Serverless edge computing adopts an event-based

paradigm that provides services and dynamically provisions

resources as needed, resulting in efficient resource utilization.

To improve the end-to-end latency and revenue, service

providers need to optimize the number and placement of serverless

containers while considering the system cost incurred by the provisioning. The

particular reason for this circumstance is that frequently creating

and destroying containers not only increases the system cost but

also degrades the time responsiveness due to the cold-start process.

Function caching is an approach to mitigate the cold-start

issue. However, function caching requires extra hardware

resources and hence incurs extra system costs.

In this paper, we study the request distribution and container caching

problem in serverless edge computing.

We jointly consider the distributed and resource-constrained

nature of edge computing and propose an optimized

request distribution algorithm that adapts to the dynamics of

serverless invocations with a theoretical performance guarantee.

We propose a context-aware probabilistic caching policy that

incorporates a number of characteristics of serverless invocations.

Via simulation and implementation results, we demonstrate the

superiority of the proposed algorithm by outperforming existing

caching policies in terms of the overall system cost and cold-start

frequency by up to 62.1% and 69.1%, respectively.