relevance lab

WHITE PAPER

The Cloud Strategy: Balancing Act of Cost Optimization & Customer Experience

Abstract

Traditionally, a legacy organization faces a lot of friction in its business due to its legacy IT infrastructure as compared with its “born in the cloud” counterpart. For instance, a global organization had a large but stagnant domestic business and a small international business which had big potential to expand globally. Relevance Lab helped this organisation in adopting cloud technology in the right manner. That gave them scalability, business agility and opportunity to increase the business velocity, reducing friction. The second stage was to have a disciplined and a well-planned IT approach/strategy to manage day-after-cloud scenario, to keep the cost of cloud infrastructure low at the optimal usage of infrastructure and improve the customer experience, by making sure business applications are always available to users.

This white paper elaborates the best IT strategy that organizations can adopt in their cloud journey to expand the business internationally for exponentially increasing user base, while keeping the Cloud IT costs optimal without compromising on the customer experience.

Business and Technology Problems

In the pursuit of expanding its business internationally, Relevance Lab helped the client migrate its entire IT estate to the cloud. This was a multi-region environment where services were hosted independently in multiple data centres catering to 0.5 million users (teachers, students, supporting staff and sales community) through various business applications. The next important step was to manage the entire IT estate (Applications & Infrastructure) in the cloud and this cloud management was extremely complex due to the following reasons:

Independent services hosted across various AWS data centres (Central, Ireland, Sydney, Seoul, China) to cater to countries around that geography

Due to time zone differences for each geographical region, IT maintenance was a challenging task.

Business and Data Privacy Law for each country was different and it was required to adhere to those requirements with risk and compliance perspective.

Maintaining the same level of user experience was very critical for the exponentially increasing user base and it was important that applications were always available and system should take care of the increasing load. Hence, Application Availability was one of the important parameters with IT perspective.

Infrastructure was provisioned in an unoptimized manner. There were 300+ EC2 instances and 20+ RDS instances to support all the production and non-production environments.

Annual Cost of IT Infrastructure was high, despite using Cloud environment.

