

# D14-050

A 16-core Mobile GPU with Power-Aware Approximation Techniques

應用功率感知近似技巧之十六核心行動裝置繪圖處理器

## 隊伍名稱

實驗室421 / Lab421

## 隊長

陳宥融 臺灣大學電子工程學研究所

## 隊員

張家銘 臺灣大學  
資訊網路與多媒體研究研究所

顏禎佑 臺灣大學電機工程學系

## 作品摘要

電池功耗是行動裝置上繪圖處理器設計最關鍵的考量。利用GPU繪圖，計算大量的頂點以及像素需耗費相當可觀的電池電量。而隨著視覺品質進步，螢幕解析度提升，像素渲染運算為處理器主要功耗的成分。由於像素渲染負責產生照明效果與材質貼圖，不僅運算量大，貼圖記憶體頻寬高。在本設計中，我們提出了130mW的應用功率感知近似技巧之十六核心行動裝置繪圖處理器。

此16核心繪圖處理器由16顆4-Channel SIMD處理器為整合渲染核心Unified Shader (US)，每四顆核心組合成一個運算核心叢集(Cluster)；每個運算叢集配有一個貼圖單元(Texture Unit)和對應的L1 Cache (快取)。近似精確度處理器技術 (APS)



2014

實現於第三個運算叢集；近似貼圖(AT)則在每個貼圖單元都有實作，透過額外的緩衝區來輔助。除了材質L1快取，材質L2快取也被用來減少外部的貼圖頻寬存取。GPU繪圖管線的最後會進行ROP測試。任務配送器將渲染器核心執行緒排程，透過調度核心叢集、螢幕空間的測試單元、Rasterization引擎與ROP引擎中的螢幕空間的測試單元，實現近似頻幕空間打光技巧(SSAL)。

由於繪圖標準使用32bits浮點數運算，需要花費相當可觀的功耗。雖然渲染器程式雖然是處理浮點數，而最終渲染的像素顏色卻會被轉成0~255的整數表示。所以我們修改向量邏輯運算單元的架構來近似精確度，將原先的32 bits利用16bits近似，讓使用者難以察覺其差異。儘管暫存器能縮短成16bits表示，我們仍需保留位置與貼圖座標暫存器的精確度，因為螢幕的像素位置對於精確度相當敏感。此技術將運算叢集面積減少為原來的61%，執行像素運算時功僅需約55%。

近似貼圖近似一些貼圖中比較變化較平緩的部分，透過存取高LOD (Level of Detail) 技術，減低執行材質貼圖時造成的記憶體頻寬。使用小波轉換來偵測貼圖的複雜度，利用高LOD的材質，取代低LOD的的材質貼圖。僅用高LOD的材質貼圖可能造成細節模糊，而我們提出的近似貼圖的技巧可以透過分析貼圖複雜度來確保細節盡可能保留。

由於局部物體表面的照明效果主要為低頻變化，高頻成分主要由物體輪廓，幾何的邊界所構成。物體表面上的光影變化大致上都不會太劇烈，我們提出近似頻幕空間打光技巧(SSAL)，在螢幕空間中做像素取樣，並且內插重建近似結果。如此一來能減少處理器運算，進而降低功耗。近似的程度與投影的三角形面積成正比。

我們所提出的架構可以應用功率感知的特性在耗電量與視覺品質權衡。並採用TSMC 45 nm製程實做晶片，最高工作時脈可達350MHz。結果顯示，我們所提出的「近似精確度處理器」與「近似頻幕空間打光技巧」，最高可以減少處理器53%功率。「近似貼圖」技術能夠減少24.57% 材質L1快取的更新。



## 指導教授

簡韶逸 / 臺灣大學電子工程學研究所

臺灣大學電機工程學研究所博士。曾任廣達電腦研究中心、臺灣大學電機工程學系暨電子工程學研究所助理教授、副教授，目前為臺灣大學電機系暨電子所教授兼任臺灣大學電機系副系主任。

