F5 White Paper

Resource Provisioning—Hardware Virtualization, Your Way

Resource allocation can be a fine line, starving services if the adequate allocation isn't precisely managed. Resource provisioning solves that problem with both partitioned and dynamic resource allocation.

by Alan Murphy
Technical Marketing Manager

Contents

Introduction

One of the most interesting and useful types of virtualization in the data center is resource virtualization: carving up or dynamically allocating compute resources to services as requested. Resource virtualization (also known as hardware virtualization) can be thought of as sandboxing: creating an isolated container on shared hardware where nothing gets in or out of that container. Originating with mainframe environments and seen today in consumer laptops with multi-core CPUs, resource virtualization typically adheres to the one-to-many (1:M) virtualization classification—presenting one physical piece of hardware or cluster of resources as individually segmented portions.

There are two types of resource virtualization: partitioning and dynamic allocation.

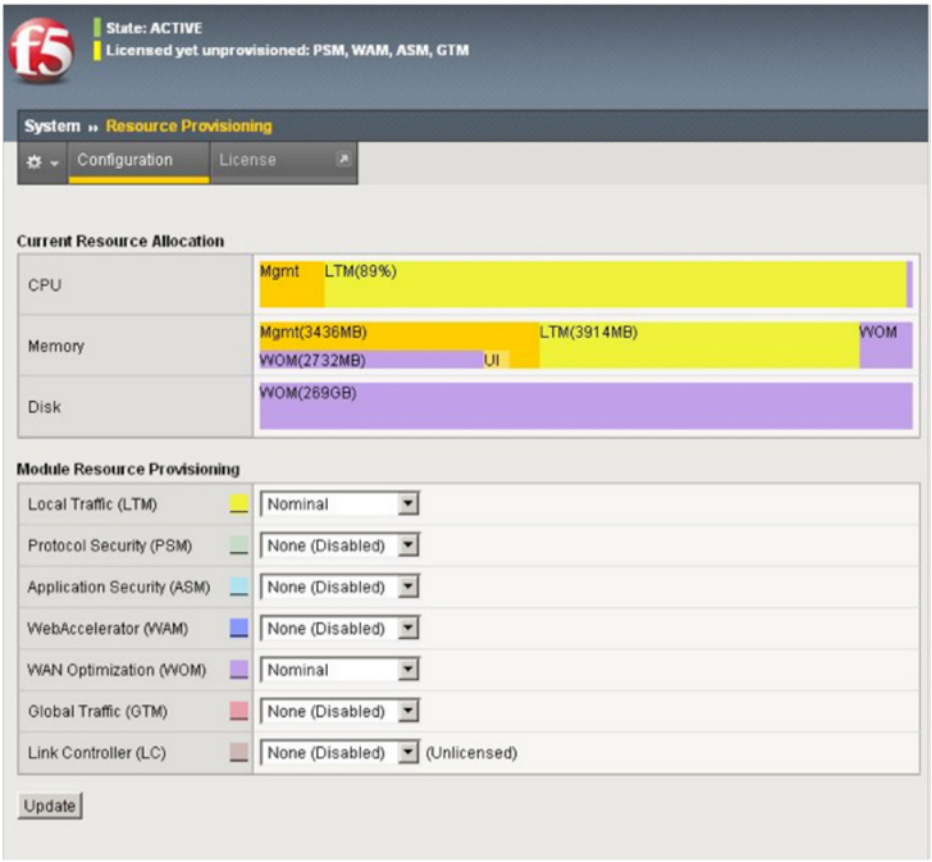
- Partitioning: Also referred to as slicing, resource virtualization partitioning pre-allocates a segment of the hardware for a set number of services. For example, a CPU may be divided into three equal parts between three difference services; each of those services can consume no more than 33 percent of the CPU. Partitioning is typically static, creating hard boundaries that do not scale as resource requirements change.
- Dynamic Allocation: The antonym to partitioning, dynamic provisioning is based on thresholds and the resource needs of a service at a particular time. Rather than carving out a fixed hardware chunk for a service, dynamic allocation will scale the hardware as needed for each service. Thresholds are set to limit the scale (if desired) and to moderate multiple services that require large portions of the hardware.

