has to be shrunk to (let us say 1 GB) and he/she is to be charged for that much only. On the other hand, when the user is running heavy activities on his/her cloud VM (such as playing heavy games) the allocated RAM has to get expanded to (let us say 4 GB).

AWS Elastic Load Balancing does not provide VM-level kind of scalability. Its aim is to keep adding/releasing VM instances based on customers' load. This is a horizontal scalability and it does not help the end user to economically rely on cloud computing in any sense. This paper aims to create the appropriate cloud services and addresses the research challenges that current cloud platforms are going to face in order to deliver the computing service to the individual users in an economically efficient scheme.

The paper is organized as follows: Section 2 discusses related work being done in vertical scalability and highlights their limitations. Section 3 elaborates on Wearable Computing and its current limitations. Section 4 explores the economical and technical barriers that prevents end users from relying on public cloud services. Section 5 proposes a high level design for incorporating vertical scalability in cloud platforms. Section 6 concludes the paper.

2. Related work

2.1. Memory ballooning

VMware, one of the leaders in the virtualization technology, has previously introduced the "Memory Ballooning" service (http://searchservervirtualization.techtarget.com/definition/memory-ballooning). It is a virtual memory management technique used to prevent the hosting server from paging to disk based on the fact that memory allocated to a VM might not all be fully actively used. When the ESXi (the VMware virtual server) runs low on memory it uses the balloon driver (installed on each VM) to determine what memory the running VMs can give up to the host (as it is not actively used) to prevent it from lacking of free memory and paging to disk. However, this technique is meant for increasing the number of VMs that can be hosted by a single host not to offer full vertical scalability. To enable vertical scalability, we need to create the appropriate services on different levels: OS, virtualization layer and cloud platforms and for all kind of computing resources such as RAM and CPU. Current virtualization platforms do not allow dynamic allocation of computing resources for a single VM, neither current OSs support this concept because they were designed for physical machines where the processor and RAM are used to be static.

Apart from VMware Memory Ballooning, several other memory management techniques were used to dynamically adjust memory allocation. Weiming Zhao and Zhenlin Wang have introduced the MEmory Balancer (MEB) in [5] which they claim it to be more accurate than VMware Memory Ballooning and achieve higher performance. However, it is meant for the same purpose as that of VMware Memory Ballooning that is to increase the number of VMs which can be hosted by a single host not to offer full vertical scalability.

VMware Memory Ballooning, MEB and all other related techniques are virtualization layer level services. They allow virtual servers to serve more clients at a time by scaling the VM allocated memory down whenever it is not fully used and without the aware of the guest OS. This is a kind of stealing resources from hosted VMs for the benefit of the service providers not for the benefit of the clients. Moreover, these techniques do not allow VMs to scale up as the guest OS was designed to run on a static RAM/CPU. We need to implement services at guest OS level to allow VMs to scale up and down in a real fashion. Also we need to implement a cloud level service that adopts a scalable pricing model according to use statistics.

As an example, by adopting the VMware Memory Ballooning, the physical memory actually allocated to a particular VM (initially configured to hold 1 GB RAM) may vary in size (bellow 1 GB) according to use statistics. However, it will never go over (1 GB). On the other hand, the change in VMs memory size happens at the virtual server level in a way hidden from the VM itself. This is due to the lack of dynamic resource management services at the guest OS level because such an OS (like Linux, Windows, etc.) was originally meant for physical machines where resources are used to be static and not expected to be changed without rebooting. On the other hand, as the guest OS is not aware of memory ballooning service running at the level of the virtualization layer, this makes the guest OS memory management not optimal because OS is not aware of the actual memory size.

This paper ignites the need to design such dynamic resource allocation services on guest OS level opening the way for making special virtualization edition of each popular OS tuned to incorporate dynamic resource management services that allow the computing resources of a single VM to be dynamically (or manually) extended/shrunk based on end user's need. On the other hand, the cloud billing services have to be upgraded to consider the dynamic resource allocation of a single VM. Current cloud billing services does not.

