

# Optimizing the cloud

Jay Kulkarni  
Sanjay Agara

Wipro Technologies

## Introduction

Cloud computing is on-demand, integrated, configured, ready to use combination of compute, storage, network, platform and application software available as a standardized set of service offerings on a pay-as-you-use pricing model.

Cloud computing is altering the IT service delivery by providing rapid service delivery – provisioning resources takes few minutes instead several weeks, on demand scaling – businesses can scale up or scale down resources based on needs instead of provisioning for peak demand, consolidation of resources by virtualization is leading to better asset utilization and reducing resourcing needs, self service provisioning is leading to reduced time for service acquisition, standardization of service offerings is leading to reduction in the cost of maintenance, flexible pay-as-you-use charging supported by metering and billing is resulting in reduced costs as businesses pay for what they consume.

With clear benefits of cloud computing, many legacy applications are being migrated to the

cloud. This paper explores the cloud migration activity and presents six levers for resource optimization during the cloud migration process.

## Cloud Migration

Cloud migration encompasses moving one or more enterprise applications and their IT environments from a “traditional hosting” type of environment into a cloud either private or public or hybrid.

Cloud migration activity is carried out in these phases:

- (1) Evaluation: Evaluation is carried out for current infrastructure and application architecture, environment in terms of compute, storage, monitoring and management, SLAs, operational processes, financial considerations, risk, security, compliance and licensing needs are identified to build a business case for moving to the cloud.
- (2) Migration strategy: Based on the evaluation, a migration strategy is drawn – a hotplug strategy

is used where the applications and their data and interface dependencies are isolated and these applications can be operationalized all at once. A fusion strategy is used where the applications can be partially migrated, but for a portion of it there are dependencies based on existing licenses, specialized server requirements like mainframes or extensive interconnections with other applications.

(3) Prototyping: Migration activity is preceded by a prototyping activity to validate and ensure small portion of the applications are tested on the cloud environment with test data setup.

(4) Provisioning: Pre-migration optimizations identified are implemented. Cloud servers are provisioned for all the identified environments, necessary platform softwares and applications are deployed, configurations are tuned to match the new environment sizing, databases and files are replicated. All internal and external integration points are properly configured. Web services, batch jobs, operation and management software are setup in the new environments.

(5) Testing: Post migration tests are conducted to ensure migration has been successful. Performance and load testing, failure and recovery testing and scale out testing are conducted against the expected traffic load and resource utilization levels.

## Optimization

Before the actual migration activity, several resource optimization activities are carried out.

There are six levers which can be used for resource optimization:

- Optimize by consolidating under-utilized resources.
- Optimize based on time of day usage.
- Optimize based on seasonal demand variations.
- Optimize based on life cycle usage variations.
- Optimize by standardizing platforms.
- Optimize by application rationalization.

Following sections detail these optimization techniques:

### **Optimize by consolidating under-utilized resources:**

In an organization with decentralized IT operations. Each brand maintains their own independent IT environments where resources are allocated to applications based on the needs of that brand. As the capacity planning is undertaken independently for each brand, resources are left with spare capacity that goes unutilized. When considered as a whole, the net spare capacity across all brands becomes significant. This provides an opportunity to consolidate applications across brands and reduce the resources based on their utilization.

To illustrate the opportunity for spare capacity utilization, resource utilization chart of applications in the production environments of two different brands under an example organization are considered here:

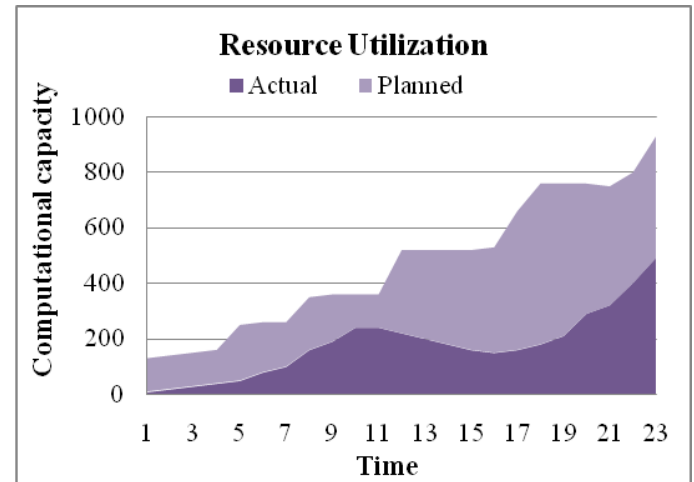
Applications	Brand 1 Instances	CPU Utilization	Brand 2 Instances	CPU Utilization
File & print	3	20%	10	40%
EDW	12	55%	2	75%
Blackberry	2	20%	4	45%
ERP	4	50%	4	65%
WMS	4	35%	2	40%
POS	4	40%	4	40%
Oracle	4	40%	8	55%

In the production environments across the brands there are common applications for POS, BI with Enterprise data warehouse, WMS, Collaboration and ERP. When considered together, the average utilization of WMS application instances is 36%. Rest of the capacity is left unutilized.

