Early-stage Thread Culling Unit for Mobile GPGPU Applications

Yu-Jung Chen Pai-Shun Ting Meng-Lin Yu Shao-Yi Chien
Media IC & System Lab, National Taiwan University

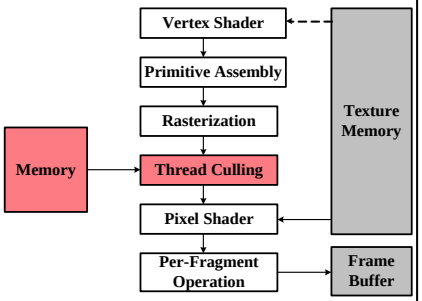
Introduction

Motivating Problem:

- Redundant or unnecessary threads execution devastates parallel computing performance in GPUs.
- High-level image and vision applications where object-based processing and *Region of Interest* (ROI) are often involved can not be effectively parallelized by GPUs.

Background:

- Early-z culling, early stencil culling and scissor test greatly enhance the rendering performance.



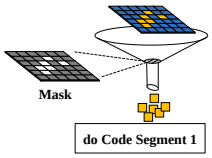
New graphics pipeline with the proposed TCU for GPGPU

Proposed Solution - TCU:

A simple yet efficient configurable early-stage thread culling unit is proposed and integrated into our mobile GPU to mitigate the execution efforts of divergent threads with identified conditions.

```
Read MaskValue from texture memory
If MaskValue is larger than threshold{
    do Code Segment 1
}
Else{
    do Code Segment 2
}
```

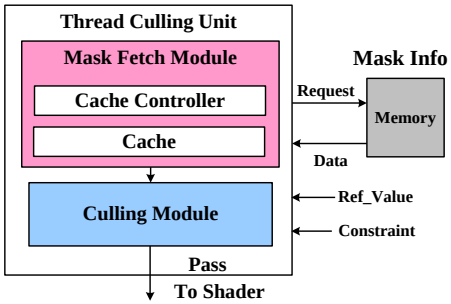
(a) Program example



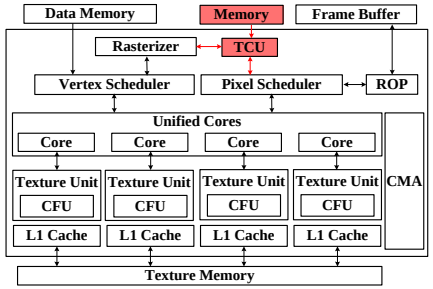
(b) With thread culling

(a) Shader program example.
(b) Illustration of thread culling.

TCU & GPU Architecture



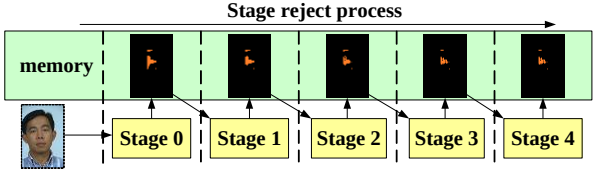
(a) TCU architecture



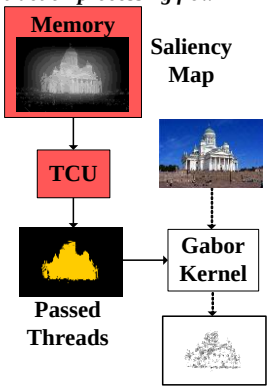
(b) GPU architecture

TCU Applications

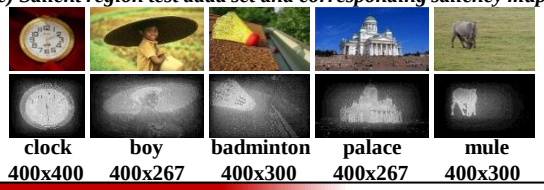
(a) Viola-Jones face detetion processing flow



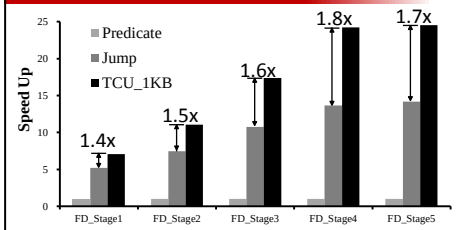
(b) Salient region Gabor feature extraction processing flow



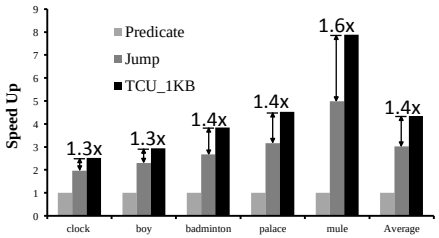
(c) Salient region test dada set and corresponding saliency map



Results



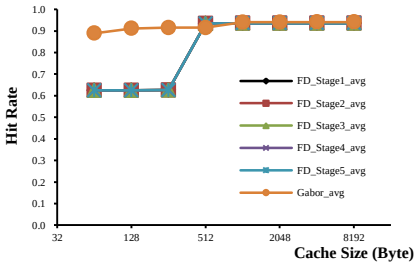
(a) Viola Jones face detection



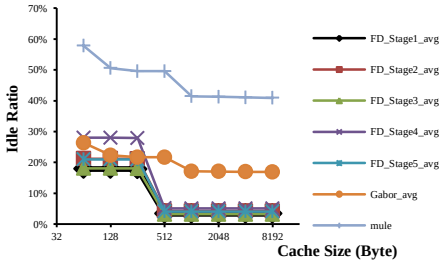
(b) Salient region Gabor feature extraction

With our TCU, GPUs can improve up to **24.5x** and **1.8x** performance for Viola-Jones face detection framework and up to **4.3x** and **1.4x** improvement in salient region linear feature extraction compared to predicate execution and branch instruction.

Cache Issues



(a) Hit rate of TCU cache



(b) Processor idle ratio

Processors are possibly stalled to wait for dispatched threads, as the throughput of the culling unit is not high enough. For the purpose of efficiency, the culling unit is employed with a cache. The cache size is evaluated by the hit rate and processor idle ratio which can indicate the activated degree of processors. To preserve both cost and performance, a 1KB cache is selected after analysis.

Acknowledgement

Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

Reference

[1] DEL BARRIO, V., GONZALEZ, C., ROCA, J., FERNANDEZ, A., AND ESPASA, R. 2006. ATTILA: A cycle-level execution-driven simulator for modern GPU architectures. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software.*

[2] HASSELGREN, J., AND AKENINE-Möller, T. 2007. PCU: the programmable culling unit. In *Proc. of SIGGRAPH '07.*

[3] PHILLIPS, P. J., MOON, H., RIZVI, S. A., AND RAUSS, P. J. 2000. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell. 22 (October), 1090–1104.*