Encompass: Managing Functionality

Oleg Goldshmidt, Benny Rochwerger, Alex Glikson,
Inbar Shapira, and Tamar Domany

IBM Haifa Research Laboratory
University Campus, Haifa 31905, Israel
{olegg, rochwer, glikson, inbar_shapira, tamar}@il.ibm.com

Abstract

Today's system management tools focus on a computer as a visible enclosure of both computational resources (CPU and memory) and the functionality and data that reside in the storage subsystem. Recent technological trends, such as shared SAN or NAS storage and virtualization, have the potential to break this tight association between functionality and machines.

We describe the design and implementation of ENCOMPASS *— an image management system centered around a shared storage repository of "master system images", each representing different functionality. The functionality is provisioned by "cloning" master images, associating the resulting "clone images" with specified physical and/or virtual resources ("machines"), customizing the clone images for the specific environment and circumstances, and automatically performing the necessary operations to activate the clones. "Machines" — physical or virtual — are merely computational resources that do not have any permanent association with functionality.*

ENCOMPASS *supports the complete lifecycle of a system image, including reallocation and re-targeting of resources, maintenance, updates, etc. It separates image creation from image management from resource allocation policies — an emerging trend that is manifested in particular by proliferation of turn-key "virtual appliances".*

1-4244-0910-1/07/$20.00

1. Introduction

The functionality, the data, and indeed the identity and the "personality" of a computer system are contained in the storage subsystem, while the CPU and the memory are merely incidental resources that have no independent business function of their own. Moving the storage to a different (compatible) machine will provide functionally identical service, performance being the only differentiator.

Traditional system management views the (usually local) storage subsystem as an integral part of the computer. The concept of "functionality" or "service" is firmly associated with the computational resources embodied in a particular machine, and is never the object of management. Today's high level "orchestration" tools translate the policy-directed decisions straight into operations on physical entities such as "machines", "disks", etc.

Modern trends, in particular shared SAN or NAS storage [9] and virtualization [8, 7, 6], can potentially detach both the storage and the CPU and memory resources from a particular physical container. This is widely expected to lead to a revision of the traditional machine-centric system management approach. Initial attempts have proven to be unwieldy [2], but recent advances and proliferation of "virtual appliances" [5] offer new promises.

We propose below a new system management paradigm that focuses on the functionality encapsulated in bootable system images residing in shared SAN or NAS storage. The other components, such as CPUs and memory of physical or virtual servers, are explicitly treated as incidental resources.

The new mechanism is suitable for data centers and other large installations that provide a variety of different computing services that have to be provisioned and managed dynamically. We describe the design and implementation

of a prototype system, dubbed ENCOMPASS, some interesting usage scenarios, and directions for future work.

2. Design and Implementation

2.1. Image Repository

ENCOMPASS is centered around a repository of "master images" residing on NAS or SAN storage. Each master image representing certain business functionality such as a "web server", a "database server", etc. Master images are never used directly. Instead, they are "cloned" (cf. Section 2.3), creating multiple independent instances. "Clones" can be "customized" for a particular platform or environment (cf. Section 2.4) and deployed on physical or virtual machines. Master and clone images carry metadata describing the functionality, the relationships, and the associations between images and resources (cf. Section 2.6).

2.2. Creating Master Images

ENCOMPASS does not deal with image composition or installation, separating the image creation function from the provisioning, resource allocation, and other day-to-day management tasks. New master images, installed by any available means, can be "imported" into the Encompass system for subsequent cloning. A new master can also be "captured" from an existing clone, normally after an update or modification (cf. Section 2.5).

2.3. Creating Clone Images

ENCOMPASS images consist of either files on NAS or logical volumes on SAN. With NAS cloning is implemented as file copying, while with SAN storage the capabilities of the storage controllers ("volume copy") may be utilized. File copying may be made more efficient if the image files are sparse.

Copy-on-write (COW) techniques may be more efficient still. On SAN storage controllers the number of COW copies is often limited by implementation. NAS environments may provide unlimited COW.