Optimization is achieved by creating a shared production environment with the common applications across brands. Shared resource pool for the applications reduces the number of resources allocated and improves the capacity utilization. Any burst in traffic can be managed by using auto-scaling to automatically increase or decrease the number of instances.

### Resource optimization based on demand:

In a traditional data center computational capacity is over-provisioned considering peak loads and thus under-utilized.



Traffic flow analysis is conducted to determine the resource demand variations for the applications that use the most bandwidth, time of day usage, average time users are connected to the service, peak season and off season usage, CPU, memory and storage resource utilization.

There are three types of demand variations that can be considered for optimization: Demand based on time of day, on season and lifecycle of usage.

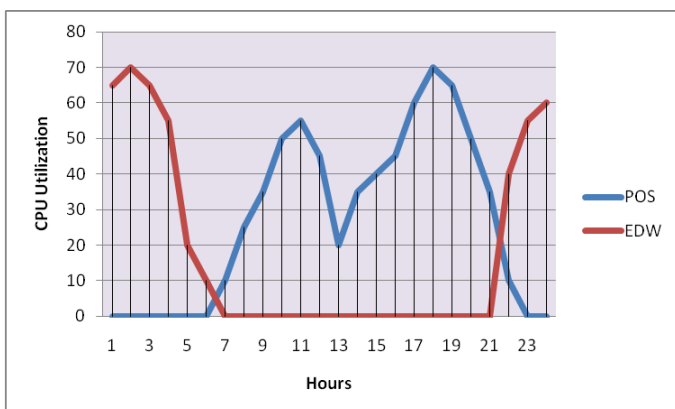
### Time of day demand variations

Certain applications demonstrate periodic spurt in traffic aligned with time-of-day variations based on factors which are unique to the organization or industry. Some examples are morning spike in load on attendance application due to employee

attendance swipe, end of day sales reconciliation, catalog item description updates during the night in an ecommerce site, search indexing every 8 hours on an intranet search application.

Traffic profiling of all applications is performed during the technical evaluations phase, provides the needed information to choose applications for optimization. Applications are chosen such that peak traffic loads are non-overlapping with each other. Resource usage on such applications can be optimized by allocating a reduced common pool of shared resources.

As an illustration, consider the Point-of-sale (POS) and Enterprise data warehouse (EDW) applications used in a production environment of a brand during a week day of usage. Time of day and their CPU utilization is as represented in the chart below:



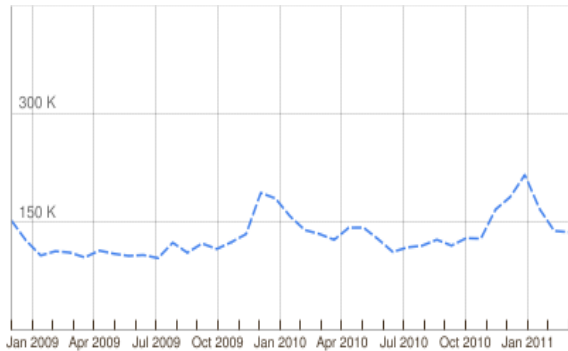
Currently, POS and EDW applications are allocated 10 and 22 instances all through the day. Instances are left unutilized during certain periods.

By pooling the POS and EDW into a virtual machine cluster and reducing the hardware for 10 instances used by POS, 30% reduction in number of instances can be achieved.

### Seasonal demand variations

Certain applications based on the industry and geography in which they are deployed, are exposed to seasonal variations in traffic. Traffic flow into eCommerce applications in North America are a classic example of this, during the holiday season (Nov, Dec, Jan) alone ecommerce sites like Amazon.com, Walmart.com receive 3X burst in traffic than their regular season traffic. This sudden burst in traffic would need the retailers to provision resources statically to cater to the peak loads and maintain them throughout the year. However, much of this capacity is under-utilized during the non-peak season. Retailers need to bear the costs of hardware, software licenses, network, power and people costs for maintenance. To reduce this overhead of costs, optimizing the resources based on such seasonal demand variability is considered.