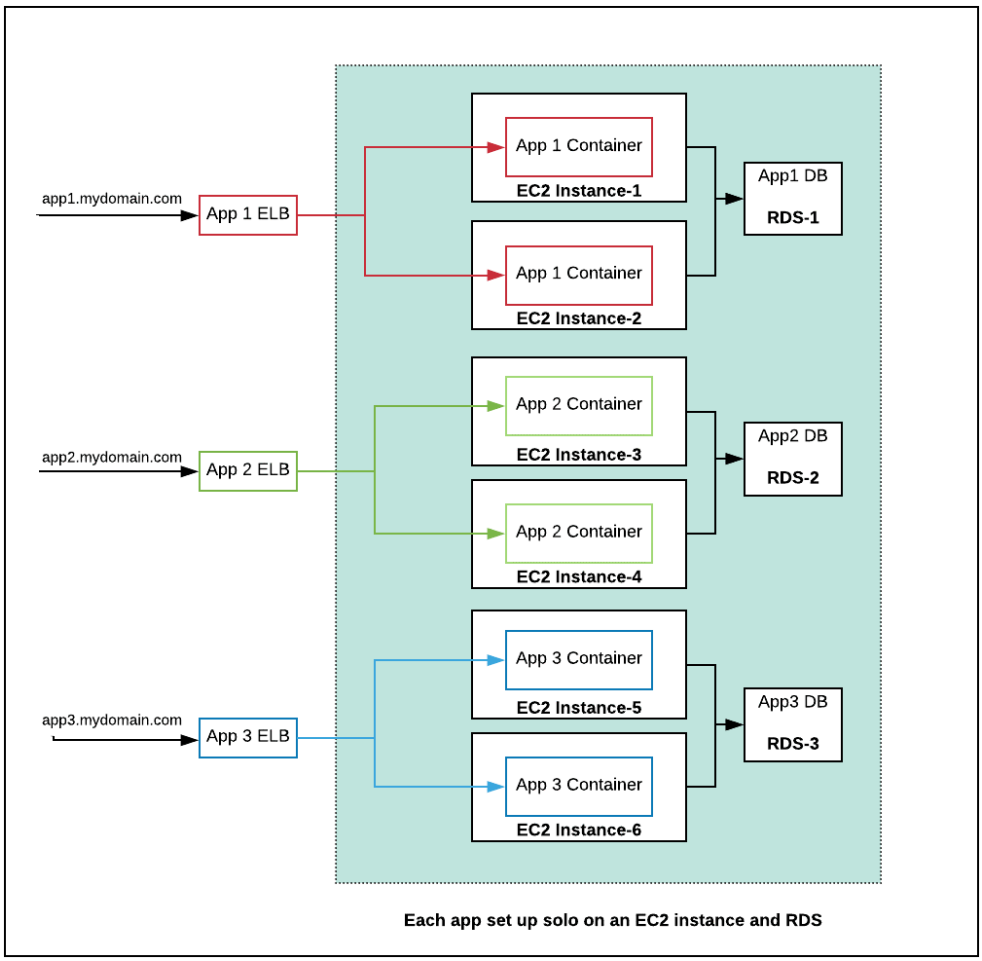
Approach and Solution

Relevance Lab suggested the following 4 main optimization initiatives to keep the cost of Cloud IT minimal without compromising on the customer experience for the exponentially increasing customer base.

1. Server Consolidation

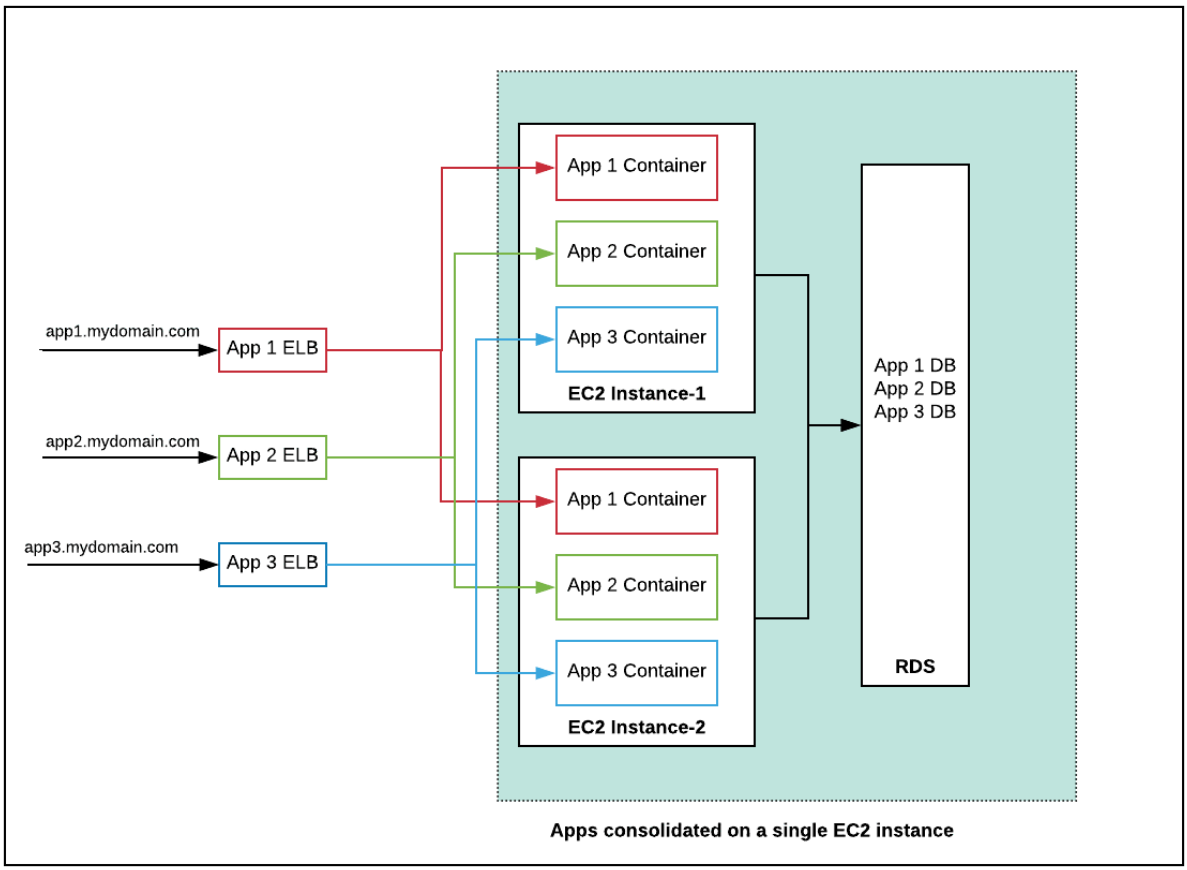
The Need: A typical application in the cloud is setup on an EC2 instance with its database on an RDS. To ensure redundancy and high availability on the compute layer, multiple (typically two) such EC2 instances are required which are attached to the load balancer for equal distribution of load among themselves. Similarly, production RDS are typically configured for Multi-AZ deployments to enhance availability and durability of the database layer. This creates a minimum footprint requirement for each application which is most likely over-provisioned for capacity but promises increased service availability. For 3 such applications, 6 EC2 instances and 3 RDSs were to be provisioned.