One can attempt to separate a master image into read-only and read-write parts, and clone only the latter. A distributed filesystem may alleviate the concern that the read-only part will become a bottleneck. Feasibility of this approach can only be determined by implementation and measurements (cf. Section 5).

ENCOMPASS is capable of performing a large number of simultaneous cloning operations by farming them out to multiple nodes of a parallel filesystem or to multiple SAN controllers. There are numerous scalability issues that still remain to be resolved (cf. Section 5).

2.4. Clone Customization

A clone image may have to be "customized" for deployment. For scalability any customization operation must be performed in an unattended manner. It should be possible to operate a "zero configuration"[1] data center today, using technologies such as NIS, DHCP, DNS, SLP, LDAP, etc., but the possibility is often eschewed for various policy reasons. Besides, applications often require complicated customizations — a problem that can only be solved by vendors via creating well-behaved and appropriately packaged products.

In general, an image can be customized by modifying or replacing files in the image's filesystems. For many systems it is applicable directly, and when it is not the vendor normally provides configuration tools. By requirement, such a tool must work unattended, and therefore must be able to read its input from a file. By providing appropriate input files and invoking the tool (e.g., during boot), the problem is reduced to that of file replacement or modification. Microsoft's `sysprep` [1] is an example of such utility.

ENCOMPASS performs arbitrary file replacement and/or regular expression search and replace. An ordered list of "customization items" is a part of each master image metadata, while each clone image metadata include the corresponding customization data. A pointer into the customization item list tracks the changes that have been made, allowing to split the procedure into stages, roll back, etc.

2.5. Clone Image Lifecycle

The state diagram of a clone image is shown in Figure 1. A clone may be in one of 3 "metastates": "clean", "operational", or "maintenance".

A newly created clone is "clean" until the first boot, at which point it becomes "operational". An "operational" image may contain data and an internal state that are specific to the server's identity, and cannot return to the "clean" metastate.

The "assign" operation associates the image with a particular machine, and includes customization (cf. Section 2.4). The "attach" operation includes all the configurations of network, storage, fabric, machines, etc. "Attached" images can be "activated" (i.e., booted).

An operational image may be "reconfigured" as a part of the process to create a new master image. The transition moves the image to the "maintenance" metastate, in which the operator may perform various operations such as software installation and upgrades. The only way out of "maintenance" is capture of the image as a new master.

[1] This means "off-line configuration of infrastructure servers only".

