Demonstration: Predicting Distributions of Service Metrics

Forough Shahab[†], Rolf Stadler[†§], Andreas Johnsson[‡], Christofer Flinta[‡]
[†] KTH Royal Institute of Technology, Sweden – Email: {foro, stadler}@kth.se , [‡]Ericsson Research, Sweden – Email: {andreas.a.johnsson, christofer.flinta}@ericsson.com, [§] RISE SICS

***Abstract*—The ability to predict conditional distributions of service metrics is key to understanding end-to-end service behavior. From conditional distributions, other metrics can be derived, such as expected values and quantiles, which are essential for assessing SLA conformance. Our demonstrator predicts conditional distributions and derived metrics estimation in real-time, using infrastructure measurements. The distributions are modeled as Gaussian mixtures whose parameters are estimated using a mixture density network. The predictions are produced for a Video-on-Demand service that runs on a testbed at KTH.**

***Index Terms*—Service Engineering, Service Management, Machine Learning**

I. Background/Concepts

Understanding and predicting the performance of telecom services is difficult due to the complexity of software systems which run these services over general-purpose platforms and operating systems. Recent approaches to performance prediction are based on statistical learning, whereby models are trained with data collected from infrastructure and services. These models are then used to predict service quality, the likelihood of component failures, etc.

Such predictions are generally modeled as *point values of continuous variables*, for instance, the mean response time of a service or the mean time for a system component to fail. Regression methods, based on linear regression, random forest, or neural networks, for instance, are used to predict the mean of a target variable (e.g., the response time), conditioned on the input (i.e., infrastructure measurements). Predicting point values like the conditional mean, however, provides only a limited description of the target variable, as explained in [1].

In our recent work [1], we present a method for predicting and evaluating *conditional distributions* of service metrics. From such distributions, key statistics, including mean, variance, or percentiles can be derived. In our work, distributions are modeled as Gaussian mixtures, whose parameters are predicted using *mixture density networks (MDN)*, a class of neural networks [2].

Fundamental to this demonstration is the concept of *Conditional Distribution Estimation (CDE)*. We use MDN models to predict the conditional distribution of the video frame rate for a Video-on-Demand (VoD) service in real-time. Figure 2 shows three derived metrics: expected values, quantiles, and aggregated density over a time interval. Such metrics are essential to understand the service behavior and predict SLA conformance.

II. Testbed

The demonstration uses a system that implements the above approach in an on-line setting as illustrated in Figure 3. A management station provides access to the KTH testbed and displays, in real time, measurements and predictions from the the service and the testbed infrastructure [3] [4].

The testbed includes a server cluster in our laboratory at KTH. It comprises ten high-performance machines interconnected by Gigabit Ethernet. Nine of them are Dell PowerEdge R715 2U servers, each with 64 GB RAM, two 12-core AMD

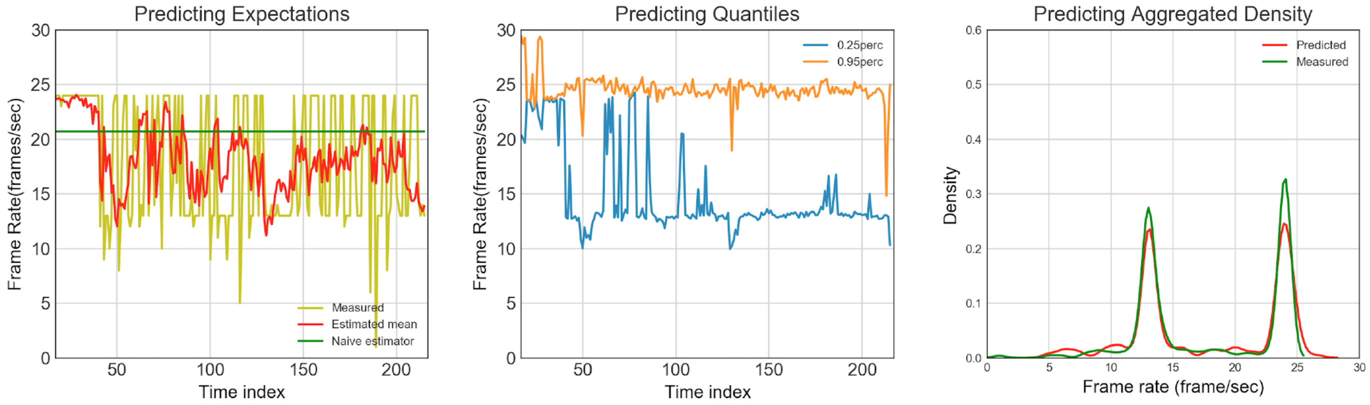

Fig. 1: Demonstration screen 1: The evolution of derived metrics—Predicting expected values, percentiles, and aggregated density of frame rates for a VoD service.

