ARM Mobile GPU Compute Accelerates UX Differentiation

GPU Compute Enables Engaging Mobile User Experiences (UX)

Executive Summary

Users continue to demand more from their mobile devices and many mobile device designers are using Mali-T604-based SoC products today to meet those ever-increasing demands. Additionally, designers are starting to enable and enhance mobile device user experiences through GPU compute. OEMs and software vendors are investing to accelerate image processing, computational photography, game physics, and video processing for both internal and external high-resolution displays.

ARM's Mali-T600 series of graphics processing units (GPUs) are ARM's first GPU compute cores designed for license by a wide range of system-on-chip (SoC) manufacturers. Mali-T600 series GPUs enable software developers through mainstream OpenCL and Android Renderscript compute software development environments. Additionally, Mali-T600 series GPUs are the first OpenCL 1.1 Full Profile conformant mobile class IP licensable GPUs shipping in widely available consumer products.

The Mali-T622 GPU is the latest addition to ARM's second-generation Mali-T600 Series, which has the same performance using half the power consumption of ARM's first-generation Mali-T604. The Mali-T622 GPU broadens the product range of the Mali-T600 Series from performance tablets down to midrange smartphones. Software developers can use existing Mali-T604-based products to develop new user experience applications for all second-generation Mali-T600 Series GPUs.

ARM's Mali-T600 series GPUs will enable mobile device designers to offer higher value, differentiated user experiences within a balanced power budget. The first generation of target user experiences include:

- smoother and seamless entertainment,
- real-time audio and video communications and recording,
- better security, voice and gesture control, and
- augmented reality.

This balance of compute performance and low power consumption will enable superior performance over prior generations of mobile devices without generating a lot of heat and without reducing battery life.

Hardware Enables New Mobile User Experiences

Over the last few years, mobile computing has enabled new user experiences while voice and text have become commoditized. Social media incorporates location-based services; video communications happen via wireless broadband services; both physical and online retail interactions are impacted – from in-store product research to finding a restaurant to customer loyalty and rewards programs. Mobile processor SoCs are becoming more powerful and incorporating more features directed at enabling new user experiences and usage models. They enable smartphones or tablets to displace single-function consumer technologies, such as cameras, music players, handheld video games, maps, TVs, book readers, etc.

Despite all these usage requirements, mobile devices have a fixed power budget. That budget is determined by the design team's choice of battery and sensors to meet their user experience goals. In many cases, designers are increasing battery power, but they are not increasing the power budget for their compute SoC. The additional power is going to higher resolution displays, better touch interfaces, more and higher quality cameras, better microphones, and a host of other sensors, including GPS/GLONASS, accelerometers, magnetometers, proximity detectors, and new radios and radio capabilities (NFC, WiFi Direct, Bluetooth, and others).

The Balance of Power

Mobile GPUs designed to offload compute cycles from CPUs seek balance between system ***power consumption*** and useful system ***compute cycles*** within a fixed bill of materials budget and a fixed power envelope. The goal is to enable multiple power vs. performance profiles. Designers generally strive to:

- *Maximize battery life* – extract maximum performance from a system at a fractional power budget, perhaps by shutting some CPU cores and some GPU cores but balancing app performance demand between them for the best experience.
- *Maximize performance* – tune for the best performance possible for a given app while burning the maximum amount of power allowed for fastest battery drain.
- *Balance* – optimize user experience while extending battery life as much as possible: a dynamic mix of the above, dependent on what apps are being used and user preferences for power saving modes.

One of the primary enablers for power-efficient GPU compute is shared memory architecture – CPU and GPU cores share a common pool of memory, so that they can pass pointers to memory locations instead of explicitly moving data and code from CPU memory space to offload compute memory space (and vice versa). ARM is a founder of the HSA (Heterogeneous Systems Architecture) Foundation. HSA is defining vendor and offload technology neutral standards for parallel programming, including shared memory architecture features and functionality.

Mali GPU Compute Today

ARM's Mali-T604 GPU compute capability has already been deployed by system vendors like Samsung and will be used to accelerate mobile image processing and photography applications. These applications will improve the quality of mobile users' photography experience by using software and GPU-based parallel computing to compensate for the limitations of mobile camera sensors and physical constraints on lenses and their placement.

Interestingly, clamshell and notebook designs mount webcams in the display bezel and face the same challenges. In the PC market, OEMs long ago shifted high quality video to external webcams to keep costs down, but smartphone and tablet designers do not have the luxury of adding large external peripherals to their devices' user experience. As mobile designers pioneer higher quality audio and video communications experiences, the clamshell and notebook markets will follow.

Smartphone vendors and game developers will be using GPU compute to improve the realism of mobile gaming by running game physics on the GPU to increase the realism of game play within the existing smartphone power budget.

In addition, GPU compute can be used to dynamically resize the playback resolution of web-based video to improve the quality of social media on both the built-in mobile screen and on external high-resolution displays.

