Architecture and micro-architecture support for virtual machines and datacenters
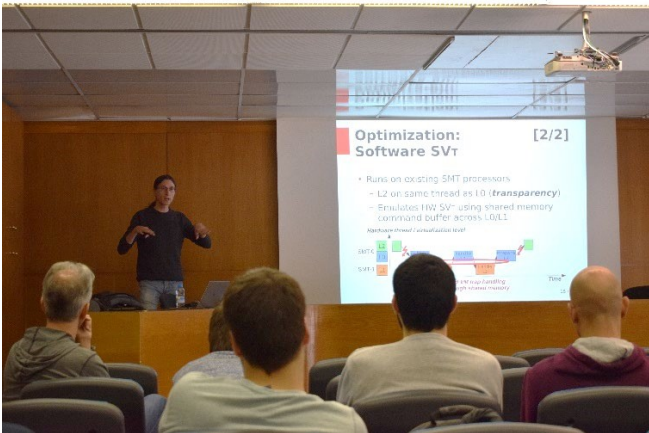
Lluís Vilanova
Technion, Israel

Abstract

Virtual machines have become a de facto standard of computing for many enterprises and, at the same time, large datacenter operators offer time on their infrastructure as a computing substrate for these virtual machines. This mix comes with a wealth of inefficiencies, and resource management

challenges that stem from the conflicting requirements of the datacenter operator and their clients. On one hand, third-parties want to solve problems that have ever-increasing memory and I/O requirements, making the impact of inefficiencies in virtualized environments ever more important. On the other hand, datacenters are largely over-provisioned to fit the requirements of their most demanding workloads, wasting a lot of processing cycles and storage bandwidth and capacity on the way. In this talk I will present some ongoing efforts in tackling all these problems. First, I will describe how

we can remove inefficiencies in memory translation and I/O in virtual machines using existing hardware, and how minimal micro- architectural changes can bring efficiency even further. Then, I will describe how we can pull some of the resource management logic present on each node of a rack into the top-of-the-rack switch to make more well-informed and fine-grained decisions to better utilize the resources that we already have available. Finally, I will show how, in the long term, we can efficiently manage an entire disaggregated rack as a single machine, shifting some of the operations typically present in a node's operating system into the network infrastructure.

Short bio

Lluís Vilanova is a post-doctoral fellow at Technion. His research interests are in operating systems, resource and device virtualization, analysis and design of hardware/software interfaces and resource management in heterogeneous and parallel systems. He has worked in a broad spectrum of topics, like efficient security-oriented hardware architectures, dynamic binary translation and instrumentation, information-secure web server infrastructures, and efficient storage virtualization, to name a few.