978-3-903176-15-7 © 2019 IFIP

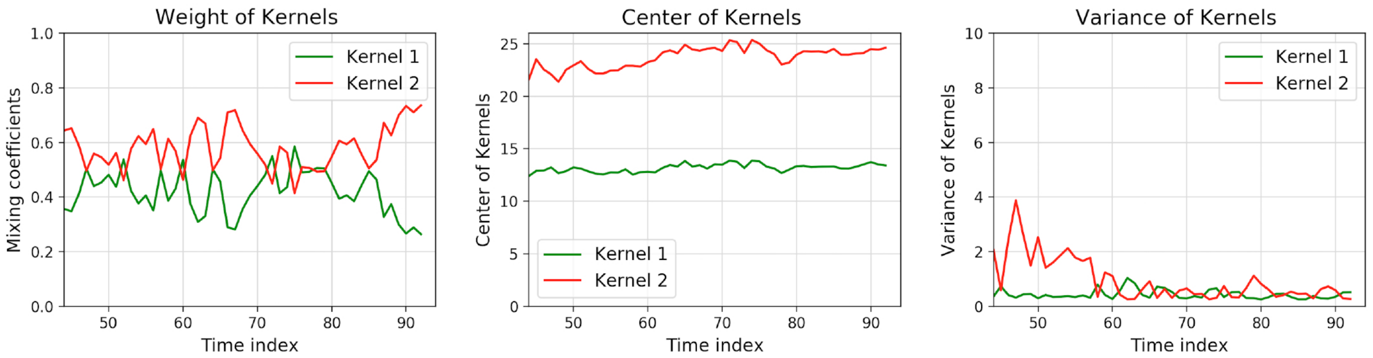

Fig. 2: Demonstration screen 2: The evolution of predicted model parameters of Gaussian mixtures.

Opteron processors, a 500 GB hard disk, and four 1 Gb network interfaces. The tenth machine is a Dell PowerEdge R630 2U machine with 256 GB RAM, two 12-core Intel Xeon E5-2680 processors, two 1.2 TB hard disks, and twelve 1 Gb network interfaces. All machines run Ubuntu Server 14.04 64 bits, and their clocks are synchronized through NTP.

The VoD service is deployed on six PowerEdge R715 machines: one HTTP load balancer, three web server and transcoding machines, and two network file storage machines. The details of load balancer, web servers, transcoding machines, and storage machines are given in [4].

Request generators, emulating client populations, and the VoD web servers are connected via an emulated OpenFlow network. The network is virtualized on the PowerEdge R630 described above. We use Virtual Box as hypervisor. Each OpenFlow switch and controller runs in a virtual machine with 1 core and 4 GB RAM for switches and 4 cores and 8 GB RAM for the controller. The network topology has 14 switches with total of 44 ports [4]. The Real-time Analytics Engine in Figure 3 is written in Python and makes use of the Keras package [5].

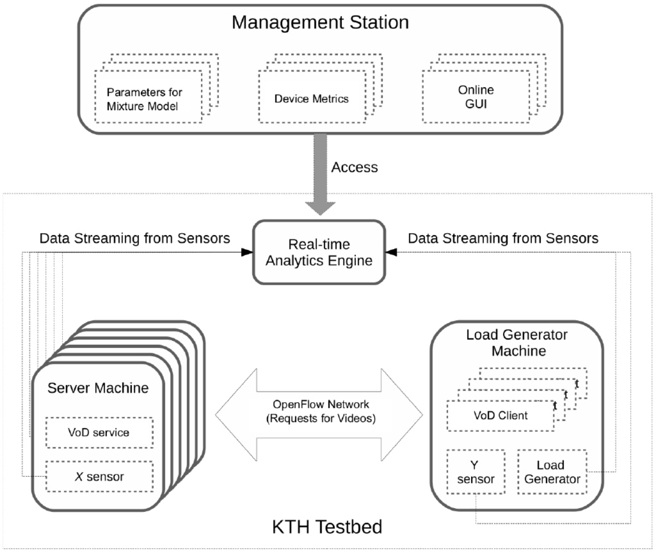
Management Station
Parameters for Mixture Model
Device Metrics
Online GUI
Access
Data Streaming from Sensors
Real-time Analytics Engine
Data Streaming from Sensors
Server Machine
VoD service
X sensor
OpenFlow Network (Requests for Videos)
Load Generator Machine
VoD Client
Y sensor
Load Generator
KTH Testbed

Fig. 3: Architecture with components from testbed, service, analytics engine.

III. Demonstration

We demonstrate the real-time computation of model parameters and the real-time prediction of service metrics based on those parameters. The real-time analytics engine shown in Figure 3 reads from a trace file (instead of from data feeds from the test bed) and produces a continuous stream of estimations and predictions. The trace is taken from the VoD scenario described in [1]. The models for prediction have been precomputed and the demo shows the execution of these models on unseen data.

Demo screen 1 shows three service metrics that are derived from the parameters of the mixture models as displayed on screen 2. We show the the expected video frame rates and compare them with measurements. Second, we show predicted quantiles, specifically 25 and 95 percentiles of the video frame rate. Lastly, we show predicted and measured aggregate density distributions of the video frame rate over the window size of 200 seconds [1].

Demo screen 2 shows the evolution of the model parameters as they are computed. To make the graphs better readable, we produce Gaussian mixture models with two kernels only. The screen shows the weight parameters for the kernels, the center of the kernels, and the variance of the kernels.

Acknowledgements

This research has been supported by the Swedish Governmental Agency for Innovation Systems, VINNOVA, through projects SENDATE-EXTEND and AutoDC, as well as the KTH Software Research Center CASTOR.

References

[1] F. Shahab Samani and R. Stadler, “Predicting distributions of service metrics withneural networks,” in *Network and Service Management (CNSM), 2018 14th International Conference on*. IEEE, 2018.

[2] C. M. Bishop, “Mixture density networks,” 1994.

[3] R. Yanggratoke, J. Ahmed, J. Ardelius, C. Flinta, A. Johnsson, D. Gillblad, and R. Stadler, “Predicting service metrics for cluster-based services using real-time analytics,” in *Network and Service Management (CNSM), 2015 11th International Conference on*. IEEE, 2015, pp. 135–143.

[4] R. Stadler, R. Pasquini, and V. Fodor, “Learning from network device statistics,” *Journal of Network and Systems Management*, vol. 25, no. 4, pp. 672–698, 2017.

[5] Keras, 2018. [Online]. Available: https://keras.io/