Zoom-In on Camera Lenses

Mobile device designers must contend with many constraints. Smartphones and tablets are physically defined by very thin enclosures. That constrains the distance between their camera sensor and the lens structure mounted over it, unlike most single-function cameras. The sensor (a semiconductor-based photo-detector chip) is attached to the smartphone motherboard. The sensor is typically an integral part of a lens assembly. Typically there is only enough separation distance for one lens, as the lens assembly becomes part of the exterior shell and may determine the total device thickness.

There are two aspects to improving users' smartphone and tablet photographic experience: **image processing** to improve the quality of a photograph before it has been written to a file, and **computational photography** to artistically change a photograph after it has been recorded.

Image processing can incorporate color correction (artifact of both inexpensive sensor and lens), lens aberration correction (distance from sensor, quality of lens), distortion compensation (helpful if the lens is inexpensive and/or dirty), noise filtering (especially in low light environments), and a host of other quality enhancement techniques.

GPUs elegantly handle image processing by offloading the cycles from the main processor and in some cases the image signal processor (ISP).

Application-Specific Acceleration

The mobile platform design space is not a simple tension between CPU and GPU cores. There are a range of technologies available to SoC designers. These

technologies enable designers to fine-tune offloading and accelerating many kinds of processing from the CPU cores, while at the same time balancing compute acceleration against power consumption.

GPU Compute enables an efficient combination of faster execution times and low power for highly (often referred to as “embarrassingly”) parallel applications. For mobile systems, applications that generate and process pixels are good candidates for acceleration, such as image processing, computational photography, and general applications support for higher resolution displays. So is increasing the realism in simulating 3D spaces by accelerating game physics.

The spectrum of compute offload options:

Logic	Parallelism	Pros	Cons
CPU core	Low, via multi-threading and multiple cores	Completely general purpose; widely available software development tools	Cost and power draw depends on size of core, manufacturing tech, and how many are used; instruction set extensions can accelerate, but at die size and power consumption cost
Programmable	Moderate	Programmable; change functions on-the-fly; deploy new functions quickly	Large area; expensive; higher power consumption; 3rd parties typically not enabled to reprogram
Custom	Potentially Highest	Smallest area; inexpensive; low power; 3rd party sources for high value functions	Fixed functions locked years in advance; function updates are long lead time
DSP core	Moderate	Programmable; smaller, faster, lower power than CPU cores for some tasks	Most are not exposed to 3rd party developers, as they are dedicated cores wrapped in custom logic for specific tasks
GPU core	High	Programmable; very high density and parallelism with low power per work done	Very expensive to develop competitive in-house intellectual property; few licensable sources

Examples of each type of SoC integrated logic:

Logic	Examples	Licensable
CPU core	ARM and its Architecture Licensees, Intel and AMD x86, Imagination Technologies MIPS, Cadence Tensilica, Synopsys ARC, and others	ARM, MIPS, Tensilica, ARC
Programmable	ARM AMBA embeddable FPGA interface	Yes
Custom	ARM AMBA based audio and video codecs and transcoders	Yes
DSP core	Qualcomm Hexagon, Texas Instruments Keystone	No
GPU core	ARM Mali-T600 series, NVIDIA Kepler, AMD GCN, Vivante ScalarMorphic, Imagination PowerVR	Mali, Kepler, PowerVR, ScalarMorphic

There are a few high-level rules for choosing compute offload technology:

- Where an application must adhere to a standard communications or file protocol, then prototype with FPGAs and then spend time and resources to commit optimized designs to custom silicon.
- Where sensor and user experience technologies are changing rapidly and there are not yet standards, prototype in software on CPU cores and then spend time and resources to commit to GPU or DSP software acceleration.
- When a de facto standard emerges as vendors converge on a technology, evaluate shifting from software acceleration to FPGAs and custom logic.

ARM Mali GPUs

ARM's Mali-T600 Series GPUs are ARM's first compute offload cores, and as such are designed for license by a wide range of SoC manufacturers. Because ARM licenses intellectual property to SoC design companies and does not manufacture or sell its own SoCs, ARM's business model and market timing are different than most of the traditional computing SoC supply chain.

For instance, Mali-T604 graphics IP was announced two and a half years ago. The first customer SoC incorporating that IP shipped six months ago (Samsung's Exynos 5250), almost two years after its introduction. Based on that Samsung SoC and Samsung production consumer devices built around it, ARM submitted its Mali-T604 graphics for OpenCL 1.1 Full Profile conformance with Khronos and was certified "conformant" in August of 2012. ARM announced the availability of much of their second generation of Mali product IP –Mali-T624, Mali-T628 and Mali-T678 – just as Samsung shipped the first instance of Mali-T604 to market.

Samsung shipped their Exynos 5250 SoC in a couple of end-use products over the last six months, including their own Series 5 550 Chromebook, and Google's Nexus 10 tablet. Samsung's BD-F7500 Blu-Ray Player and F8000 Smart LED TV both use other Samsung SoCs based on ARM's Cortex-A15 core and Mali-T604 GPU.