In a global environment, there were 5 such independently hosted setup to serve various regions/countries across the globe, replicating the same infrastructure in each data centre. Thus, the costs of all inefficiencies get multiplied based on number of such hostings.



The Approach: These being legacy applications, weren't designed to be stateless. Thus, containerization was not possible without significant architectural changes. Each such instance was 4GB/2CPU as the application needed ~3GB of memory. However, when their CPU utilization was observed, the average consumption was mostly around 30%. Though occasional peaks were observed, not all applications had the spikes getting raised at the same time. This paved the way for server consolidation, considering the average CPU utilization of all the 3 apps at any point of time will rarely exceed 90% when set up on a single instance.

In place of spawning an individual instance for each app totalling to 6 instances (each 4GB/2CPU), all 3 apps were made to run on a single instance with higher memory (12GB/2CPU) on different ports. Each application had its maximum memory capped through the JVM options and all the other resources were shared. All ELBs were configured to attach to the same set of instances but were forwarding the traffic on the respective application ports. This reduced the requirement from 6 (4GB) instances to 2 (12GB) EC2 instances with the same compute capacity. Similarly, all application schemas could be created under a single RDS and the IOPS were provisioned keeping the consolidated application requirement in mind. This approach brought drastic reduction in the consolidated minimum footprint, thereby reducing the costs.

