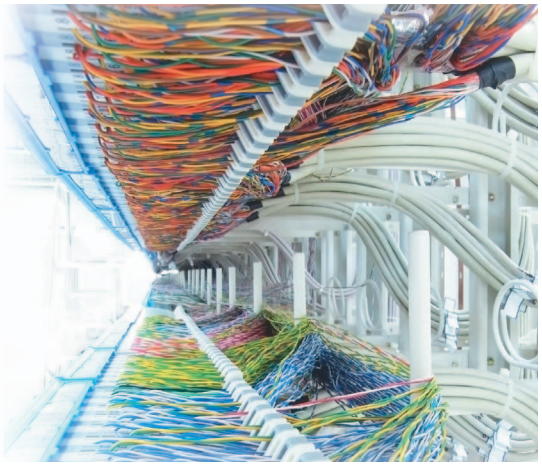

GPUs Go Mobile

Lee Garber

Mobile GPUs will need substantial improvements to enable wireless devices to perform the complex graphics-related functions that many manufacturers, developers, and users want.

As mobile devices become increasingly popular, users and manufacturers are calling on them to do more tasks, many of which involve graphics. For example, people are using smartphones to play games, view complex video and images, and even edit photos. Industry observers say activities such as mobile augmented reality could become common in the not-too-distant future.

All of this has brought increased attention to the mobile GPU.

There are numerous mobile GPUs on the market. And, said Eric Demers, telecommunications vendor Qualcomm's vice president of engineering, "The demand is growing rapidly."

However, noted Dave Shreiner, director of graphics and GPU computing for chip designer ARM, the processors and supporting technologies are still relatively immature and thus will not be able to improve without many changes.

These changes could include greater performance, reduced energy consumption, better APIs for software developers, and more efficient use of memory. As these technologies change, mobile devices could become much more functional than they are now.

However, there is still considerable debate within the industry of how best to make these changes and whether all of them are good ideas. The results of these discussions will profoundly affect the future of mobile technology.

THE STATE OF THE MOBILE GPU

"Mobile graphics are currently at the forefront of consumer computing," said Peter McGuiness, director of business development for chip designer Imagination Technologies. "This decade has seen mobile devices like smartphones and tablets becoming the key drivers for the exponential growth in worldwide computing device shipments, surpassing PCs."

Now, smartphone users want the same type of performance, image quality, and functionality that they get from PCs. Providing this will require considerable innovation because handsets have fewer resources—particularly in terms of processing power, memory, energy supply, and real estate—to work with than desktop machines.

Thus, wireless technology is the focus of considerable innovation. Much of the innovation is focused on the GPU, which the "A GPU Primer" sidebar discusses in detail.

WANTING MORE FROM THE GPU

Chip designers, application developers, and mobile device users want to see changes that could improve GPUs. For example, there is a growing demand for more powerful mobile devices that provide richer, higher-quality graphics. This could require GPUs that offer more performance.

According to Tim Leland, senior director of product management for Qualcomm subsidiary Qualcomm Technologies, "Consumers want user interfaces on their mobile devices that are easy to use, look terrific, and respond very quickly to their touch. [The] GPU is responsible for a substantial amount of the underlying processing that

Published by the IEEE Computer Society

makes this experience enjoyable on some of the most popular smartphones and tablet devices."

For developers, said Renaldas Zioma, phone and handhelds shepherd for Unity Technologies—which makes Unity, a platform for developing interactive 3D and 2D content—"GPUs' power and architecture directly affect which tools and solutions we can provide. They are the driving force behind the ability to push cutting-edge visuals to the screen. The better the GPU, the higher quality the visuals."

Performance and power consumption

"There will be a demand to increase the performance of GPUs for new products until GPU-rendered graphics are as photorealistic and smooth as the images that come from an HD video recorder," said Qualcomm's Leland. "We have a long way to go until that happens."

"Mobile devices have to be very careful to deliver computation performance using the minimal possible power, not just to extend battery life but to enable applications to run inside the thermal design point, the amount of power that can be consumed by a mobile device without it becoming too hot to operate or handle," said Neil Trevett, GPU maker Nvidia's vice president for mobile content and president of the Khronos Group industry consortium.

"GPUs are more efficient when it comes to power consumption than general CPUs at the same tasks," said Unity's Zioma. "However, processing complicated geometry, calculating realistic lighting, and drawing and shading millions of pixels 30 times per second are very computation-hungry processes that in turn require a lot of energy."