2.2. VMware vSphere hot-add RAM and hot-plug CPU

This feature provided by VMware vSphere (http://searchvmware.techtarget.com/tip /VMware-vSphere-hot-add-RAM-and-hot-plug-CPU-Not-so-hot-but-still-cool) allows adding additional virtual hardware resources to a running virtual machine without bringing it down. Though this feature looks very close to the one that we would like to propose in this paper, yet there are many differences:

- This feature lacks to support by many guest operating systems. OSs that may support this feature are few server editions not suitable for end users. They support it for the sake of reducing the downtime of servers while enhancing their hardware. In this paper, we are concentrating on the end users' operating systems not the server ones.
- vSphere has implemented this feature at virtualization layer level. It is also supported on OS level by few server systems as discussed earlier. However, this feature lacks for support at cloud level. The cloud billing services have to be upgraded to consider the resource hot-plugging. Cloud level support is crucial for having VM-level vertical scalability implemented.
- This feature is manual. It means that administrators have to watch hardware resources consumption on guest VMs and manually they have to add additional resources if required. We need to have this process fully automated by implementing appropriate resource consumption watchers on guest OS level as will be discussed later in Section 5.
- To the best of our knowledge, VMware vSphere is the only virtualization platform that support this feature. This paper spurs open source virtualization platforms to support this feature as well.

2.3. Service level vertical scalability

dotCloud platform offers vertically scalability not on VM-level but on service level (SaaS). That is, if you host a particular service on their cloud, such as a web server service, it allows you to change, up and down, the quantity of memory allocated for that service at any moment, by using the dotCloud scale command (http://docs.dotcloud.com/guides/scaling):

```
#dotcloud scale www:memory=512M
```

They state in their web portal that: “The ability to scale the CPU and disk limits is being worked on!” (http://docs.dotcloud.com/guides/scaling). So their offered scalability is limited only to memory allocation and operates only on service level.

Joyent, a high performance cloud infrastructure company, on the other hand does offer a vertical scalability service. However, its vertical scalability seems to be limited to service level as well not to end user VM-level. They adopt vertical scalability through the introduction of their SmartOS. SmartOS is a hypervisor lean enough to run entirely in memory, powerful enough to run as much as you want to throw at it (http://smartos.org). Thus, services hosted on SmartOS can take the advantage of its dynamic resource allocation. We have approached Joyent support to make sure that our understanding to their offered dynamic resource allocation gained through reading their online documents is true; they replied: “A resize on a KVM (Linux/Windows) instance does require a reboot”. So their vertical scalability service is limited to services hosted on SmartOS not to end user VM-level.

3. Wearable computing

Mann describes wearables as constant and always ready, unrestrictive, not monopolizing of user attention, observable and controllable by the user, attentive to the environment, useful as a communication tool, and personal devices [6]. Many scientists have imagined how the future computers are going to be based on the concept of wearable computing [7]. As an example, the Pen Computer concept shown in Figure 3 (a) and the HOLO Computer concept shown in Figure 3 (b). However, these have remained as a conceptual view with no place in real time as it is still infeasible, technology-wise, to have efficient processors and high capacity hard disks in such a small size devices that can meet the computing resources requirement of a personal computer.

Moreover, even if it becomes possible in the future to accommodate the computing resources of current personal computers in such small devices, at that time current personal computer won't be even sufficient to meet the required computing resources of the future.

If cloud computing reaches the end users, it can take wearable computing to the next stage. When it is not possible to place big enough hard disk and fast enough processor along with its cooler in such small devices, cloud offers to host the computing and storage services in the cloud premises and keep those wearable devices to be used as a thin client only. This makes it possible for such conceptual devices to replace the personal laptops in the near future. Then we may find VM service providers in each country as common as the cell phone service providers. In fact, VMs can be offered as a new service of current cell phone service providers where the cell phone itself is used as a thin client to the VM.