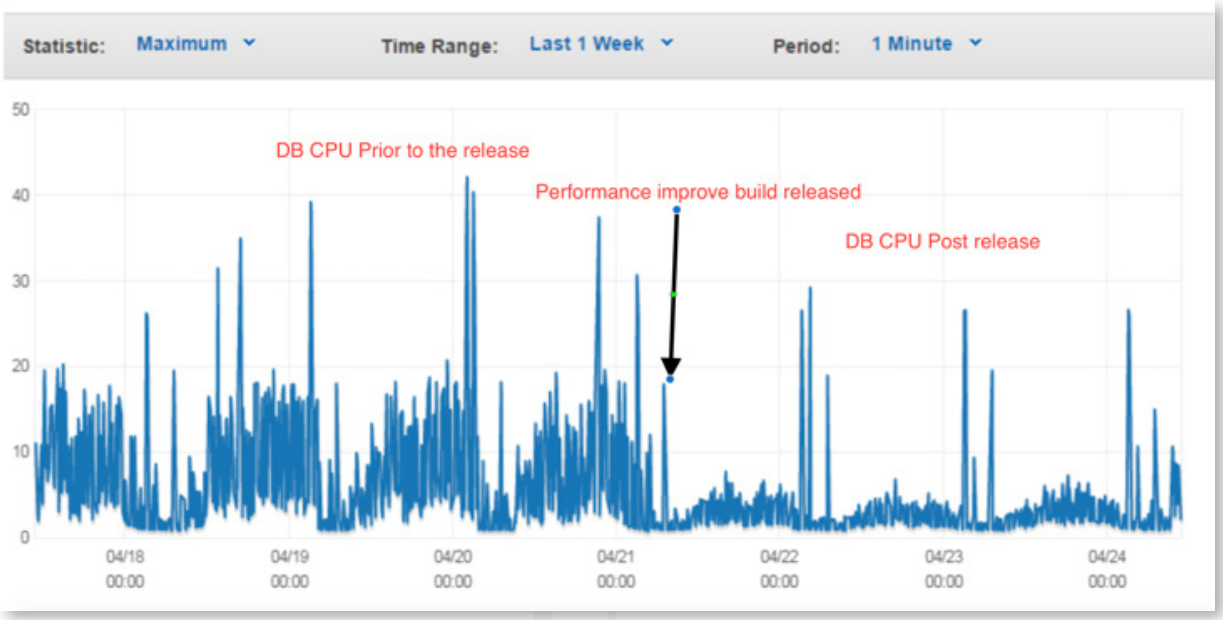


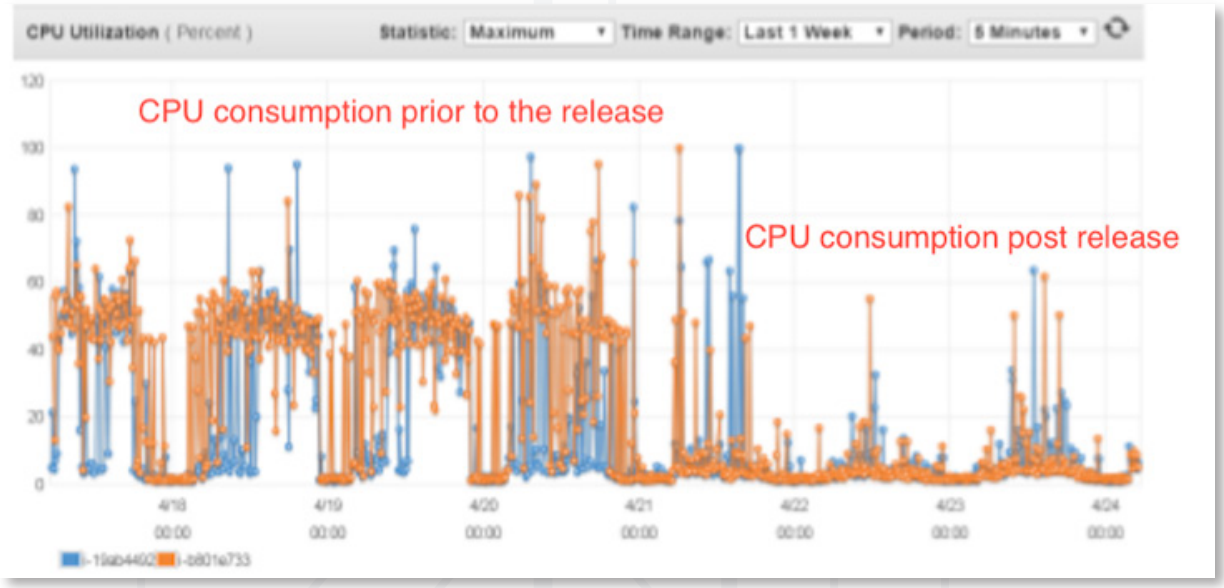
Apps consolidated on a single EC2 instance

2. Application Performance Monitoring (APM) – Tool-Driven Performance Improvements

The Need: With applications setup in consolidated mode, each application being performant became all the more important. Any inefficiency in one application causing resource consumption spikes had good chance to adversely affect the others. Additional capacity had to be provisioned and frequent reboot was required to handle geographical regions with high usage and application crashes. These situations demanded applications to be highly performant and efficient, in order to increase their service availability.

The Approach: Correctly identifying the inefficiencies and measuring improvements with elimination of inefficiencies in every build has been the way to address this. APM tool-based analysis and a rigorous performance test suite were designed depicting a typical peak business day.





3. Switch to Reserved Instances

The Need: Applications that got consolidated, optimized and had a consistent traffic pattern reached a steady-state usage. Major portions of the computation and storage requirements became predictable, enabling us to get into a long-term usage commitment on the infrastructure footprint with the cloud provider. When switching to reserved instances, much lower hourly rate can be availed for the same infrastructure as compared to the on-demand mode billing.

The Approach:

Identify the list of infrastructures that had consistent and predictable usage throughout the year—these included not only the production EC2 instances and RDS but also the support infrastructure such as build machines, VPN servers and non-prod environments.

Selection of appropriate RI mode based on available options. If the instance was used for 7 months a year, its cost under RI got fully recovered when compared to its on-demand cost. Choosing one year term with a pay-all-upfront option resulted in ~40% reduction in the bill.

4. Optimize Ongoing Cloud Operations

The Need: Discipline was needed in the team to define the right processes and keep a close watch on the ongoing operations so that no inefficiencies get back into the system.

The Approach: The following measures were initiated and made a part of the ongoing operations:

Optimize Ongoing Cloud Operation measures

- Eliminated all swim lanes (replicas spawned for resolving production defects), fixed the number of non-prod environments (dev, qa, stage) per application. Teams were asked to coordinate within the available environments instead of getting new swimlanes created.
- Optimized the backup (retention) strategy and got rid of unwanted backups and snapshots, thereby considerably reducing dead-weight storage.
- Weekly utilization and performance reports were created to monitor the provisioned infrastructure. Approval workflow was set up demanding justification for spawning any new instances. Audit checklists were prepared and a process was put in place for performing regular audits against these checklists.