## 研究領域

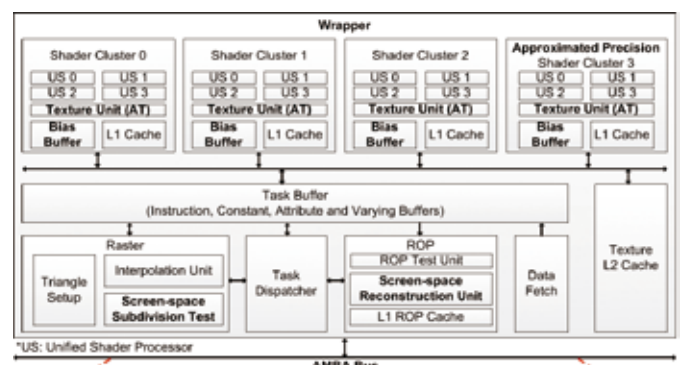
多媒體訊號處理系統、多媒體積體電路設計、晶片系統設計方法研究。

## Abstract

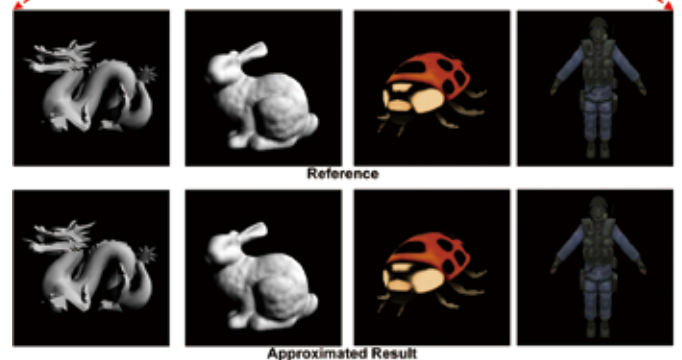
Power consumption is the most critical design consideration for mobile Graphics Processing Units (GPUs). Due to its parallel architecture for processing tons of vertices and pixels, a mobile GPU drains significant energy from battery. Intensive pixel shading dominates the power dissipation of the graphics pipeline as the screen resolution grows. Since pixel shading is responsible for producing lighting effects and object surfaces texturing, it notoriously consumes computation efforts and texture memory bandwidth. In this work, we propose a 130mW 16-core mobile GPU with three pixel approximation techniques.

The architecture of the proposed 16-core mobile GPU is mainly composed of four shader clusters. Each shader cluster contains four unified shader cores, a texture unit and a corresponding texture L1 cache. Therefore, there are sixteen concurrent active vector threads operating in parallel for either vertex program or pixel program. Each unified shader core adopts a four-channel SIMD (Single-Instruction-Multiple-Data) processor architecture. The proposed Approximated Precision Shader (APS) architecture is implemented in Shader Cluster 3. In addition, the proposed Approximated Texturing (AT) mechanism is employed in each texture unit. Besides texture L1 cache, an extra bias buffer is auxiliary for Approximated Texturing technique. A texture L2 cache is included as well for reducing external texture access bandwidth. Instruction and constant caches store vertex instructions, pixel instructions and corresponding constant data. Pixel varying buffers reside the rasterized pixel data. The data fetch unit is responsible for filling the instruction, constant and vertex attribute caches from accessing external memory. Raster engine includes a triangle setup unit for computing coefficients of the plane equation, and interpolation unit can interpolate pixel data on screen space. To reach the maximum pixel throughput for our design, the interpolation unit can generate sixteen sets of pixel varyings for shader cores at the same time. At the end of the graphics pipeline, the ROP test unit takes charge of the ROP tests. Task dispatcher schedules threads among shader cores, and it realizes the proposed Screen-space Approximated Lighting (SSAL) technique by coordinating these shader clusters, the

screen-space subdivision test unit in the raster engine and the screen-space reconstruction unit in the ROP engine.



### Proposed 16-core Power-aware Mobile GPU



Average PSNR can reach to 36 dB ~ 42 dB  
Maximum processor power consumption can reduce 50% !