The laser keyboard and the mini-projector concepts have already been manufactured and made available in markets but not in a combined project and not for the sake of wearable computers. The laser keyboard was manufactured as a portable mini keyboard (http://www.celluon.com/products_prodigy_overview.php) and the mini-projector was designed to easily share photos taken by mobile cameras with a set of friends sitting in the same room (http://www.aiptek.com.tw/c0_1.php?bid=18&pid=61). Also there are many thin client apps available as smartphone applications which incorporate RDP or VNC protocols for machines remote access (https://play.google.com/store/apps/details?id=com.softmedia.remote). Figure 4 shows an example of manufactured laser keyboard and mini-projector attached to an iPhone device. With the introduction of VM-level vertical scalability, a laser keyboard, a mini-projector and a smartphone can be used in a combined project to work as a thin client accessing scalable cloud VM bringing the concept of efficient wearable computer to reality.

4. Public cloud VMs vs personal computers - economical and technical challenges

In this section, we discuss the economical and technical barriers that prevent end users from relying on the cloud. Current public cloud pricing plans are yet not suitable for end users for regular and long-term usage. As an example, according to Amazon Simple Monthly Calculator, the cost of running a single VM of (2 x 2.6 GHz CPU, 7.5 GB RAM and 750 GB hard disk) for single month under the "On-Demand" billing option and for 35% usage over the time (1/3 running time over the month) costs 71.98$. This means the user can purchase a personal laptop of almost the same configuration with the cost of using his/her cloud VM for about 7 months only. After that he/she can keep using his/her laptop for free.

On Joyent, the cost of running a single VM of (2 x 1 GHz CPU, 7.5 GB RAM and 738 GB hard disk) for single month and for 35% usage over the time as well costs 0.240*24*30*35/100 = 60.48$. That is 84% of Amazon's rate, however, it is still considered expensive enough not to replace the personal laptop.

Horizontal scalability, discussed earlier, is crucial for enterprises that are willing to host their business services on the cloud. It provides them an excellent mechanism to manage their computing resources consumption efficiently with nearly optimal cost by dynamically adding or removing VM nodes to the AutoScaling group according to the business load. However, it does nothing for individual users who are willing to use a single VM through their wearable clients.

On the other hand, current Remote Access Protocols are meant to access machines remotely mostly for administration purposes not for personal use. As an example, RDP is not meant to play high quality videos on remote machines neither to run heavy games (http://etutorials.org/Microsoft+Products/microsoft+windows+server+2003+terminal+services/Chapter+3+Communication+Protocols+and+Thin+Clients/Remote+Desktop+Protocol+RDP). For the remote VM to replace the user's personal laptop, we have to tune the RDP parameters (http://www.donkz.nl/files/rdpsettings.html) to make it capable of doing such kind of activities remotely. Perhaps the following RDP parameters need to be tuned in order to make RDP effectively sufficient to run the non-administrative tasks:

- **Videoplaybackmode:** RDP efficient multimedia streaming has to be set up in order to play videos efficiently through the Remote Desktop Connection.
- **Session bpp (bits per pixel):** Determines the color depth on the remote computer when you connect. The higher value of session bpp the best quality of videos obtained over RDP session.
- **Audioqualitymode:** Determines the quality of the audio played in the remote session. Need to be adjusted to get the best audio quality with the minimal bandwidth.
- **Compression:** Choosing the efficient compression algorithm that keeps the required bandwidth and the computational cost to play heavy games or high quality videos at the lowest possible level.

Resolving these issues in current cloud platforms by adopting various services which incorporate VM-level vertical scalability and by tuning the RDP parameters to encounter the difficulties of running the non-administrative tasks helps to make cloud services more reachable to the individual users and affects the industry of wearable computing to an extreme extent. So that we can expect in the near future to have devices such as Amazon Cloud Pen or Cloud Glasses etc. which are used as thin clients to a personal cloud VM that replaces the personal laptop. These devices are to be purchased for one time and need not to be upgraded every couple of years like our personal laptops as the computing configuration is relevant to the cloud VM not to the thin client.

