**White Paper | AMD POWERTUNE TECHNOLOGY**

### Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal Design Power and Performance Constraints on Modern GPUs</td>
<td>2</td>
</tr>
<tr>
<td>AMD Powertune Technology - Intelligent Power Monitoring for Higher Performance</td>
<td>4</td>
</tr>
<tr>
<td>The Dynamic Nature of AMD Powertune Technology</td>
<td>5</td>
</tr>
<tr>
<td>Summary</td>
<td>7</td>
</tr>
<tr>
<td>AMD ZeroCore Power - Enabling the World's Most Power Efficient GPUs</td>
<td>8</td>
</tr>
<tr>
<td>Introduction</td>
<td>8</td>
</tr>
<tr>
<td>Background</td>
<td>8</td>
</tr>
<tr>
<td>Scalable Energy Efficiency with AMD CrossFire™ Technology</td>
<td>9</td>
</tr>
<tr>
<td>Summary</td>
<td>10</td>
</tr>
</tbody>
</table>

March 23, 2012
Today’s modern GPUs incorporate highly advanced mechanisms for power management during active workloads. For example, if parts of the graphics engine are not fully stressed under a particular rendering or compute workload, the GPU will work to reduce power in that portion of the graphics engine through clock, or power gating, techniques. Over the course of a full workload, this leads to varying levels of instantaneous activity for the GPU. In some cases, the GPU will be very heavily loaded with little opportunity to clock or power gate; while in other cases, components of the GPU may be waiting on data from the CPU, framebuffer, or some other bottleneck, and use the latent time as an opportunity to manage the power down to lower levels to enable lower average power levels under load.

As a result of this GPU power management under active workloads, it can be demonstrated that all applications tend to have their own unique power ‘signature’ based on how a particular application stresses the graphics architecture and how much opportunity the GPU has to reduce power. While these applications tend to run in the highest power state (defined by engine core voltage and frequency) available to the GPU, they exhibit a fairly large spread in terms of the actual power consumed in the GPU. Figure 1 highlights the measured spread in load power for a wide range of applications running on a 225W discrete GPU.

![Figure 1: Max Total ASIC Power (W) without AMD PowerTune technology](image-url)
Measurements on modern GPUs also show that there is a relatively small subset of peak applications (referred to sometimes as “power viruses”), which tend to consume significantly higher power when compared to most other applications. GPUs must accommodate these peak applications in their design while still delivering meaningful performance on typical applications (which consume significantly less dynamic power).

The need to accommodate higher power applications has traditionally led to a compromise in performance. Any applications which result in long-run excursions above the GPU Thermal Design Power (TDP) will trigger a “thermal event”. Thermal events arise when the thermal sensor on the GPU exceeds a maximum pre-set value which forces the GPU to take immediate action to greatly reduce voltage and frequency in an attempt to keep the GPU within its operating temperature. Clearly a thermal event is not desirable as it results in much lower overall GPU performance and limits the opportunity for the GPU to move back into a higher performance band. The established design compromise on GPUs is to have a high degree of design margin – in the form of lower clock frequencies – to ensure that high power performance sensitive applications do not trigger a thermal event. This serves to generally avoid thermal events on most applications, but does so at the expense of lower overall performance across all applications.

As a result of this compromise, typical applications that consume significantly less power are not able to use the thermal headroom of the GPU to maximize their performance within the GPU TDP. Without an intelligent mechanism to adaptively manage clocks in response to active power during workloads, the GPU loses a very considerable performance opportunity as shown in Figure 2.
AMD Powertune technology (“PowerTune”) addresses this TDP Power/Performance compromise by introducing two important capabilities to GPUs power management:

1. The ability for the GPU to dynamically calculate its runtime power based on workload activity; and
2. The intelligence to control engine clocks based on the power calculations

PowerTune dynamically manages the engine clock speeds based on calculations which determine the proximity of the GPU to its TDP limit. The ability of PowerTune to calculate how close it is to the TDP delivers significantly higher performance for power constrained applications. PowerTune is very different when compared to existing discrete GPU power management policies. Rather than compromising maximum clock frequency to settings based on high power applications and TDP, the GPU can be enabled with much higher maximum clock frequencies which can be adjusted in real time to ensure that the GPU is contained to the TDP envelope with all applications it may encounter. As outlined in Figure 3, the maximum clock frequency in a GPU with PowerTune is significantly higher while the containment control mechanism is very fine grained compared to the traditional method of thermal throttling to much lower intermediate power states.
The end result is higher performance across the board for both typical and higher power applications. Typical applications with thermal headroom enjoy increased performance, in some cases significantly more performance, since these applications can run at the raised clock speeds. High power applications also enjoy higher overall performance. While PowerTune clock control may incrementally lower the engine clock during some intervals of the high power application to keep the GPU safely within its TDP limits, this is still much preferred to the legacy approach of relying on thermal triggers to force the GPU into a much lower overall performance state for longer time periods. The fine-grain and incremental nature of PowerTune’s clock control works to keep the engine clocks at the highest clock available within the TDP limit and allows the GPU to dynamically move up to higher clock rates when thermal headroom exists in subsequent power measurement intervals. Figure 4 demonstrates PowerTune’s ability to enable higher clocks with leverage the thermal headroom of the GPU to enable higher performance, while at the same time intelligently managing clocks for better performance with peak apps.

