A Large-scale Service System with Packing Constraints

Dr. Alexander Stolyar
Lehigh University

Wednesday, April 20, 11AM, SERC 306

Abstract: The model is motivated by the problem of efficient real-time "packing" of virtual machines into physical host machines in a network cloud data center. There is infinite number of servers (physical machines) and multiple types of customers (virtual machines). Each server can simultaneously serve several customers, subject to some packing constraints. Service times of different customers are independent - even if they share a server. The objective is to minimize the number of occupied servers. We show that some versions of greedy packing are asymptotically optimal as the system scale goes to infinity. [In part, this is joint work with Y. Zhong (Columbia Univ.)]

Bio: Since 2014 Alexander Stolyar is a Timothy J. Wilmott Endowed Chair Professor in the Department of Industrial and Systems Engineering at Lehigh University. His research interests are in stochastic processes, queueing theory, and stochastic modeling of communication, information and service systems. He received Ph.D. in Mathematics from the Institute of Control Science, USSR Acad. of Science, Moscow, 1989, and was a research scientist at the Institute of Control Science in 1989-1991. In 1992-1998 he was working on stochastic models in telecommunications at Motorola and AT&T Research. From 1998 to 2014 he was with the Bell Labs Mathematical Sciences Research (Murray Hill, New Jersey), working on stochastic networks and resource allocation problems in a variety of applications, including service systems, wireless communications, network clouds. He received INFORMS Applied Probability Society 2004 Best Publication award, SIGMETRICS'96 Best Paper award. He currently serves on editorial boards of Advances in Applied Probability and Queueing Systems.