5. Incorporating VM-level vertical scalability in cloud platforms - high level design

Vertical scalability involves dynamic allocation of different computing resources, such as RAM and CPU, according to user needs. Current virtualization platforms do not support dynamic memory allocation at VM-level due to the lack of support at end user's guest OS installed on the virtual machines. Neither the cloud billing services are meant to support VM-level vertical scalability. It is only the virtualization layer which currently supports dynamic resource allocation for the sake of increasing the number of clients that can be served by a single virtual server. Also few server editions OSs, which are not meant for end users, support this feature partially as discussed in Section 2.2. Techniques used in these layers can still be adopted in the proposed solution, however, we still need to implement VM vertical scalability on cloud level and other end users' guest OSs.

Traditionally, once a VM boots up, the virtualization layer requests the hosting OS (the virtualization server) to reserve a memory block for the VM RAM. Consecutive blocks are allocated for other VMs in the same way as depicted in Figure 5. The limitation of this method arises if a VM wants to extend its memory at runtime, for an example consider VM_2, it is bounded with other VMs memory blocks, such as VM_1 and VM_3. Thus it cannot scale up unless it is powered off then powered back on with a new memory block allocated to it according to the new configuration.

To overcome this limitation, a virtualization layer level service called as the VM Dynamic RAM Manager (VMDRM) has to be designed and implemented to allow dynamic memory expansion at runtime. Figure 6 illustrates a working example. The DRMT stores VM IDs associated with their discontinuously allocated memory blocks. In this way, the RAM allocated to a VM can be expanded dynamically at runtime by allocating the next available memory block. The set of discontinuous memory blocks associated with the VM ID in the DRMT table forms the total VM RAM.

Some virtualization platforms have a sort of DRMT implemented for the sake of increasing the number of VMs that can be served by a single server by reclaiming the unused blocks of certain VMs and re-allotting them to another one as discussed earlier in Section 2. However, this change is not recognized by end user's VM itself for whom the memory is statically provisioned. For that, we have to implement other services on various levels. Figure 7 shows high level design of different services that need to be incorporated into different levels in order to achieve fully automated VM-level vertical scalability. Hereby we define these services and their functionalities:

- **VM-Watch agent** has to be implemented and installed in guest OS level to monitor resource consumption and fire the appropriate alarm according to a predefined policy.
- **vAutoScaling cloud service** is the master service that manages and provides the VM-level vertical scalability. It is the means by which users can define their scalability policies, such as their VM's maximum and minimum allowed RAM capacity, according to which the VM-Watch alarms are going to be handled. Figure 8 depicts a presumptive screen shot of the vAutoScaling service control panel for the scalability policy shown in Figure 9.

  User can specify the initial resource's units under which his/her VM is supposed to boot up. He/she can also define the minimum and maximum allowed units for a resource to shrink or expand according to the workload. Users can also view current resource consumption and manually acquire or release hardware resource units as shown in Figure 10. The cloud billing service has to be informed about every update being done in this service in order to charge the end user accordingly.

- **Dynamic VT services** are responsible for dynamic allocation/release of computing resources, such as RAM and CPU, at virtualization layer level. It is instructed by the cloud level service "vAutoScaling" that works according to the user's predefined scaling policy. Dynamic VT services adopts techniques such as the DRMT table to perform their job.

Once the resource consumption on a VM reaches a predefined threshold (e.g. 75% of RAM is utilized), VM-Watch agent fires the corresponding alarm (e.g. Overloaded_RAM). vAutoScaling service responses to the alarm according to the user's predefined policy (e.g. RAM can be extended up to 8 GB) then it instructs the virtualization layer services (e.g. VMDRM) to perform the needed actions (e.g. allocates new memory block and associates it with the VM ID in the DRMT table) and intimates the VM-Watch agent in the VM with the new resource configuration. The guest OS installed on the VM has to detect the changes of the allocated resources. For that, VM-Watch agent invokes the (RAM Probe and CPU Probe) system calls after each change in order to update the OS kernel with the new computing resources. Similar example can be given for CPU resources.