Each of these resource virtualization models has their benefits and their drawbacks, and neither choice typically accomplishes the goal alone. Only when they're used together can true resource provisioning be a benefit.

Resource Provisioning

F5® enables you to take advantage of both types of resource virtualization on F5 BIG-IP® devices with resource provisioning. Resource provisioning gives administrators the ability to manage hardware resources—CPU, memory, and disk—for each individual module. Resources can be allocated with a fixed percentage of each (partitioning) or a relative threshold (dynamic allocation) that can grow and shrink, within configured limits, as required by the module. By

default, resources grow with each module as needed. Once allocated, resources are dedicated to a particular module and cannot be taken away or used by other modules or services, unless the resource provisioning configuration is changed to reflect the new resource allocations.



An example of BIG-IP Local Traffic Manager's dynamic resource allocation configuration

A Modular Approach

The resource provisioning engine treats an installed set of modules as a whole system, adjusting the available resources based on what the installed modules require. Modules that are licensed and actively running on the system can be allocated resources; modules that are not enabled cannot be given resources. As modules are added and removed, enabled and disabled, the resource provisioning engine will reallocate available resources to accommodate the change.

For example, let's say an F5 BIG-IP® Local Traffic Manager™ (LTM) device is currently licensed for BIG-IP LTM and BIG-IP® Application Security Manager™ (ASM). The resources are allocated, based on percentages, between BIG-IP LTM and ASM. When the administrator tries to enable the BIG-IP® WAN Optimization Module™, the resource provisioning engine will determine what resources the BIG-IP WAN Optimization Module will need and if there are enough resources to allocate to it (based on the existing BIG-IP LTM and ASM requirements and resources available). If there are enough resources, the engine will allocate the configured amount to the BIG-IP WAN Optimization Module and the module will be enabled. If there are too few resources, the engine will not allocate anything to the BIG-IP WAN Optimization Module and will notify the administrator that there aren't enough available resources for the three modules to run in parallel. The resource determination is based on available physical resources (CPU and RAM) and the amount of minimum resources a module will require.

One of the most useful benefits of resource provisioning is the ability to dynamically add new modules—both fully licensed modules as well as evaluation modules. This enables the administrator to test new modules and their impact on available resources in a very granular, fully controlled manner. If BIG-IP LTM has an evaluation license for the BIG-IP ASM module, the resource manager will allocate resources to that module while the module is licensed (via the resource provisioning configuration and either with a fixed set of resources or a scalable amount). Then, it can remove those resources and place them back in the shared pool once the license expires. The resource provisioning and licensing engines are disassociated; modules can be added and removed independently of resource provisioning, enabling resources to be allocated and used by the team regardless of whether a module is licensed or not. Since all modules have a minimum resource requirement, resource management can be handled dynamically by BIG-IP devices, bringing new modules in and taking old modules out. The only limitations are the amount of physical resources available to the BIG-IP device and the amount of existing resources being consumed by each module.

Conclusion: Complete Control

Resource provisioning was designed to give administrators total control over the hardware usage of BIG-IP traffic management modules. One of the traditional complaints with fully partitioned systems is that the partitions can't scale. If a service absolutely needs more resources, there aren't more resources to give; or if a service is underutilizing its resources, it can't give them to another service. Homogeneous dynamic allocation isn't without fault, either. A dynamically allocated system with configurable limits or the ability to limit services can scale beyond the available resources, enabling a resource-demanding service to run unfettered and starving other services of their resources.

BIG-IP resource provisioning combines the best features of each virtualization model by enabling BIG-IP LTM modules to run sandboxed without sharing resources, while giving them the ability to dynamical scale if/when needed. Resource provisioning is a new model in hardware virtualization, one that will change the face of application delivery and Application Delivery Controllers moving forward.

F5 Networks, Inc. 401 Elliott Avenue West, Seattle, WA 98119 888-882-4447 www.f5.com

F5 Networks, Inc. Corporate Headquarters	F5 Networks Asia-Pacific	F5 Networks Ltd. Europe/Middle-East/Africa	F5 Networks Japan K.K.
info@.f5.com	info.asia@f5.com	emeainfo@f5.com	f5j-info@f5.com

WP-Resource-Provisioning 02/09