A key to this will be developing more efficient algorithms for GPUs to run, explained ARM's Shreiner.

The smart use of transistors, including the ability to turn off those not needed for a task, will also be important because fans and other types of active cooling approaches aren't practical for chips in mobile devices. Imagination's McGuiness noted that his company already provides chip designs with power islands that can be turned on or off dynamically to optimize power consumption.

Developers are also using approaches such as texture compression on images so that chips can use less bandwidth processing them.

Unified memory

Memory usage is a key to making GPUs operate more efficiently. The more cycles and energy that GPUs must expend to get data from memory, the less efficiently they operate.

Unlike PCs, mobile systems on chip (SoCs) have the CPU and GPU

A GPU PRIMER

GPUs are specialized, programmable processors that efficiently manipulate computer graphics, conducting calculations in a highly parallel manner, noted Peter McGuiness, director of business development for chip designer Imagination Technologies.

They are also good for nongraphics-related calculations that can be performed in parallel, in which case the chips are considered general-purpose GPUs (GPGPUs).

GPUs rapidly manipulate memory to accelerate the building of images in a frame buffer for subsequent output to a display. In the process, they offload work that a system's CPU would otherwise have to do.

GPUs perform tasks such as texture mapping, pixel shading, polygon rendering, color support, and MPEG decoding. They also create lighting effects, show motion within a scene, and transform objects every time a scene is redrawn.

"GPUs have gained critical momentum in mobile computing as their per-core multithreaded capabilities far exceed what a traditional CPU can offer," added McGuiness.

All mobile devices and tablets, as well as some laptops, netbooks, TV set-top boxes, and handheld game consoles use mobile GPUs, said Renaldas Zioma, phone and handhelds shepherd for Unity Technologies, which makes Unity, a platform for developing interactive 3D and 2D content.

These rapidly growing markets have generated demand and interest from consumers, developers, and manufacturers for mobile GPUs, he said.

A brief history

GPUs have been around since the early 1980s. The initial versions were essentially integrated frame buffers, which drove a video display from a memory buffer containing a complete data frame. The Commodore Amiga, released in 1985, was one of the first PCs to include a GPU.

However, even into the late 1990s, many computers still used either traditional CPUs or CPUs with on-chip graphics capabilities to handle graphics processing.

Nvidia, one of the early leading GPU vendors, coined the term "graphics processing unit" in 1999. Over time, the chips have increased in processing power and functionality.

Only within the past few years have companies begun making GPUs that could function well within mobile device's resource restrictions, allowing vendors to regularly include them in their products.

Today's major mobile GPU makers and designers include AMD, ARM, Imagination Technologies, Intel, Nvidia, Qualcomm, Samsung, and Texas Instruments.

Applications

GPUs are primarily for applications such as graphics in mobile devices, games, user interfaces, and browsers, said Tim Leland, Qualcomm subsidiary Qualcomm Technologies' senior director of product management.

In addition, said Imagination's McGuiness, mobile GPUs could also help with activities such as image processing and computer vision.

Video processing, various types of visualization, geographic information systems, and architectural drawing would also benefit, noted Daniel Wexler, founder and a corporate officer of the 11ers, an application development firm.

on the same die and they share the same memory subsystem, said Nvidia's Trevett. In PCs, he said, the CPU is separated from the GPU, and the processors and memory communicate via slow busses.

"Mobile GPUs have the opportunity to leapfrog the problems that held back the range of their usage on the desktop by providing direct support for unified memory between the CPU and GPU," said Daniel Wexler, founder and a corporate officer of the 11ers, an application development firm. "Memory transfer is the single biggest performance bottleneck."

Explained Kevin Krewell, senior analyst with the Linley Group consultancy, "By sharing the memory-management structure, the two units can pass tasks and pointers without a lot of software overhead." However, noted Unity's Zioma, "Current mobile APIs do not allow programmers to leverage such architecture to the full extent."

The GPU programming model still uses the legacy approach from PCs, which doesn't incorporate unified memory, stated Wexler.

APIs

"As graphics operations steadily became more complex to generate increased visual realism, GPU architectures became increasingly programmable," said Nvidia's Trevett. "Graphics and compute APIs have evolved to expose and catalyze the increasing capability of GPUs."

The principal graphics APIs include the Open Graphics Library (OpenGL, primarily for PCs and workstations), OpenGL-ES (for embedded systems), WebGL (a JavaScript API for interactive graphics), and OpenCL (for heterogeneous processor platforms), all managed by the Khronos Group, and Microsoft's DirectX (for Windows-based platforms).

