Towards a Predictable and Scalable Cloud

Abstract
As the vast majority of online applications and services are moving to the cloud paradigm, properties such as predictability and high-performance have become key requirements for major cloud providers. Yet, these properties have been notoriously elusive in modern massive data centers where infrastructure and computing resources are shared across applications and users. In this talk, I will provide an overview of the efforts I was involved in at Microsoft Research towards ensuring predictable and high-performant data centers, across all layers of the stack, from the physical to the application layer. I will briefly describe Microsoft’s planet-scale cloud infrastructure, and then discuss how simple abstractions can help towards predictable performance for shared cloud resources like the network and storage. Finally, I will describe our recent efforts to ensure the scalability of cloud architectures by taking advantage of advances in optical technologies.