Assured Behavior for the Software-Defined-Infrastructure

For applications to perform as expected, a Service Level Agreement (SLA) is often established for the infrastructure. The SLA typically specifies required compute performance, I/O throughput and latency, amount of memory, security objectives, and other requirements. For the service provider or IT group, implementing this SLA can be challenging. For example, an application on a bare-metal server is not impacted by applications on other (bare-metal) servers. However, in a virtualized, multitenant environment, the applications are co-residing with other applications on shared infrastructure and the performance of applications meets the target service level objectives (SLO) that other applications do not interfere, is not a simple task.

Cloud service administrators (whether service providers or IT managers) require a means to specify the service level objectives of an application, and depend on hardware instrumentation adherence to the prescribed SLA.

Performance: To ensure that the application gets the allocated performance and suffers from less than adequate performance, even across CPU generations with varying cache and frequencies. This applies even if other co-resident applications are being “noisy neighbors,” where one or more applications negatively impact the performance of other nonrelated applications on the same server.

Trust Attestation: To ensure virtual machines are safe while running on shared, multitenant infrastructure, IT administrators need to assure that BIOS and the hypervisor have not been compromised (trust attestation). IT Administration wants applications placed and maintained on infrastructure that meets the specific trust attestation requirements for that application. For example, Peripheral Component Interconnect (PCI) applications are only placed on systems which are approved for PCI applications.

OpenStack* installations can be enhanced with OpenStack “flavors” (i.e., an OpenStack virtual hardware
Agility, Automation, and Efficiency

By using Intel® SAA, the service provider or IT manager can confidently create a multitenant environment supporting multiple business units with a wide variety of business applications. By assigning a specific SLO to a virtual machine instance or application—regardless of other activities in the environment—the application meets the expected performance criteria and operational parameters. Additionally, there is confidence that the infrastructure is running the designated hypervisor.

These capabilities allow more applications, along with legacy and cloud architecture to be hosted in a multitenant environment, while allowing older hardware to be phased out. Intel SAA provides the key capabilities necessary to host more applications in this shared environment. It brings the confidence that SLOs will be met, or if this is not possible for various reasons, such as over-subscription notification.

For the service provider, Intel SAA enables the ability to create new cloud service offerings potentially generating incremental revenue. For IT, cloud administrators can confidently use their approved cloud infrastructure.

**How it Works**

The Intel SAA software consists of three components: controller, agents, and plug-in.

- The controller periodically collects deep (low level) platform telemetry data from compute node agents to enhance virtual resource scheduling decisions.
- The plug-in for the OpenStack resource scheduler connects to a controller virtual appliance. This controller then routes machine instance provisioning requests for Intel SAA flavors to compute nodes that match the required trust attestation.

The SAA software also provides monitoring, remediation, and reporting capabilities to help assure that the cloud machine instances continue to run and meet the service level objectives.

In an OpenStack environment, a flavor defines the compute, memory, and storage capacity of a virtual server that users can launch. As an administrative user, you can create, edit, and delete flavors. An administrator can also enhance the flavor by using Intel SAA.

When a virtual machine is presented to an OpenStack user for instantiation, flavor is associated with the virtual machine (a normal process for OpenStack). That flavor will map to the virtual resource scheduler in the controller, which will then perform two actions:

- Determine the specified performance requirements, measured as Service Compute Units (SCU), based on the flavor and the required level of trust attestation.
- Determine the compute node suitable to host that application, based on the load level of the compute nodes, current levels of applications that are competing for resources (sometimes referred to as “contention”), and trust-attestation.

The SCU is a unique capability in Intel SAA. This simple metric, roughly equivalent to 2 billion instructions per second, defines compute performance that can be measured at runtime, without relying on coarse operating system or hypervisor utilization metrics. The measurements utilize deep platform telemetry (hardware in the CPU and chipset) to determine the level of performance available to that application, and are independent of the overall level of CPU utilization (high or low overall CPU utilization does not affect the measurements). This telemetry and the calculations that follow provide a uniform result regardless of CPU generation, cache size, or clock frequency.

For example, a standard benchmark such as Unixbench®, run in a VM managed by Intel SAA and with a target of 1 SCU, would show similar results across CPU generations, across varying cache sizes, and across varying clock frequencies.

One of the factors that can prevent an application from obtaining its allocated performance is the impact of other virtual machines running on the same socket (the noisy neighbor problem). The platform telemetry allows Intel SAA to compute, in real time, a contention metric for the compute node. Compute nodes with a high contention score are likely having noisy neighbor problems, such that additional applications should not be placed on those nodes. As new Intel® CPUs and chipsets are released, Intel SAA is expected to improve the ability to identify specific virtual machines that are causing the contention, thereby allowing more prescriptive actions to be taken.

Finally, the data gathered from the measurements is exported to a time series database where real-time analytics are run to maintain the contention scores and keep current on compute node utilization. The results of the data collection and analytics are accessible through a representational state transfer (REST) API.
Effective Operations
Cloud operations and cloud administrators often suffer from the problem of “blamestorming.”

Blamestorming occurs when application owner encounters performance issues and blames the infrastructure. However, the owners of the infrastructure often do not see issues and they place blame back on the application.

Is it the application? Is it the cloud software? Is it the infrastructure? Using deep platform telemetry and visual tools available in the administration console provided by Intel SAA, administrators are able to identify potential under-performance of SLO before they happen and take remedial action. The deep platform telemetry of Intel SAA acts to detect cache contention and memory channel bandwidth contentions that can negatively impact application performance. The end result: a responsive infrastructure delivering performance in an automated, efficient, and approved environment that can be monitored.

Easy Administration Console and API
Intel SAA presents an administration console (an extension of the OpenStack console), to allow inspection of compute node operation and to manage overall compute node configuration (for example, specifying SCU to flavor translations).

This console is an essential tool for the administrator to manage the infrastructure with the goal of meeting SLOs. The deep platform telemetry enables communication with the console through a published REST-based API.

One of the key elements to be managed is the health of the OpenStack infrastructure. Intel SAA maintains a powerful set of metrics to help monitor the performance and availability of the OpenStack cloud environment.
White Paper: Intel® Service Assurance Administrator

Intel SAA v1.1 supports OpenStack Grizzly* and Havana* releases and is available now.

Learn more at: intel.com/assurance