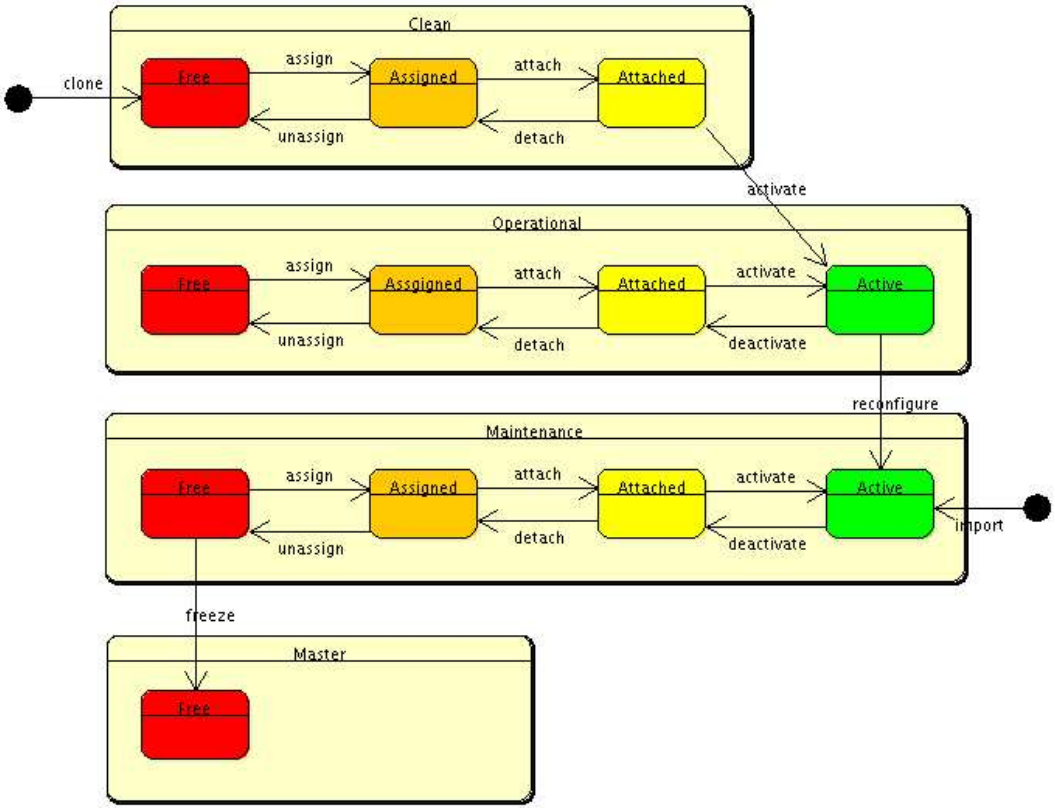

Figure 1. Clone image state diagram.

2.6. Bookkeeping

ENCOMPASS maintains metadata for images and resources (storage devices and volumes, physical and virtual machines). Master image metadata specify the service and software profile of the image, the machine requirements (architecture, minimal configuration, etc.), the list of existing clones, the required customizations. The clone image metadata point to the right master image, the machine to which the clone is assigned, the image state, the customization data, etc. The metadata are updated upon completion of each operation, and the management application can be stopped and restarted as needed.

2.7. Resource Discovery

ENCOMPASS is able to discover storage volumes and physical and virtual (VMware ESX or Xen) machines. Discovery requires that the physical machines be manageable remotely even when no software is running. Technologies that satisfy the above requirements include IBM BladeCenter [3] and Intel AMT [4].

In the current implementation, discovery is seeded with minimal information about the managed BladeCenters and storage devices. The discovery results are combined with the persistent image metadata that only exist inside ENCOMPASS, using heuristics to take into account that the physical world might have changed since the previous metadata update. More sophisticated discovery capabilities (e.g., via SLP) can be integrated without affecting the overall design.

Virtual machines offer additional capabilities. Physical computers equipped with wake-on-LAN may boot a hypervisor from flash or via PXE boot. The hypervisor will effectively form a part of the platform's firmware and will facilitate both remote management and discovery.

3. Usage Scenarios

Deployment of Functionality: To deploy a service the master is cloned (possibly more than once), and each clone is assigned to a physical or a virtual machine, customized, attached, and activated.

Migration of Functionality: Functionality may be moved from one set of resources to another by deactivating, detaching, and unassigning an image from a (physical or virtual) machine and deploying it on a different machine.

Reassignment of Resources: A physical or virtual ma-

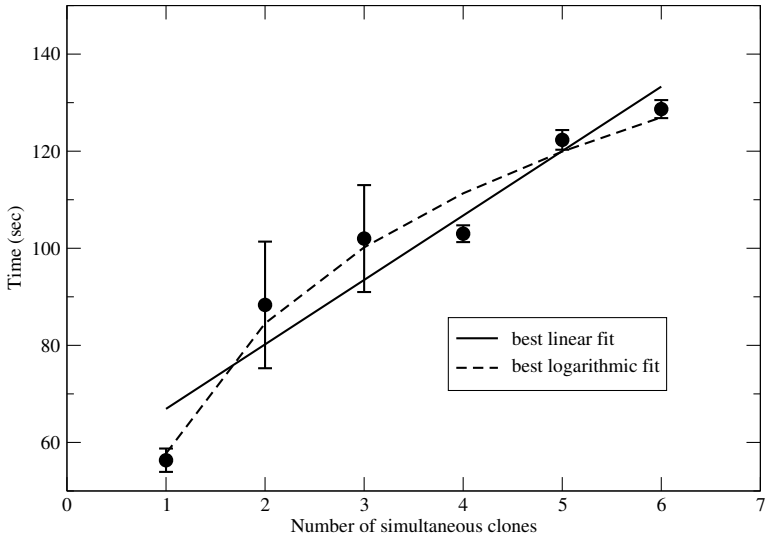

Figure 2. Parallel cloning performance

chine may be "reassigned" to support a different functionality. The active image is deactivated and detached from the machine, and another image is deployed. The original image may remain "on standby" or may be deployed elsewhere.