This helped in further cost reduction by reducing the number of

Load Balancers

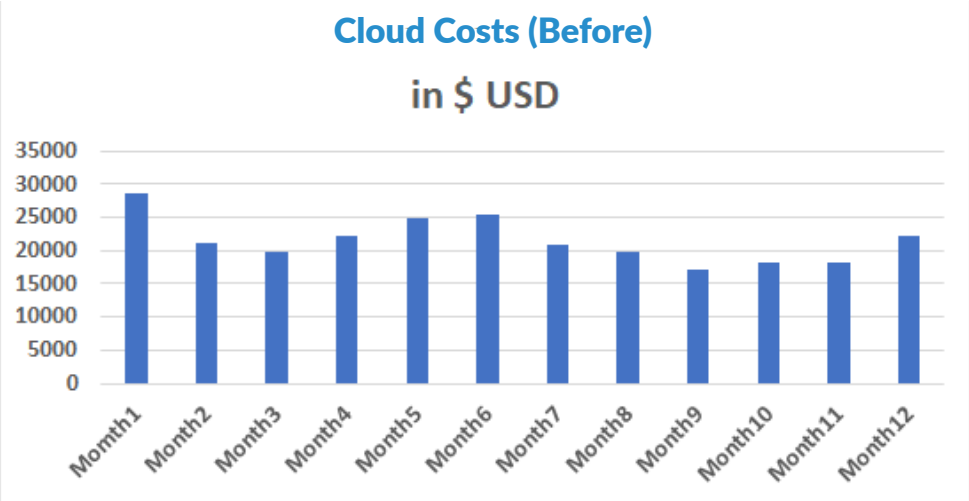
Storage & Data Transfer

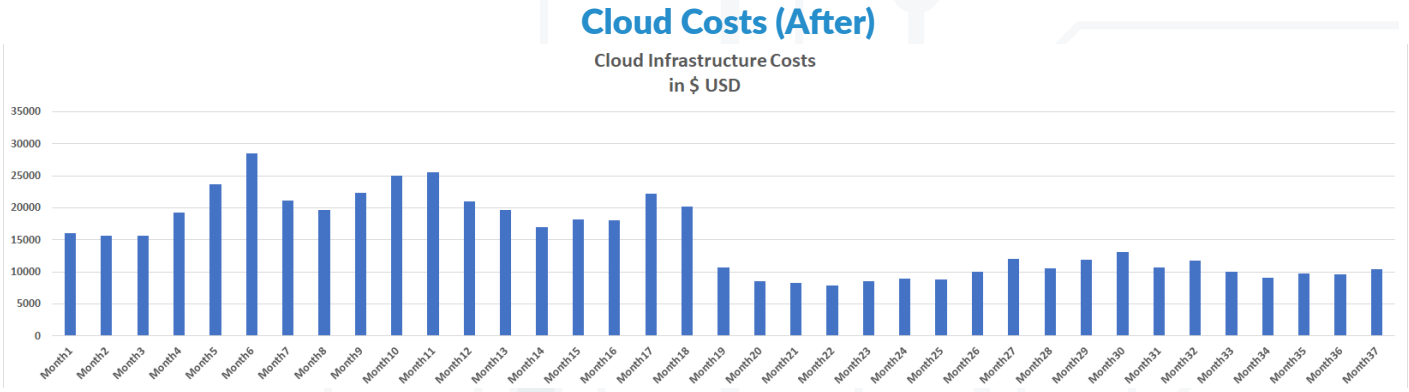
Cloud-Front Distribution costs

Benefits Delivered

- **The overall EC2 and RDS instances dropped down by 58% ensuring optimal utilization of Cloud Infrastructure.**

- **As depicted in the graphics below, the disciplined approach helped the client in reducing the overall hosting cost by 41% and per-user IT Infrastructure cost by 65%.**





While the user base was increased from 0.5mn to 1mn, system could take the added load of increased demand. The overall Application Performance was improved ensuring 100% availability of all applications.

Key Learnings:

Here are the key learnings that we would like to elaborate for organizations that have infrastructure in the cloud:

- **Invest in tools to provide the information on unused/under-utilized infrastructure.**
- **Invest in application performance management and monitoring tools. Inefficiencies in the application is directly proportional to its infrastructure requirement.**
- **Do not spawn a new instance/environment for every new request. Consolidate and share as much as possible. How would you have operated if you weren't in the cloud?**
- **Control the provisioning of new infrastructure through an approval process. Have processes for regular audits and utilization reports.**

About Relevance Lab:

Relevance Lab is a DevOps and Automation Specialist company, driving frictionless business and enabling large enterprises to disrupt traditional delivery models of Infrastructure Management, Application Development and Service Delivery Operations with faster velocity, better quality and optimized spend using new technology. Relevance Lab is a partner of choice for a number of large enterprises and technology companies globally in their adoption of Cloud Computing, Big Data Analytics and Digital Solutions for their business transformation initiatives

www.relevancelab.com

For more information on our product offerings or solutions to cloud adoption please write to us at *marketing@relevancelab.com*