6. Conclusion and future scope

In this paper, a novel cloud service enhancement has been proposed by introducing VM-level vertical scalability service. Current cloud platforms offer horizontal kind of scalability that best suits for enterprises. With the VM-level vertical scalability, cloud services can reach individual end users and compete with personal computers industry. It can also take wearable computing industry to the next stage so that in the near future we may see tiny wearable devices that can access efficient cloud VMs very popular.

With this technological advancement, end users need not to worry about backing up their data to be protected from any kind of disk failure as this will be the job of the cloud service providers.

They also need not to worry about upgrading their personal devices' hardware as they can do so by just improving their cloud VMs configurations through the cloud portal. Additionally, they can have more efficient computing services with devices much more portable than their personal laptops as they can access cloud VMs equivalent to supercomputers whenever required through their tiny wearable devices. Work in progress to implement the proposed design in an open source OS (Debian Linux), virtualization layer (KVM) and cloud platform (Apache CloudStack).

References

[1] Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems 2009; 25:599-616.

[2] Jansen W, Grance T. Guidelines on security and privacy in public cloud computing. National Institute of Standards and Technology Special Publication 2011; SP800-144:0-80.

[3] Buyya R, Vecchiola C, Selvi ST. Mastering Cloud Computing: Foundations and Applications Programming. 1st ed. San Francisco, California, USA: Elsevier, 2013.

[4] Ali KH, Ian S, Ilango S. Research challenges for enterprise cloud computing. LSCITS Technical Report 2010.

[5] Zhao W, Zhenlin W. Dynamic memory balancing for virtual machines, In: ACM 2009 SIGPLAN/SIGOPS International Conference on Virtual Execution Environments; 11–13 March 2009; New York, NY, USA. pp. 21-30.

[6] Mann S. Introduction: On the bandwagon or beyond wearable computing?. Personal Technologies 1997; 1:203-207.

[7] Mann S. Wearable computing: A first step toward personal imaging. IEEE Computer 1997; 30: 25-32.

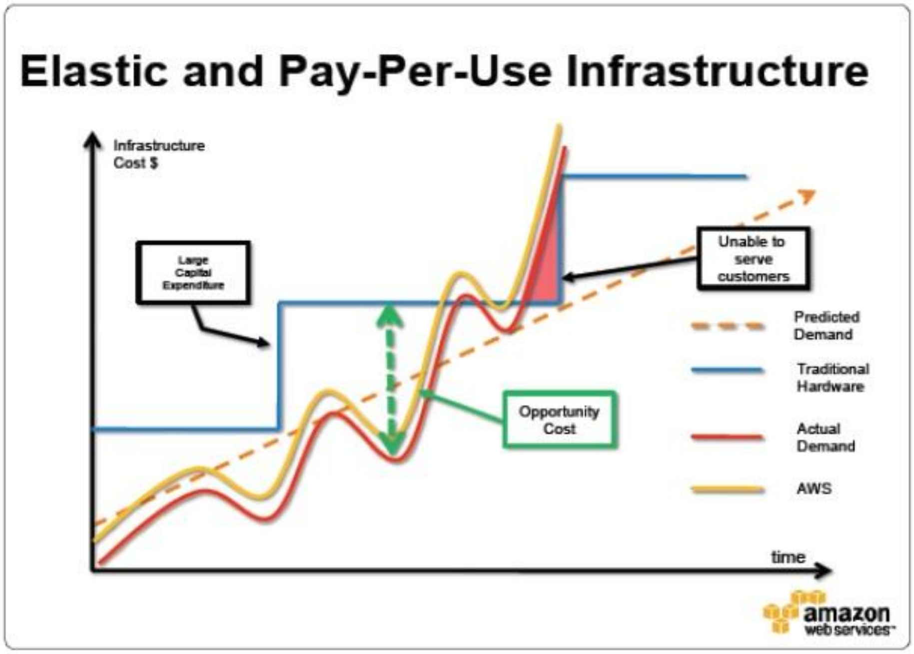

Figure 1. Elastic and Pay-Per-Use pricing model of AWS

(http://www.edureka.co/blog/why-aws).

