Selective QoS Guarantees for Heterogeneous Applications in Android Memory Management

Eunji Lee, Sunghwan Kim
Chungbuk National University
Cheongju, 362-763, Korea
http://oslab.cbnu.ac.kr

Minah Kim, Hyunkyoung Choi, Hyokyung Bahn
Ewha Womans University
Seoul, 120-750, Korea
http://home.ewha.ac.kr/~bahn

1. INTRODUCTION

With the recent advances in mobile embedded system technologies, computing paradigms are shifting from desktop PCs to smart embedded devices. Such transitions require an embedded device to provide general-purpose computing functionalities like desktops, beyond its traditional single-task real-time features. However, Android, which is most promising operating systems for embedded devices, has no awareness of supporting heterogeneous jobs, i.e., real-time and non real-time tasks [1]. As for the memory management, Android manages memory spaces in an on-demand manner that allocates memory upon a request and reclaims it if not used for a long time. This is efficient in terms of space-efficiency, but it cannot support real-time applications as memory access time is not predictable. Note that traditional real-time operating systems reserve entire memory spaces necessary for a real-time task, thereby preventing unpredictable delay due to page faults. However, this pinning policy essentially degrades the performance of non real-time jobs. Motivated by this, we present a new memory management technique that guarantees the deadline requirements of real-time application and also minimizes the performance penalty of non real-time applications.

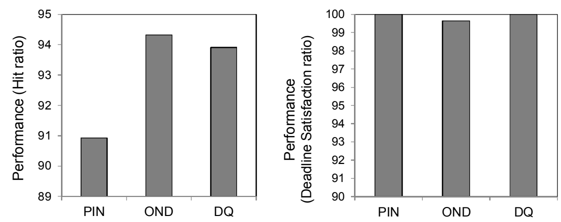
2. THE PROPOSED POLICY

Our policy predicts the minimum memory size required for a real-time application to guarantee a certain level of QoS. The prediction relies on the Belady's lifetime model which obtains the minimum memory size for achieving the given hit ratio, by monitoring past behaviors of applications [2]. Then, we ensure the memory space for real-time programs as required, while allowing remaining spaces for non real-time applications. This policy strikes a balance between supporting a real-time application and overall system performances in a judicious way.

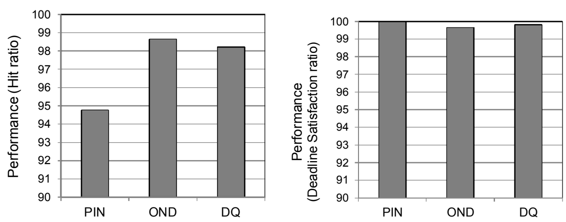
3. PERFORMANCE EVALUATIONS

Performance evaluation is performed through trace-driven simulations. Figure 1 shows the performance of the proposed policy (DQ) compared to on-demand memory allocation (OND) and pinning (PIN) policies. We assume two scenarios where heterogeneous jobs are running together, Facebook and Youtube, and Navermap and Tictoc (real-time chatting program), respectively. As shown in the figure, the proposed policy improves the performances of non real-time applications to a comparable level with an on-demand policy without nearly incurring deadline misses of real-time programs. Specifically, the performance gap between on-demand policy and the proposed policy is within only 0.4% in non real-time applications. As for the real-time applications, the proposed policy satisfies deadlines by almost same ratio with the pinning policy. This demonstrates that our memory requirement prediction is fairly accurate.

APSYS'15, July 27–28, 2015, Tokyo, Japan.



(a) Facebook (Interactive) and Youtube (Realtime)



(b) Navermap (Interactive) and Tictoc (Realtime)

Fig. 1. Performance comparisons

4. ACKNOWLEDGMENTS

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2014R1A1A3053505) and IT R&D program MKE/KEIT (No.10041608).

5. REFERENCES

[1] S. Bae, H. Song, C. Min, J. Kim, Y. I. Eom, "EIMOS: Enhancing Interactivity in Mobile Operating Systems," LNCS, Vol. 7335, pp. 238-247, 2012.

[2] E. Lee, K. Koh, H. Bahn, "Dynamic Memory Allocation for Real-time and Interactive Jobs in Mobile Devices" IEE Electronics Letters, Vol.46, No.6, pp. 401, 2010