The primary differentiation between the Mali-T604 GPU and the second-generation Mali-T600 Series GPUs is that the second-generation maintains Mali-T604 performance while consuming half the power. Performance scales with the number of cores and pipelines in the GPU – at the high-end, Mali-T678 is tuned for the demand of tablets, while the newest second-generation Mali-T600 Series GPU, the Mali-T622, is tuned to bring GPU compute down from tablets and high-end smartphones to more affordable midrange smartphones.

ARM's Mali-T600 series cores all comply natively with IEEE 754-2008 64-bit floating point standard:

Mali	Shader Cores		ALUs per Shader	Max ALUs
	Min	Max		
T604	1	4	2	8
T622	1	2	2	4
T624	1	4	2	8
T628	1	8	2	16
T678	1	8	4	32

In addition to OpenCL, Mali-T600 series GPUs already support Google's Android Renderscript, and will receive Microsoft DirectCompute support as DX9, DX10, and DX11 drivers come online in 2013 – when Microsoft includes Mali support in their Windows RT 8.1 distribution. Low cost Exynos-based development platforms such as the InSignal Arndale are publically available from third party suppliers.

Mali-T600 Series software development support:

	Publisher	Availability
Development Target		
OpenCL 1.1	Khronos	Now
Android Renderscript Compute	Google	Now
DirectCompute DX9, DX10, DX11	Microsoft	2013
Development Environment		
DS-5 and Streamline Performance Analyzer	ARM	Now
Android Developer Tools with Java/Dalvik	Google	Now
Native Development Kit and Java Native Interface (JNI)	Google	Now
GCC 4.4.3	GNU	Now

GPU Compute in the Future

Current GPU compute usage models – image processing, computational photography, improvements in web video, and improving mobile gaming experiences – will continue to see R&D investment, and they will show substantial improvements between mobile product generations. They are nowhere near their upper performance limits in terms of algorithm choices, performance improvements, and reduced power consumption.

To start, higher resolution mobile displays and "4K resolution" content formats and panel displays are being introduced to the market now. Within the next few years they will become commonplace. We can no longer assume that web-based video is always higher resolution than an average user's mobile display. Up-scaling web-based video content for higher resolution mobile displays is as important as downscaling high resolution content for smaller, lower resolution displays.

As higher resolution video file formats emerge, mobile cameras will be able to capture motion video at higher quality. Video and audio pre- and post-processing for recording and local viewing will consume increasing amounts of GPU compute processing power.

Media compression will also take advantage of increasing GPU compute. Services and device vendors will be able to offer much better quality within an existing bitstream bandwidth ceiling. Video conferencing quality will improve through both much better quality at full motion and through reducing transmission bandwidth while still seeing improved quality over today's video call experience.

As media compression, media file formats, and media apps and services evolve, vendors will want to deploy new algorithms with GPU compute as soon as they are developed. And they will want to deploy well in advance of standards and specialized custom silicon.

Looking past extending current uses for media processing, computer vision will enable a host of new image recognition-based usage models, including gesture control, security, and augmented reality applications. These applications span a wide range of consumer market segments outside of mobile computing, including the broader consumer electronics market, digital TV, entertainment, and security markets. The next generation of mobile SoCs will use both application processors and GPU compute to enable new consumer user experiences across mobile and home consumer electronics.

Call to Action

GPU compute is an emerging set of technologies that can accelerate the performance of high-value consumer usage models for hardware designers, software developers, and service providers.

AMD is now the only GPU compute IP source that has not announced plans to license their GPU cores to third parties. ARM's advantage is that they are a one-stop-shop for all of the critical compute IP required to build an applications processor SoC that includes GPU compute.

ARM Mali-T600 series graphics enables licensees to design competitive GPU compute capability into their products without the need to become an expert in every skill set required to bring a complete solution to market, specifically: heterogeneous compute architecture, software infrastructure design, applications design, and parallel programming.

Subject Matter Experts Interviewed

- Jem Davies, ARM, Fellow and VP of Technology, Media Processing Division, and member of HSA Foundation Board of Directors
- Peter Hutton, ARM, GM & EVP Media Processing Division
- Roberto Mijat, ARM, Visual Computing and GPU Computing Marketing Manager
- Mikaël Bourges-Sévenier, Aptina, Director High-Performance Imaging
- Andrew Richards, Codeplay, CEO & Founder
- Jin-Aeon Lee, Samsung, VP Multimedia

Important Information About This Paper

Author
Paul Teich, Senior Analyst at Moor Insights & Strategy

Editor
Patrick Moorhead, President & Principal Analyst at Moor Insights & Strategy

Inquiries
Please contact us here if you would like to discuss this report and Moor Insights & Strategy will promptly respond.

Citations
This note or paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

Licensing

Disclosures
Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies, including ARM, who commissioned this paper. No employees at the firm hold any equity positions with any companies cited in this documented.

Disclaimer
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness or adequacy of such information and shall have no liability for errors, omissions or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the

results of any revision to these forecasts and forward-looking statements in light of new information or future events.