[Poster session] Technical Report

Cost saving recommendation for workloads with burstable performance

Satish Kumar Jaiswal[1,a)] Satoshi Kaneko[1,b)] Shinichi Hayashi[1,c)]

1. Background

1.1 Burstable instance

Public cloud vendors provide burstable instances [1] types which offer a certain fraction of CPU utilization at lower cost. They can also support full CPU usage whenever required based on the availability of CPU credit (performance is capped if CPU credit is 0). If appropriately selected, burstable instances can provide upto 25%.

1.2 Problem

Optimal selection of burstable instance is not an easy task. First of all, there is a risk that if a burstable instance type is not selected appropriately then there could be various types of penalty such as performance capping or incurring of additional charges. Secondly, CPU credit-based performance and pricing model (standard mode, unlimited mode) of burstable instance type is complex, and it differs from one vendor to another adding an extra layer of complexity. Thirdly, forecast data is required for making recommendations which will be valid in the future. Due to uncertainty in forecast there is uncertainty in penalty estimation. Uncertainty in penalty estimation makes it difficult to decide appropriate specification of burstable instance types.

1.3 Our contribution

We propose a method to create a computer program for making reliable cost saving recommendations for workloads with burstable performance by matching them to suitable burstable instance type in the public cloud. Our method can recommend burstable instances for 95% of on-premises instances in our customer data center.

2. Proposed Method

The proposed method provides the instance specification, cost and penalty associated with the burstable instance in the public cloud when a workload with burstable performance is migrated onto it.

First of all, we narrow down the on-premises instances for which forecast accuracy for CPU utilization is beyond certain threshold (decision box in Fig. 1). This is to ensure that our recommendation is accurate (solution to third problem in Section 1.2). We use Random Forest [2] Regressor for timeseries forecasting. Then we select burstable instances in public cloud which have similar configuration (no. of vCPU and size of RAM). We obtain a list of candidate burstable instances each with their own baseline.

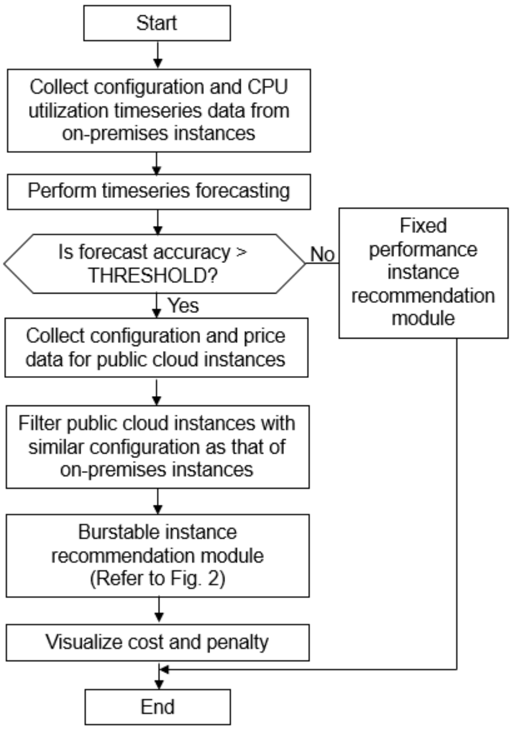
Start
Collect configuration and CPU utilization timeseries data from on-premises instances
Perform timeseries forecasting
Is forecast accuracy > THRESHOLD?
No
Yes
Fixed performance instance recommendation module
Collect configuration and price data for public cloud instances
Filter public cloud instances with similar configuration as that of on-premises instances
Burstable instance recommendation module (Refer to Fig. 2)
Visualize cost and penalty
End

Fig. 1 Flowchart of recommendation program

Next, we determine which burstable instance among the candidates obtained above can be used and how much is the

[1] Services Computing Research Department, Hitachi Ltd. Kokubunji, Tokyo 185–8601, Japan
[a)] satish.jaiswal.ty@hitachi.com
[b)] satoshi.kaneko.bc@hitachi.com
[c)] shinichi.hayashi.ez@hitachi.com

expected penalty (top 4 steps in Fig. 2). The burstable instance is determined such that its baseline is sufficient enough to support its forecasted CPU utilization (forecast series in Fig. 3) without any penalty. However, penalty may occur when forecast is inaccurate. We consider 95% confidence interval (upper series in Fig. 3) to calculate expected penalty. We also recommend another burstable instance such that its baseline is sufficient enough to support upper series without penalty (bottom 2 steps in Fig. 2). Choosing to migrate on this instance means one can be 95% sure not to get any penalty. Thus, we recommend two burstable instances for a single mode of operation.

