Recent advances in load balancing: replication, speculation and auto- scaling

Jonatha Anselmi*[1]

[1]INRIA – INRIA – France

Résumé

In this talk, we will discuss modern approaches for load balancing in large-scale parallel-server systems: replication, speculation and auto-scaling.

Replication sends multiple copies of a given job, simultaneously upon its arrival, to multiple servers and then uses the results from whichever copy responds first.

Speculation sends one or multiple copies of a given job only once the system smartly detects it as a "straggler", i.e., as a job taking longer than expected to complete because of some unfortunate runtime phenomenon.

Auto-scaling allows the net service capacity, or overall number of servers, to scale up or down in response to the current load and within the same timescale of job dynamics.
We will review the state of the art and present some recent results in the flavour of mean-field and fluid limit theorems.

*Intervenant