APIs for general-purpose GPUs (GPGPUs) include Google's RenderScript compute and Microsoft's DirectCompute.

Developers, who must design versions of their applications for different APIs, say it is increasingly difficult to do so with so many APIs. They are thus hoping that the number can be reduced.

Typically, graphics APIs for resource-constrained mobile devices have fewer features than APIs for PCs. Qualcomm's Demers said there is a need for full-featured APIs so that developers can build in more complex and richer graphics functionality. However, running such APIs on a mobile device often comes with a performance penalty.

According to ARM's Shreiner, developing full-featured APIs for wireless devices would be complex and might be overkill for the mobile platform.

Distributed rendering

"With 3D gaming," said Shreiner, "people want a console experience." However, he noted, the GPUs in phones don't offer a console's performance in rendering images. Instead, he suggested, a more powerful cloud-based computer could render the images people want to see in a game and send them across the network to appear on a mobile device's screen.

In fact, said the Linley Group's Krewell, numerous companies are already developing cloud-based game servers to work with mobile devices. "The difficulty with cloud gaming is that the server-to-client delay must be kept low enough so that gamers perceive little or no lag between the time they make a move and the time the resulting image appears on the display," he noted.

"I'm not a big supporter of distributed rendering," said the 11ers' Wexler. "Until network performance and accessibility improve by two to three orders of magnitude, I don't believe distributed rendering is the solution."

CHALLENGES

Making the improvements that users, developers, and chipmakers want to see in mobile GPUs presents several challenges.

According to Qualcomm's Leland, "Delivering leading performance over a very wide range of product tiers for each of the major operating systems and for increasingly complex mobile graphics and compute applications that run concurrently with each other, within the stringent power consumption and cost constraints for mobile devices, is becoming more [difficult] each year."

Another issue is making GPUs' processing capabilities easily accessible to programmers within mobile OS frameworks that are not always designed for high performance, added Nvidia's Trevett. "An additional challenge will be memory bandwidth for higher-performing GPUs," said the Linley Group's Krewell.

"Legacy is the enemy," stated the 11ers' Wexler. "Millions of man-hours have been poured into the development of the current programming models and chip architectures. Major changes, like adding unified memory, will undercut these investments. Programmers used to working with the legacy model need to learn the new unified model. Chip architects with existing designs need to modify those designs. Major change is hard, but you have to start sometime."

PROCESSING THE GPU'S FUTURE

"As mobile GPUs get faster and more capable, and are able to run desktop graphics and APIs, their uses will expand beyond just processing gaming and UI graphics," said Nvidia's Trevett. "For example, GPUs will be used to process the input from the camera and other sensors on mobile devices. Vision processing

will enable tracking of features in the surroundings to enable new user interfaces such as gesture processing and emerging applications such as augmented reality."

According to Trevett, "Mobile GPUs will be used to provide parallel-computation capabilities on devices that will soon provide the power of a supercomputer in the palm of your hand."

Leland predicted that better GPU performance "will obviate the need for the cloud for some types of user experiences, such as console-quality gaming on mobile devices." He added, "APIs will continue to evolve to make it easier for application developers to program GPUs to process more tasks concurrently and more effectively with less power consumption. GPU APIs will continue to evolve well into the next decade and possibly longer than that."

Stated Trevett, "The next generation of mobile GPUs that will begin to appear over the next one to two years will have the full functionality of the current desktop APIs."

In the future, said Krewell, GPUs will continue to be implemented more often as part of SoCs, rather than as freestanding chips.

"This trend will continue for many years to come, agreed Leland. "A GPU is most useful when it is tightly integrated on the same chipset, with the other necessary components of mobile platforms."

ARM's Shreiner said, "There's a lot of potential for the mobile GPU. It's in the beginning of its performance curve."

However, noted the 11ers' Wexler, "It will take architectural and process breakthroughs to enable higher levels of performance. Fortunately, the vendor roadmaps indicate huge performance improvements are coming to mobile GPUs over the next few years."

"Mobile GPUs are already very important products with hundreds of millions of devices shipped every year," said Unity's Zioma. "In a couple of years, non-mobile GPUs will become a niche product powering only hardcore gamer rigs and computers in CG movie production houses."

"Consumers demand the most visually immersive experiences and longest possible battery life from their mobile devices," Leland said. "Well-designed mobile GPUs help customers and partners meet consumer demands."

Lee Garber is the IEEE Computer Society's senior news editor. Contact him at lgarber@computer.org.