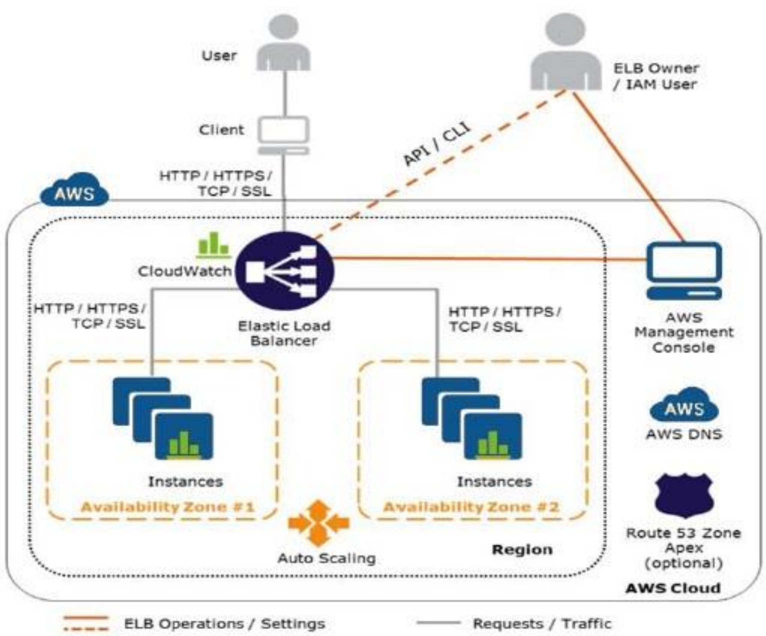

Figure 2. Architecture of the AWS Elastic Load Balancing service

(https://aws.amazon.com/articles/1636185810492479).

Figure 3. Wearable computer concepts.

(a) Pen Computer (http://www.hoax-slayer.com/pen-computer.shtml).

(b) Holo Computer (http://www.tuvie.com/holo-2-0-future-wearable-computer-for-2015).

Figure 4. Prodigy phone laser-keyboard and mini-projectors.

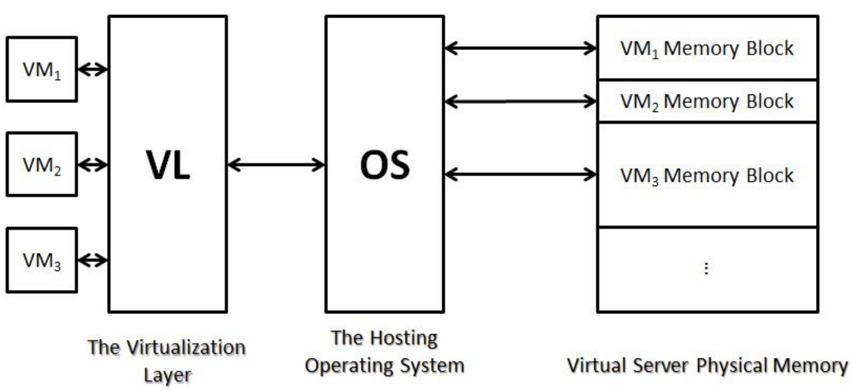

Figure 5. Traditional VM memory allocation.

