Will Computer Systems with Performance Guarantees Ever Go Mainstream?

Juan A. Colmenares
Computer Science Laboratory (CSL)
Samsung Research America - Silicon Valley (SRA-SV)
<juan.col@samsung.com>

Keynote Speech
15th IEEE Int'l Symposium on High Assurance Systems Engineering (HASE 2014)
Miami, Florida, USA. January 10, 2014

Abstract

A key requirement for cyber-physical systems, especially those of the mission-critical type, is to offer performance guarantees (e.g., hard or high-confidence upper bounds for service times). But most computer scientists and engineers consider that those systems belong to an ultra-specialized sub-area, and they do not work in it. In fact, they rarely need to show that the systems they develop offer high degrees of performance assurance; usually performance evaluations based on average values suffice for their needs and interests. An inflection point, however, has been latent for more than a decade. Ever-increasing demands for high-quality multimedia applications (e.g., multi-party video conference and video/audio on demand) have typically motivated the need for (probabilistic) performance guarantees. More recently, Internet-based service providers, such as Google and Facebook, have started to show interest in offering predictably responsive interactive services; an obvious motivation is to try to differentiate themselves from the competition in order to retain existing users and attract new ones.

Developing a distributed computing system with performance guarantees is a difficult, costly, and time consuming task. For that reason, this task is only carried out when it is strictly necessary. Naturally, mainstream computer science and engineering community usually defer such hard problems until there is no other choice but to face them head on. A notable, recent example is parallel computing, catapulted to the masses of software developers by the offering of multi-core processors due to the diminished gains in processor performance from increasing the operating frequency.

In this keynote talk, we discuss whether or not current trends may force us to develop massively used computer systems with performance guarantees. Motivating examples include cloud-computing systems, mobile client applications, as well as applications enabled by swarms of sensors and actuators at the edge of the cloud. The talk covers the types of performance guarantees those systems will likely require as well as some techniques that can help provide those guarantees fast enough to meet market demands. We conclude with a set of key open problems and future research directions in this area.