Powering the Service Responsiveness of Deep Neural Networks: How Queueing Models can Help

Evgenia Smirni
College of William and Mary
Williamsburg
VA, USA
esmirni@cs.wm.edu

ABSTRACT

Deep neural networks (DNNs) enable a host of artificial intelligence applications. These applications are supported by large DNN models running in serving mode often on a cloud computing infrastructure. Given the compute-intensive nature of large DNN models, a key challenge for DNN serving systems is to minimize user request response latencies. We show and model two important properties of DNN workloads that can allow for the use of queueing network models for predicting user request latencies: homogeneous request service demands and performance interference among requests running concurrently due to cache/memory contention. These properties motivate the design of a dynamic scheduling framework that is powered by an interference-aware queueing-based analytic model. The framework is evaluated in the context of an image classification service using several well known benchmarks. The results demonstrate its accurate latency prediction and its ability to adapt to changing load conditions, thanks to the fast deployment and accuracy of analytic queuing models. This work is in collaboration with Feng Yan of the University of Nevada at Reno, and Yuxiong He and Olatunji Ruwase of Microsoft Research. The interested reader is directed to [1] for details.

Categories and Subject Descriptors

C.4 [**Computer Systems Organization**]: Performance of Systems

Keywords

scheduling; performance guarantees; workload characterization, deep neural networks

ICPE '17 Companion April 22-26, 2017, L'Aquila, Italy

ACM ISBN 978-1-4503-4899-7/17/04.
DOI: http://dx.doi.org/10.1145/3053600.3053620

Bio

Evgenia Smirni is the Sidney P. Chockley Professor of Computer Science at the College of William and Mary, Williamsburg, VA, USA. Her research interests include queuing networks, stochastic modeling, Markov chains, resource allocation policies, Internet and multi-tiered systems, storage systems, data centers and cloud computing, workload characterization, and modeling of distributed systems and applications. She has served as the Program Co-Chair for DSN'17, ICPE'17, QEST'05, ACM Sigmetrics/Performance'06, and HotMetrics'10. She also served as the General Co-Chair for QEST'10 and Numerical Solutions of Markov Chains 2010 (NSMC'10). She is an ACM Distinguished Scientist, a senior member of IEEE, and a member of the Technical Chamber of Greece.

Acknowledgments

This work has been partially supported by NSF grants CCF-1218758 and CCF-1649087.

Bibliography

[1] Feng Yan, Yuxiong He, Olatunji Ruwase, Evgenia Smirni. *SERF: efficient scheduling for fast deep neural network serving via judicious parallelism*, in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, Supercomputing 2016: 26:1-26:12