DEOPLI

Increasing Revenue Through Continuously Optimized Data Center Management

UNIVA

Executive Summary

Most businesses are completely dependent on technology, and the impact of using non-optimized technology can be directly measured in their revenue. Why? Because as widespread dependence on technology has grown, the size and complexity of the technology's environment has grown in a non-linear manner. This non-linear growth creates hindrances in operation that become costly, as well as resource consuming. The best solution to this growing discrepancy between our business's technology dependence and its operational environment is to program infrastructure policy.

Just as it no longer makes sense for us to write programs in machine code, optimizing a technology's environment is not something that can be accomplished through manual manipulation. Fortunately, a high level of optimization can be accomplished by defining the policy that automates the process - continuously reassessing the total workload, environment resources, and policy criteria for optimal configuration and workload placement. This optimization and reassessment are accomplished through solving numerous, very complex matrices of multi-variable equations. Alone, humans are very poor at solving these types of equations, but with the assistance of computers, they can be solved at speeds faster than the mind can comprehend. Thus, enabling programmed policy to execute the appropriate events in the right sequence, at the right time, and only when necessary, is the best method for optimizing our business technology.

Today's Typical Data Center

Data centers grow larger and larger based on ever-increasing business demands, with technology continually introduced to help manage their increased size and complexity. In the past, new data center requirements were remedied with the help of system monitoring solutions, provisioning solutions, asset management solutions and supplier provided reporting data on power and WAN consumption. These solutions, however, only served as partial fixes to larger, emerging problems. And perhaps worse, they were individually designed to address independent issues, and not intended to work in collaboration.

The efficiency of a data center translates directly to the work that the infrastructure is performing. When separate solutions are forced to coexist, the inefficiencies compound, preventing maximum value from being delivered. The value is intrinsically linked to the core product of the business - and measured in revenue.

Moreover, as data center capacity and complexity increase, the number of personnel needed to manage this emerging environment is no longer a linear progression of how it was done in the past. Using historical methods and tools, the number of personnel required to maintain control of the increasing environment only grows geometrically larger.

Note: This geometric increase is due to the environment growing linearly larger, as well as linearly more complex, causing both variables in the equation to grow.

Targeting Inefficiency in Infrastructure

It is vital for companies using technical infrastructure as part of their product development cycle to get the absolute most out of that capability. Falling short of this goal will have a direct impact on revenue, often at multiple times the cost of the infrastructure investment. Inefficient infrastructure on a small scale is understandable, if the effort taken to address the inefficiency risks being more expensive than the gained value. However, as data center capacity scales up to address demand, the inefficiency can become multiplicative and significant. At this point, the growing inefficiency must be targeted and dealt with.

To maximize use of the entire data center, our equation for optimization must be made up of all the variables and equations that comprise the resource as a whole - not just a bunch of smaller equations solved individually for the sake of simplicity. To do this, we gather all data related to resources associated to the data center, system and infrastructure. Next, we layer on top a framework that can house and maintain the data, but also be capable of detecting and tracking any changes performed once the initial data has been gathered. Additionally, the framework must create and track relational associations between resources in order to understand all existing provisioning potential. Using this method, the process and framework will always be customized to each and every deployment, based on the data fed into the solution.