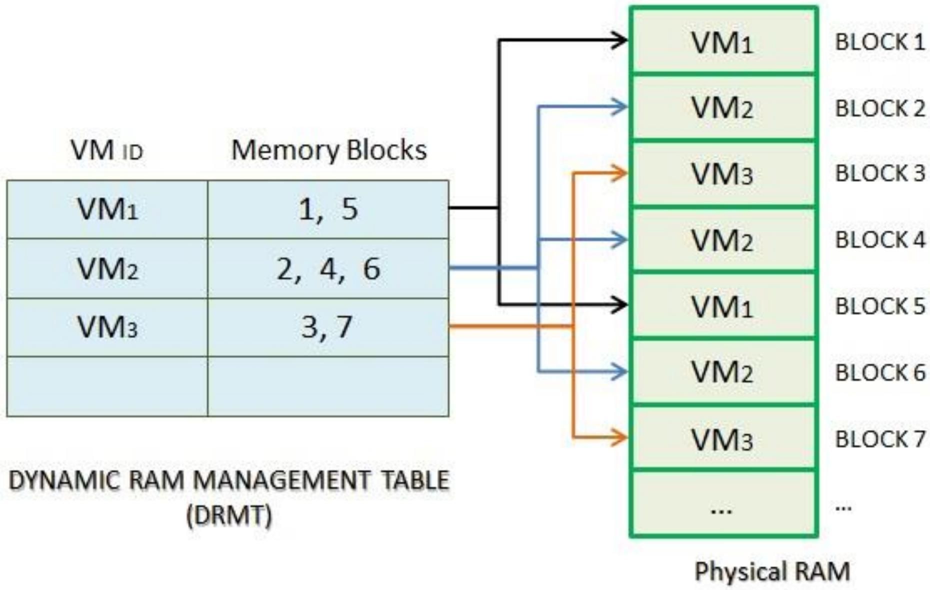

Figure 6. VM dynamic RAM management.

