

The two main objectives of this problem statement is to reduce cost and efficiently deploy resources as needed.

Here's a few of my suggestions::

1. Since booking requests come in real-time, we could use the data we have accumulated over the years and perform a trend analysis. We could analyze the bookings made by the customers and identify the most regular and predictable ones in terms of frequency and quantity of workloads requested. Also figure out the minimum number of resources that have been requested at any given time by all the customers put together.
2. We could then book dedicated instances on AWS based on the number of workloads being in use at all times. Dedicated instances on AWS will cost us significantly less than deploying on-demand instances. Compared to on-demand, reserved ones will spare us between 40%-60% of expenses. Costs can be lowered further if an upfront payment is made in partial or, even better, full; partial upfront will fetch us discounted hourly rates while full payment involves no additional cost, regardless of hours used. AWS offers 1 or 3 year commitment; so this kind of analysis will render us to choose the best option.
3. Unexpected/unpredictable requirements can be addressed using the instances in the on-demand pool. They can be deployed using predefined AMI templates that have been tried and tested to host the required applications and configurations needed for load testing.
4. Deploying on-demand instances must be dynamic and reduce as much operational overhead as possible. This could be achieved using auto scaling groups. The group will start with a minimum number of instances (this number could be derived from the pattern that we have seen in the past) and automatically scale up or down depending on the demand. Once a load test is completed, the instances can be automated to terminate.
5. It is important to prioritize using reserved instances over on-demand. All reserved instances must be occupied before launching new on-demand ones.
6. Taking inspiration from AWS, we could encourage our customers to plan their load tests and book workloads in advance for an incentive of a discounted price. This will aid us to foresee our own future requirements.
7. Analysis of trends must become an ongoing process for us to stay in the best shape. It's important to constantly look for changes in requirements and act upon them accordingly.