Rapid growth of machine-to-machine (M2M) communications necessitates the reevaluation of the Long Term Evolution-Advanced (LTE-A) performance, since the current standard is not optimized for intensive M2M traffic. A serious issue is that massive M2M arrivals can overload the LTE-A random access channel, resulting in a significant access delay. There have been a number of proposals to control this overload; however, there are no studies on the mathematical characterization of delay bounds to the best of our knowledge. Here, we derive lower bounds for the LTE-A average random access delay for both a regular traffic pattern (uniformly distributed arrivals) and for a traffic pattern, indicating a serious congestion (beta-distributed arrivals). The proposed delay bounds, which predict the minimum delay with less than 6% error, present the fundamental limits of delay that can be achieved by a practical load-balancing algorithm. This paper is also one of the first attempts toward the mathematical analysis of beta-distributed arrivals. We also analyze the effect of estimation accuracy, frequency of random access opportunities, and the number of preambles on the access delay. We show that it is possible to reduce the access delay by several orders of magnitude using an appropriate configuration of these system parameters.