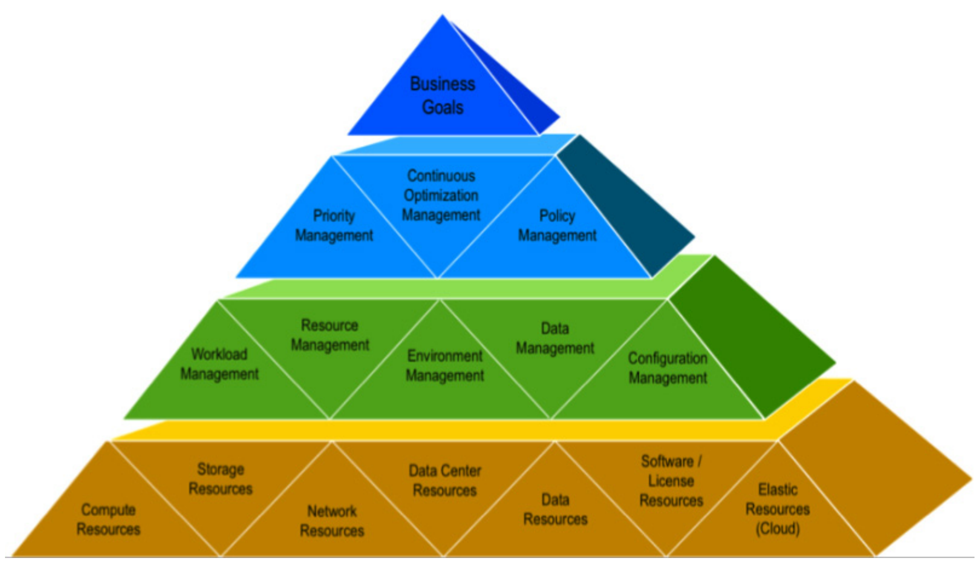

Figure 1—Data Center Optimization Solution Structure

Identifying an Appropriate Solution

Implementation of a data center optimization solution must be carried out systematically. The following steps are an overview of the process:

- Start by defining the business objectives that will translate into policy criteria.
- Correlate business objectives to infrastructure resources to define policy.
- Capture and store all aspects of existing operational resources, tracking them over time.
- Measure the resources for utilization, capacity, performance, and availability for use in provisioning. Record those measurements for historical analysis.
- Define the policy based on business objectives, available resources, and tactical capabilities, and have the automation execute the proper steps to accomplish configuration.
- Tune when, and how, environment configuration happens based on policy.
- Correlate workload-to-resource capability to estimate effectiveness and efficiency.

Compiling Operational Resources

The solution should follow the solution structure in **Figure 1** by collecting material assets and separating them into the following categories:

- Compute Resources - Computational nodes, including physical systems (e.g., processors, memory, local attached storage, add-in cards, network connectivity, special capabilities).
- Storage Resources - Storage subsystems, including everything that can be dynamic (i.e., not locally attached). Accounts for storage head node (server) capacities, performance, and dynamics.
- Network Resources - Local switch, network topology, switch resources (e.g., processing, memory, buffers, routing).
- Data Center Resources - Racks, network cables, power cables, power circuits, PDU, UPS, generators, chillers, cooling towers, air handlers, outside air economizers, filtrations systems, temperature sensors, geographic location (e.g., row, column, latitude, longitude), special location (e.g., rack unit), data center geographic location (e.g., city, street, building, floor).
- Data Resources - Data sets (i.e., associations), meta-data (e.g., asset inventory, name service data, accounting data, authorization data, access controls).
- Software / Licenses Resources - All available software packages, license keys and constraints.
- Elastic Resources (i.e., Cloud) - All available cloud resources, costs and access metric.

Identifying Management Tactics

The solution should then aggregate material resources into management tactics to accomplish higher level functions:

- Workload Management - Process workload requests and combine with resource management information to arrive at the best configuration of resources to execute the workload, then place the workload optimally - accounting for total workload and all resources.
- Resource Management - Data collection and processing on compute resources, storage resources, network resources and software/licenses resources with regard to capacity, performance, and capability. Track and leverage data regarding provisioning capabilities, configurations, provisioning options, management of virtual resources and configuration parameters of all resources.
- Environment Management - Configure and set dynamics related to the physical environment where workload is being executed. Take into account the location of sensors, multi-input criteria for temperature measurement, power measurement, system location, system configuration and capability, job resource utilization, data center resource utilization and capability.

  Note: Integrating this knowledge allows workload to automatically be redistributed, migrated, or quiesced based on an environment failure or loading issue.
- Data Management - Aggregate information related to data-sets such as storage location, use dynamics and access protocol.
- Configuration Management - Aggregate data related to configurations for systems, virtual machine images, network devices, storage devices, data entities, applications and application instances.

  Note: Track any changes to those configurations for backing out changes without re-creation.

Orchestrating a Strategy

The next step toward an all-encompassing solution is to orchestrate tactical functions to accomplish high level functions:

- Priority Management - Priorities should be determined based on revenue impact, or an ability to impact strategic business objectives. Define how priorities are handled and negotiated - there must be an authoritative decision tree that accommodates prioritization when conflicting requirements are encountered.
- Policy Management - Policies are derived from business goals, and are measured as a function of revenue. Create a set of rules that determine behavior based on demand and resource availability. Establish policy based on the expected interaction of systems, users, and workload. Policy must accommodate all workloads and resource utilization through a policy engine that verifies resource utilization requirements and can make priority decisions on that usage, if required.
- Continuous Optimization - Target revenue goals as compared to system and resource goals. Establish resource allocation decisions across the entire infrastructure on a continuous basis. While resetting priority on work is something that will always happen, the ability to make those decisions without throwing away the work that has been executed on workloads that are no longer on the proper resource is vital. Perhaps a new job is required, or a larger resource is needed. The continuous optimization part of the solution must be able to adjust to such requirements.

