Increasing heat density of servers causes proper thermal management a crucial aspect in modern computing systems. In this work, a fan management methodology aiming reduction of energy usage of the heterogeneous server system is presented. The method utilizes power and thermal models in order to predict the thermal state of the servers and thus enable runtime proactive fan speed adjustment. Thanks to the mapping of cooling zones to the particular system components it prevents from hotspot driven management. The method finds the trade-off between the cooling power of fans and temperature-dependent processors’ power leakage while ensuring to keep the temperatures of processors below the desired level. The presented approach is evaluated on real hardware system developed within the M2DC project. Corresponding power savings results are presented together with the validation of correctness and reliability under various load conditions.

**Keywords:** Fans management, Power and thermal management, Cooling optimization, Heterogeneous servers.