Service Update: A new clone is deployed and moved to the "maintenance" metastate. The operator applies the updates and tests the new image as needed. A new master is captured from the updated image. The resources used by the old clones are reassigned to clones of the new master: the bookkeeping engine makes tracking functionality and the associated resources easy. The old clones may remain "on standby" for rollback.

Export/Import: A snapshot of the metadata and a set of master images can be recorded ("exported") and the deployment can be easily replicated ("imported") at a different location. This is useful for creation of secondary sites or branch offices, as well as for replication of a client's setup for support purposes.

4. Cloning Performance

Image cloning is by far the most time-consuming operation related to service deployment. We timed parallel cloning of a master image in a number of storage environments. For lack of space, we only show the results of an experiment where a 3GB master image that resides in GPFS [9] is cloned by Xen hypervisors [6] running on IBM LS21 blades to their respective local `ext3` filesystems. Figure 2 presents the results, including the best linear ($t = 53.6 + 13n$) and logarithmic ($t = 57.7 + 38.6 \log n$) fits.

5. Future Work

Service deployment and migration in complicated communication and storage network topologies is an open issue. Target machines need connectivity to storage while the deployed images must belong to specified subnets, and satisfying all the constraints may be difficult.

Security issues have not been fully resolved, either. Keeping the storage network topology simple while granting images access capabilities to, e.g., specific storage volumes is a subject of future research.

More studies of scalability and performance of image cloning are needed. Massive concurrent cloning may be limited by read access to the master image. A scalable solution may involve "cascading" cloning, e.g., creating N clones of the master, then N clones of each "first generation" clone, etc. It will also be beneficial to create clones in storage locations topologically close to the target machines.

ENCOMPASS is suitable for embedding in, e.g., storage controllers, integrating image management with the basic storage functionality.

6. Conclusion

Modern technological trends, in particular shared networked storage and virtualization, sever the ties between business functionality represented by software and data and the computational resources. The trends warrant a reexamination of the traditional machine-centric approach to system management. We propose centering system management on business functionality, while treating physical and/or virtual machines as incidental resources. We have developed ENCOMPASS — a prototype functionality man-

agement system that supports the full lifecycle of a service and demonstrates viability of the new approach.

We are grateful to Yariv Aridor and Jose Moreira for support and many insightful discussions.

References

[1] How to use the Sysprep tool to automate successful deployment of Windows XP. http://support.microsoft.com/kb/302577.

[2] HP Utility Data Center. http://www.hp.com/hpinfo/newsroom/\press_kits/2003/enterprise/pdf/\7736_hpUDC_0505.pdf.

[3] IBM BladeCenter. http://www-03.ibm.com/systems/bladecenter/.

[4] Intel Active Management Technology. http://www.intel.com/technology/manage/\iamt/.

[5] Virtual Appliance Marketplace. http://vam.vmware.com.

[6] Xen Users' Manual. http://www.cl.cam.ac.uk/Research/SRG/\netos/xen/readmes/user/user.html.

[7] R. Oglesby and S. Herold. *VMware ESX Server: Advanced Technical Design Guide*. Brianmadden.Com Publishing Group, 2005.

[8] J. E. Smith and R. Nair. *Virtual Machines: Versatile Platforms for Systems and Processes*. Morgan Kaufmann, 2005.

[9] U. Troppens, R. Erkens, and W. Mueller. *Storage Networks Explained: Basics and Application of Fibre Channels SAN, NAS and InfiniBand*. Halsted Press, 2004.