The Univa Solution

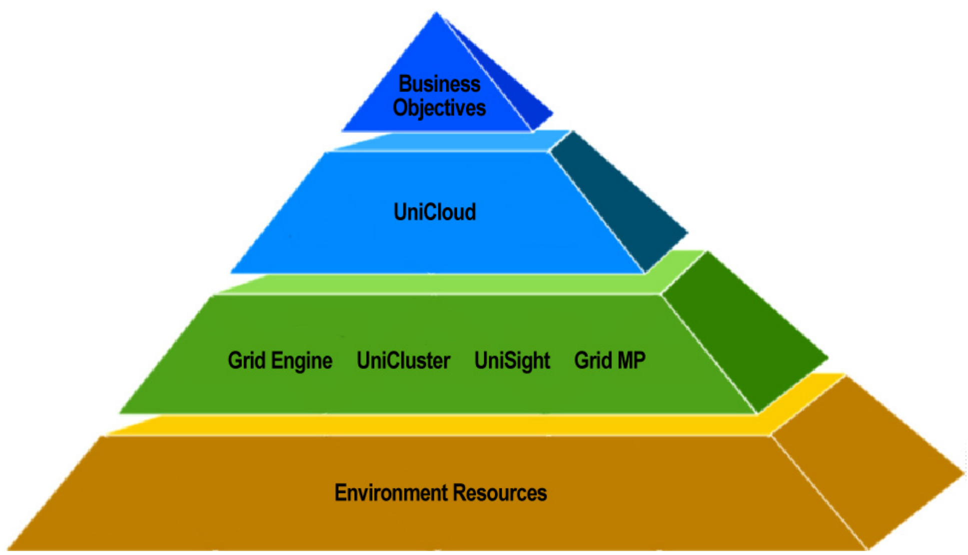

Figure 2—**Univa Product Structure**

Executing the business plan above is a daunting process. Fortunately, Univa has created a suite of software that accomplishes data center optimization with ease, while also intentionally designed to work in collaboration. Univa continually optimizes all components of a modern data center by integrating the following functions:

- Policy - ensures smart optimization by mapping business objectives to infrastructure resources
- Provisioning - optimizes infrastructure delivery
- Resource management - supports optimization with metrics and data
- Virtualization & cloud management - optimizes resource utilization including extending into dynamic capacity
- Workload scheduling - optimizes workload placement

Univa's modular approach allows layering various management software in to the infrastructure to improve specific utilization or efficiency gaps. The software dynamically creates and configures the managed elements within the computing environment. The software then initiates action as necessary to ensure optimal policy alignment - from creating physical or virtual servers, and changing their configuration, to managing capacity on storage or networking resources or securely moving workload to, or from, a cloud.

Grid Engine

Grid Engine combines the functionality of a resource manager and a workload manager, making it easy to create clusters of thousands of machines while managing workloads across the cluster. This tool delivers multiple scheduling policies for matching workload in the cluster to business and organizational objectives. These objectives include maximizing utilization across all machines, reducing turnaround time for jobs in the cluster and prioritizing workload according to group, department or company affiliation.

Grid Engine continuously collects metrics from all cluster nodes, then uses scheduling strategies configured by the administrator to evaluate active and pending workload. The tool then matches specific job requirements with appropriate resources, both utilized and available, to reallocate workload to accommodate proper placement. With Grid Engine, any resource including software licenses can be monitored, tracked and scheduled to ensure applications are automatically matched to the appropriate licenses and machines in the cluster. All finite resources are effectively managed through constraint, allowing resource quota sets to further refine the level of control for sharing resources among users, groups and departments.

Grid Engine supports a variety of workload types so that jobs can be scheduled and run efficiently in the cluster. Whether sequential, parametric, or parallel, Grid Engine houses a profile that can accommodate the appropriately matching job type for any collection of physical resources. Additionally, interactive jobs can be submitted to Grid Engine, providing a remote shell to the user while simultaneously enforcing scheduling and policy control.

Grid MP

Grid MP is the leading product for building distributed computing environments from non-dedicated resources in order to deliver increased high-performance computing (HPC) power. Grid MP is proven to increase productivity, efficiency, innovation, and performance - from small-scale desktop scavenging deployments to global enterprise-class and even Internet-based grids.