Consider the daily unique visitors on a particular online store of a brand below:



There is an 80% increase in the traffic during every holiday season. Similar variation in traffic can be seen across brands, this would mean, increase in cost of computational capacity, storage and power utilization throughout the year.

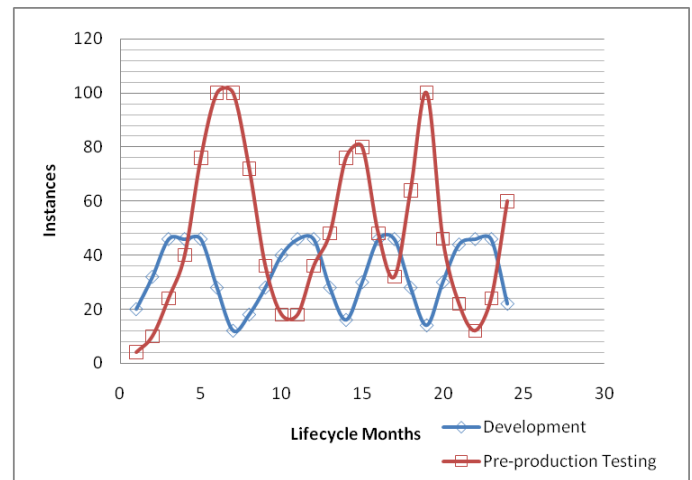
Brand	Average traffic	Peak traffic
Daily unique visitors	100000	180000
Daily number of page views	892233	1562200
Expected instance hours per annum (web servers, app server, db servers)	35040	52560
Additional instances hours used during peak season (Oct, Nov, Dec, Jan)	5760	
Total instance hours per annum	40800	52560

By resizing the ecommerce infrastructure for all the brands to average load instead of peak load, and rescaling the infrastructure during holiday season by using cloud bursting, about 22% optimization in the number of instances can be achieved.

### Optimize by life cycle usage variations

Organizations have independent environments for various software lifecycle stages, which include development, pre-production testing, staging, multiple production sites and recovery environments. The software development lifecycle and environments for each of these brands are independent today and are scheduled and sized according to their needs. By aligning the schedules of software lifecycle stages of several brands and pooling the computing resources costs can be reduced.

Below is an illustration of the lifecycles and instance utilization during the period by development and pre-production testing environments.



By aligning the lifecycle stages (for e.g. Agile sprints) and their schedules, at the peak of development, about 40- 60% of computing resources allocated for testing can be decommissioned while retaining their storage. Similarly at the peak of pre-production testing,

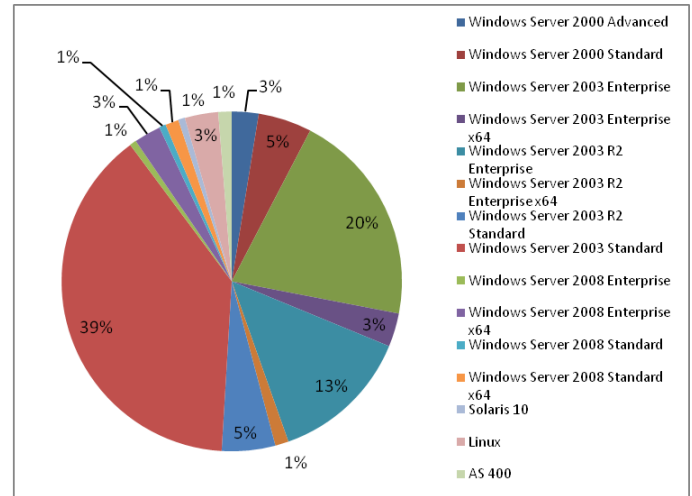
70% of computing resources allocated for development can be decommissioned.

Optimization is achieved by following these steps: Align the software development life cycle stages of multiple brands that have significant resources allocated for development and pre-production testing. Pool the computing resources for the brands into a virtual machine cluster with separate VM images as needed by development and pre-production testing. Create a resource allocation plan aligned with the software development lifecycle. Commission and decommission the computing resources as per the plan to reduce the resources.

### Optimize by standardizing platforms

Certain IT environments have high diversity of operating systems, software applications, tools, and hardware from various vendors and versions. A highly fragmented environment with different vendors and versions for applications limits the interoperability, limits the reuse of applications and components across the organization leading to high levels of redundant functionality, increases the lead time in provisioning and change management of hardware and software due to limitations in effectively managing diverse platforms.