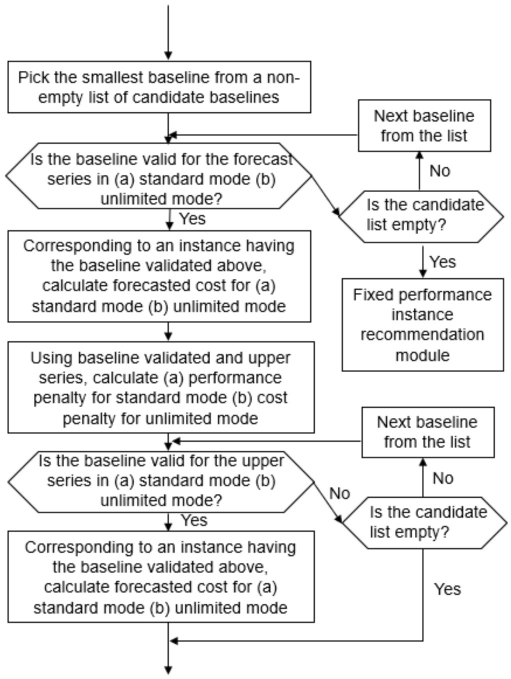
Pick the smallest baseline from a non-empty list of candidate baselines
Next baseline from the list
Is the baseline valid for the forecast series in (a) standard mode (b) unlimited mode?
No
Is the candidate list empty?
Yes
Yes
Corresponding to an instance having the baseline validated above, calculate forecasted cost for (a) standard mode (b) unlimited mode
Fixed performance instance recommendation module
Using baseline validated and upper series, calculate (a) performance penalty for standard mode (b) cost penalty for unlimited mode
Next baseline from the list
Is the baseline valid for the upper series in (a) standard mode (b) unlimited mode?
No
No
Is the candidate list empty?
Yes
Corresponding to an instance having the baseline validated above, calculate forecasted cost for (a) standard mode (b) unlimited mode
Yes

Fig. 2 Flowchart of burstable instance recommendation module in Fig. 1

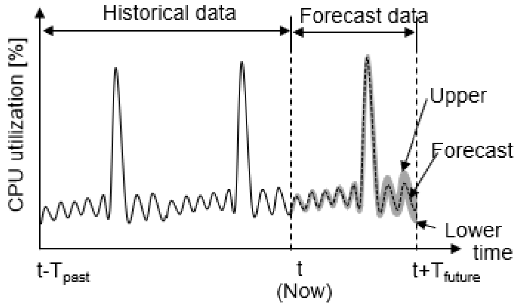

Fig. 3 Forecast series and upper series in forecast data

As shown in Fig. 4, four candidate burstable instances are recommended, two for standard mode and next two for unlimited mode of operation. In case of standard mode, the performance penalty is shown using heat map. On the other hand, cost penalty is shown using an error bar in case of unlimited mode. A vertical dotted line represents the cost of next fixed performance instance.

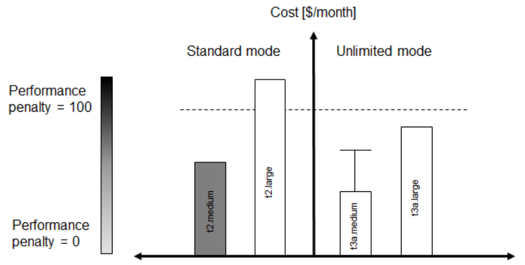

Fig. 4 Recommended burstable instances with penalty

3. Evaluation and result

According to our preliminary study on our customer ' s data, we found that 95% of VMs are overprovisioned, i.e. more than 50% of VMs have less than 5% CPU utilization. These overprovisioned VMs can be migrated to burstable instance in public cloud. By doing so, the customer can save up to 25% of their cost. This is because the price difference between fixed performance instance and burstable instance having similar configuration (no. of vCPU and size of RAM) can be as high as 25% in case of AWS [*1].

The evaluation requires that customer's on-premises VMs are migrated to recommended AWS [*1] burstable instances, and confirmed that cost and penalty are as reported by our method. Since migration of customer's on-premises VMs to AWS [*1] burstable instance is not feasible due to confidentiality. We have devised an alternative evaluation method. According to the alternative evalution method, workloads running on AWS [*1] fixed performance instances (e.g. m5.large, m5.xlarge, m5.2xlarge) should be migrated to recommended AWS [*1] burstable instances, and confirmed that the cost and penalty are as reported by our method. The evaluation method is under study.

4. Conclusion and Future Work

We proposed a method to create a computer program for making reliable cost saving recommendations for workloads with burstable performance by matching them to suitable burstable instance type in the public cloud. The proposed method provides the instance specification, cost and penalty associated with the burstable instance in the public cloud when a workload with burstable performance is migrated onto it. As a future work, we will evaluate the proposed method according to evaluation method proposed in Section 3.

References

[1] Burstable performance instances, `https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/burstable-performance-instances.html`

[2] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[*1] AWS is a registered trademark of Amazon.com, Inc. or its affiliates.