Grid MP enables Hybrid Cloud augmentation of your existing environment. The tool is fully integrated with UniCloud, providing functionality that seamlessly extends a Grid Engine cluster to an external cloud computing service such as Amazon EC2 or Rackspace Cloud. It's fast and easy to create a Grid Engine cluster that runs in the public cloud, or to establish a secure hybrid cloud that contains machines from your environment and a public, or private, cloud provider. With production installations at hundreds of Global 2000 customers, Grid MP reduces operating costs and improves productivity by harnessing unused compute cycles from local or remote resources. No other commercial technology is engineered to operate across thousands of disparate, globally-dispersed devices.

UniCluster

UniCluster is an integrated cluster management software stack that consolidates best-of-breed open source technologies to provide everything needed to run applications on a cluster. The tool is the open-source alternative to traditional proprietary/commercial cluster software. Delivering far more capabilities than a scheduler alone, UniCluster goes beyond ordinary cluster management tools to provide integrated installation of multiple components, monitoring and analytical capabilities.

UniSight

UniSight integrates directly with Grid Engine and Grid MP software, capturing and consolidating usage data for applications, jobs, resources and users to precisely measure utilization across all resources. The tool creates comprehensive views of job and resource information with ease, presenting trends across multiple clusters and schedulers from a single instance. UniSight combines robust data collection and scalability with a simplified interface and the practical features system administrators require. UniSight supports Grid Engine and Grid MP schedulers out of the box, and can be extended to support other schedulers, as necessary.

UniSight provides unparalleled advantages in its ability to report historically and remotely. While many commercial reporting products leverage only instantaneous data, UniSight can generate summary reports for any resource - both instantaneous and historical. Without the ability to incorporate historical data, a reporting tool can only reflect the current state, and not previously occurring events. UniSight's ability to reflect history allows trending analysis that can help mitigate problems before they occur. In addition, UniSight collects reporting data remotely from scheduling servers using standard JDBC, offering a distinct advantage over tools that require local data collection agents - where the act of monitoring impacts the workload, or system being monitored.

UniCloud

UniCloud optimizes resource utilization by ensuring applications are matched to their optimal data center resources. Using the tool's built-in policy engine, decisions are made regarding where applications should run in the infrastructure and what machines and memory combinations are needed to support optimal performance. If those machines (physical or virtual) do not exist, or are not available, UniCloud can create and size the resource to meet the needs of the application, ensuring memory and cores are not wasted. Once systems are made available and the infrastructure has been shaped to meet the needs of the application, UniCloud then automates the configuration of those new machines so that all software is installed and ready to deploy - including the application. Finally, UniCloud dynamically optimizes this new infrastructure to automatically respond to user and project demands by continually repeating the process described above.

Conclusion

Univa's software suite addresses all of the optimization issues found in traditional, dynamic and cloud data centers by harnessing automation, policy control and resource management in one package. They have analyzed the appropriate business model for creating an all-encompassing solution, and in doing so, have succeeded in creating a suite of software that is robust and configurable - but also client-friendly. Their software is industry proven to dramatically increase efficiency, effecting technology-dependent business revenue in the best way possible.

About Univa:
Univa, the Data Center Optimization Company, is the leading provider of optimization and management software for traditional, dynamic and cloud data centers. Their award-winning products are used by Global 2500 companies to improve resource sharing, amplify the efficiency of people and processes, and increase application and license utilization. Univa offers the industry's broadest, most innovative and integrated product set for managing shared, high-demand data center resources. From workload management to policy-driven provisioning across physical, virtual and cloud resources, only Univa provides a proven combination of enterprise-class capabilities, industry expertise, and community sponsorship. Univa is headquartered in Lisle, Illinois with offices worldwide. For more information, visit **www.univa.com**.

About Deopli:
Deopli is one of the foremost thought leaders in the EDA infrastructure and cloud computing space. Composed of highly-trained personnel, equipped with technology and experience, operating under principles of self-sufficiency, technical competence, speed, efficiency and close teamwork. Providing advisory and consulting services to EDA companies with respect to their HPC environments, they also conduct specialized operations including reconnaissance, strategy definition, tactical definition and resource training. In addition, Deopli executes non-operational, high-risk tasks to achieve significant strategic objectives. Deopli is headquartered in Irvine, California. For more information, visit **www.deopli.com**.