GTA / UFRJ
GRUPO DE TELEINFORMÁTICA E AUTOMAÇÃO

PALESTRA

Prof. Antonio Marinho Pilla Barcelos

Universidade Federal do Rio Grande do Sul

Palestra: Dia 24/11/2015 das 11:00 às 12:00 – Sala H - 301

Achieving Predictable and Work-Conserving Performance in Datacenters

***Abstract**: Avoiding performance interference in datacenter networks (DCNs) is challenging. Evidence indicates that throughput achieved by virtual machines (VMs) in current datacenters can vary by a factor of five or more, leading to poor and unpredictable overall application performance. Despite valuable efforts, there are important aspects yet to be addressed properly, including poor utilization of resources, mechanism overheads, and absence of bandwidth guarantees. Besides, there lack schemes that optimize performance considering both network and processing resources. In this talk, we'll present three proposals that address performance interference in DCNs: IoNCloud, Predictor and Packer. IoNCloud leverages the key observation that temporal bandwidth demands of cloud applications do not peak at exactly the same time. Therefore, it seeks to provide predictable and guaranteed performance while minimizing network underutilization by (a) grouping applications in virtual networks (VNs) according to their temporal network usage and need of isolation; and (b) allocating these VNs on the cloud substrate. Despite achieving its objective, IoNCloud does not provide work-conserving sharing among VNs, which hurts provider revenue. Predictor, an evolution over IoNCloud, leverages Software-Defined Networking (SDN) and uses two novel algorithms to provide network guarantees with work-conserving sharing. Furthermore, Predictor is designed with scalability in mind, taking into consideration the number of entries required in flow tables and flow setup time in DCNs with high turnover and millions of active flows. Despite the benefits, at allocation time IoNCloud and Predictor consider only network resources (while resources such as CPU and memory are allocated according to time slots). This leads to fragmentation of non-network resources and, consequently, results in less applications being allocated in the infrastructure. Packer, in contrast, aims at providing predictable and guaranteed network performance while minimizing overall multi-resource fragmentation. The key insight is that applications have complementary demands across time for multiple resources. To enable multi-resource allocation, we devise a new abstraction for specifying application requirements, called Time-Interleaved Multi-Resource Abstraction (TI-MRA), which allows the specification of multi-resource requirements across time.*

Biography: Prof. Marinho P. Barcellos received a PhD degree in Computer Science from University of Newcastle Upon Tyne (1998). In 2003-2004, he worked in a joint project between University of Manchester and British Telecomm research labs on high-performance multicast transport. Since 2008 Prof. Barcellos has been with the Federal University of Rio Grande do Sul (UFRGS), where he is an Associate Professor. He has authored many papers in leading journals and conferences related to computer networks, network and service management, distributed systems, and computer security, serving as both TPC member and TPC chair. He has authored book chapters and delivered tutorials and invited talks. His work as a lecturer has been consistently distinguished by graduating students. Prof. Barcellos was the elected chair of the Special Interest Group on Computer Security of the Brazilian Computer Society (CESeg/SBC) 2011-2012. He is a member of SBC and ACM. His current research interests are cloud computing data center networks, software-defined networking, information-centric networks and security aspects of those networks. He is the Program Co-Chair of IEEE P2P 2015, SBRC 2016, and the General Co-Chair of ACM SIGCOMM 2016.