Diversity of operating system and version variations in the IT environment of an example organization is illustrated below:

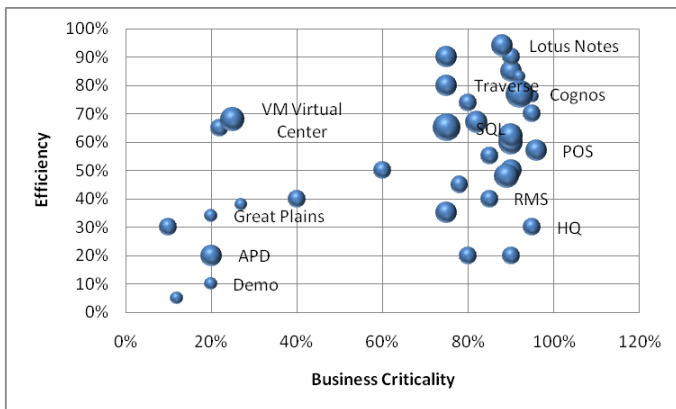


As can be seen, Windows Server 2003 and its versions have the highest deployment. This can be optimized by providing a catalog of standard virtual machine images with Windows Server 2003 as the selected platform.

### Optimize by application rationalization

Application rationalization helps to reduce overlapping and poorly utilized applications across brands in an enterprise. A current state chart of applications, their resource utilization efficiency, business criticality and cost is drawn for applications deployed in the IT environment of each brand.

Below is an illustration of current state chart of applications in the IT environment of a brand:



Applications can be categorized into four quadrants:

At the top right are the applications which are high on business criticality and resource utilization efficiency. The applications in top-right or top-left quadrants are moved to the cloud environment as-is. However, consolidation opportunities are explored where there are significant cost variations for applications with similar functionality.

At the bottom left quadrant are applications that are both low in resource utilization efficiency and low on business criticality. These applications can be significantly optimized by moving to cloud. Applications are either optimized for efficiency by resource consolidation or by application refactoring. If there are redundant applications with high costs on multiple IT environments, these applications are decommissioned and consolidated into a common pool.

At the bottom right quadrant are applications with low efficiency and high business criticality, these applications are either retained as-is or resource consolidated to gain higher efficiency.

## Conclusion

Cloud computing provides a cost-effective, dynamically adaptive, scalable, self service provisionable platform for IT services delivery. These factors are driving migration of legacy applications into cloud computing platform.

Cloud migration presents an opportunity to significantly reduce costs incurred on applications. The six levers of optimization presented here help enterprises to fully leverage the potential of the cloud and benefit from the migration activity.

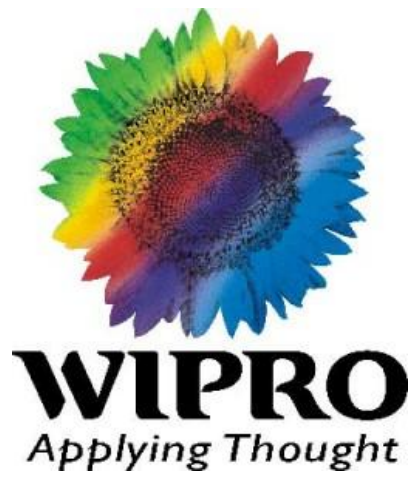
### About the authors



Jay Kulkarni is a Senior Architect with Wipro Technologies and has worked on ecommerce projects of large North American retailers like Walmart and Bestbuy. He is a core techie and has keen interest in cloud computing projects like Hadoop, Pig and Nutch. He has 12 years of experience in IT. He can be reached at [jayaprakash.kulkarni@wipro.com](mailto:jayaprakash.kulkarni@wipro.com).



Sanjay Agara heads the global practice on technology & architecture consulting for Retail customers within Wipro. He has been part of the IT industry for about eighteen years, a significant part of which spent on consulting with large retailers globally. He has played the role of a trusted advisor to CIOs and CTOs on many of their enterprise level initiatives. He comes with a sound understanding of domain combined with knowledge of enterprise-wide technologies to deliver solutions that not only meet the business objectives but also help in maximizing the value to the enterprise. He can be reached at [sanjay.agara@wipro.com](mailto:sanjay.agara@wipro.com).



**Wipro Technologies**  
*Innovative Solutions. Quality Leadership.*