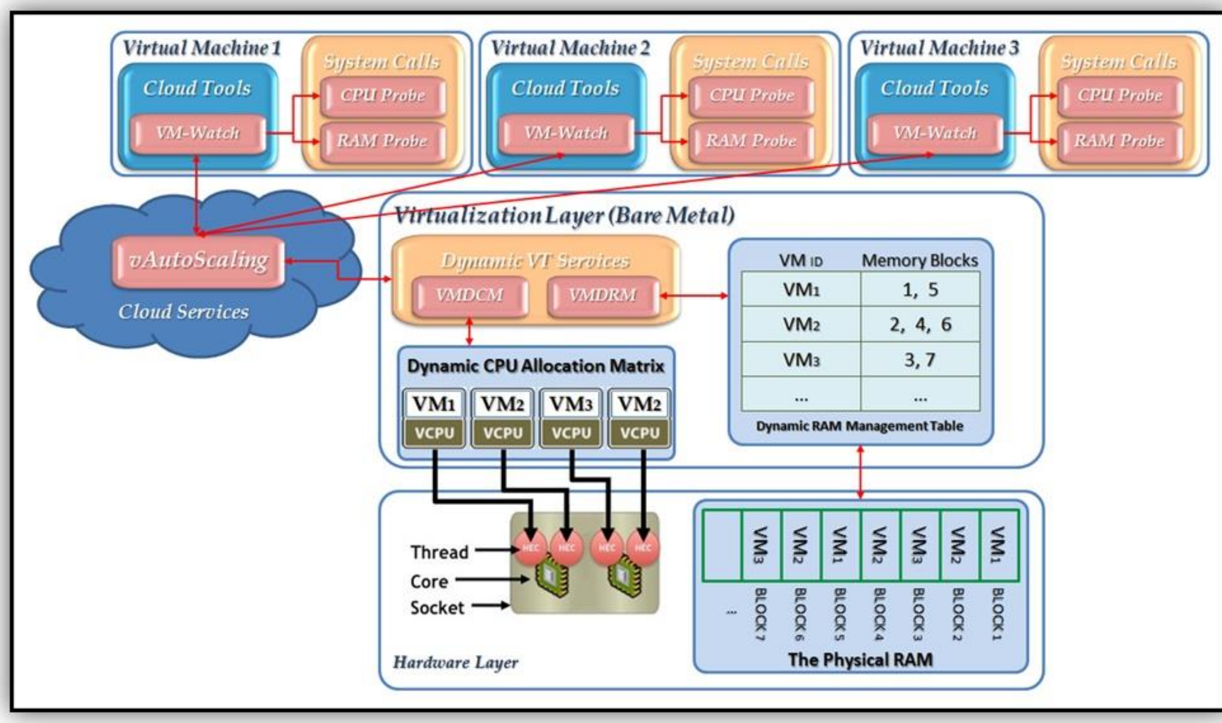

Figure 7. The proposed VM-level vertical scalability – services and architecture.

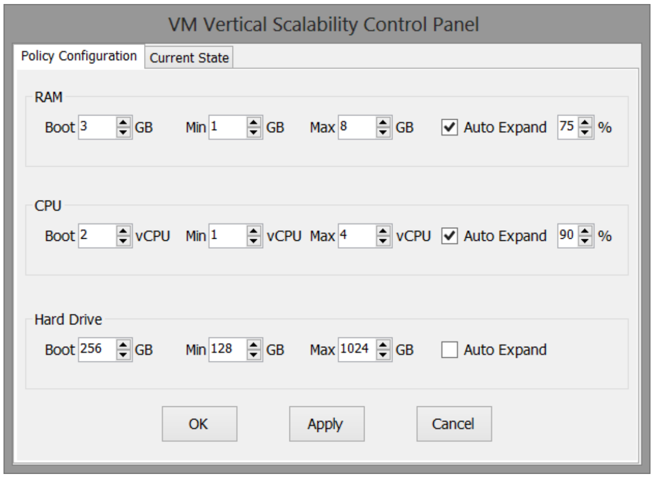

Figure 8. VM vertical scalability control panel.

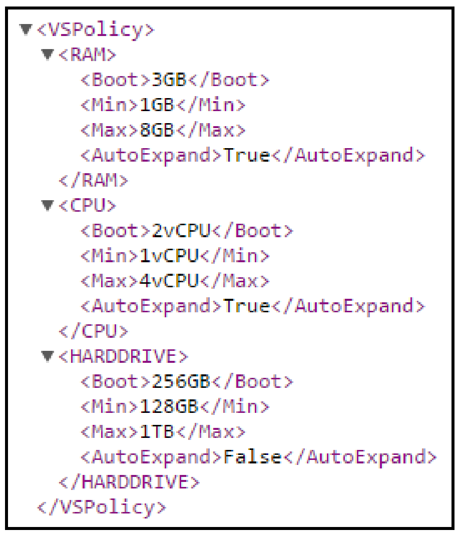

Figure 9. An example of VM vertical scalability policy.

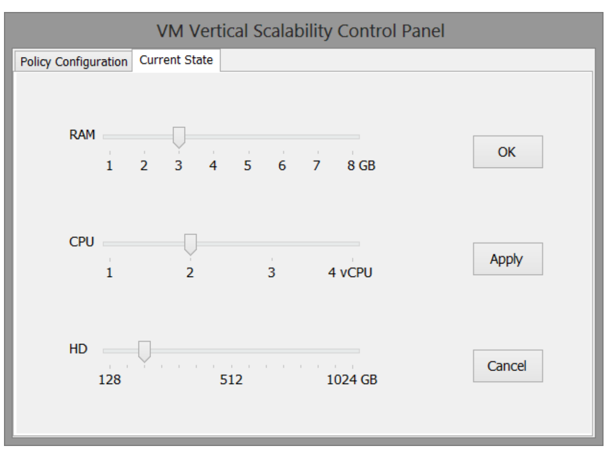

Figure 10. VM vertical scalability manual settings.