The Dynamic Nature of AMD PowerTune Technology

Some high power applications consume power that is above TDP levels for a small percentage of their total runtime. PowerTune dynamically assess GPU power at frequent sampling intervals. For thermal stability, power history per sampling interval is analyzed to ensure that power levels have not been sustained above the allowed TDP level. In addition, if power exceeds a higher threshold level in a sampling interval, PowerTune takes immediate action. This allows PowerTune to assess power for both short and long time intervals to deliver two different benefits. The short PowerTune interval is used to manage any atypical power excursions which could jeopardize the electrical design specifications of the GPU such as the power supply limitations of the voltage regulators. Any excursions which jeopardize the electrical design limitations of the GPU must be dealt with immediately to avoid failures.
From a thermal design standpoint, a GPU can safely operate above its rated TDP for relatively short periods of time. However, if the GPU exceeds TDP for too long, a thermally event will throttle the GPU to a much lower performance state. The goal of an effective active power management policy is to avoid such throttling. Traditional GPUs without PowerTune adopt an active power management policy of lower peak clocks to avoid throttling. PowerTune allows the GPU to exceed its TDP for short intervals (typically on the order of milliseconds). This has the benefit of fully maintaining the maximum clock frequencies without performance impact. If the application's dynamic profile is such that it exceeds TDP for a longer period of time (on the order of tens or hundreds of milliseconds), PowerTune takes corrective action to manage the clocks incrementally to avoid a thermally triggered event.

PowerTune is also highly granular in terms of its ability to manage clocks. While previous GPUs had only 3 or 4 power states (idle/low, medium, and peak), a GPU with PowerTune contains hundreds of intermediate states in between the primary power states to maximize performance within the TDP constraint as outlined above in Figure 4. Since the temporal measurement interval is also very small, the PowerTune algorithm keeps the GPU at the maximum allowed clock at every opportunity. The maximum allowed clock is reassessed at every interval.

The dynamic nature of PowerTune is highlighted in Figure 5. Without PowerTune, we see the GPU in Figure 5 exhibit a large spread of application power. The peak power application without PowerTune violates TDP for a period of time before a thermal event is triggered and the GPU is forced into a much lower performance state. Meanwhile, the average workload for typical applications trend to be well below the GPU TDP signaling that the GPU is not delivering optimal performance within its TDP. With PowerTune, we see a much tighter spread in power. All applications are managed by PowerTune to fit within the GPU TDP in a manner which avoids the thermal event and its associated performance drop. With PowerTune, the typical applications benefit from the higher PowerTune-enabled maximum clock frequencies to make use of the available thermal headroom of the GPU for the power profile associated with the application; delivering much higher overall performance.
Summary

AMD PowerTune technology represents a major shift in how GPUs are power managed to maximize their performance potential. With AMD PowerTune technology’s ability to intelligently monitor and manage dynamic power, GPUs can be designed to meet thermal constraints and move past the traditional tradeoffs of accommodating power heavy applications at the expense of average performance. The net result with AMD PowerTune technology is the ability to enable GPUs with higher factory engine clocks which deliver improved performance across the board.
Introduction

During static screen operation, a GPU continuously refreshes display device(s) from its frame buffer. A GPU may minimize static screen idle power by enabling a host of active power saving techniques including (but not limited to) clock gating, power gating, memory compression and stutter, as well as a number of others. Generally the same idle power savings techniques have been used when there is no display refresh required.

However, GPUs with AMD’s exclusive ZeroCore Power technology take power efficiency to entirely new levels by completely powering down the GPU core while the rest of the system is allowed to remain in an active idle state.

Background

Nearly all PCs can be configured to turn off their displays after a long period of relative inactivity and lack of user input. This is known as the long idle state; where the screen is blanked but the rest of the system remains in an active and working power state (referred to as the G0/S0 ACPI states). When the PC reaches this state and applications are not actively using background GPU resources, the GPU enters a state where the graphics core power draw is minimized. In this state, all major functional blocks of the GPU (including the compute units; multimedia, audio and display engines; memory interfaces; etc.) are completely powered down.

However, one cannot simply remove the GPU and its associated device context completely; particularly when it is the only GPU in the system. The OS, SBIOS and rest of system cannot function without a primary graphics device and must still be aware that a GPU is logically present in the system. The innovation of AMD ZeroCore Power technology is that it maintains a very small hardware - level bus control block to ensure that the GPU context is still visible to the OS and SBIOS (the “ZeroCore Power state”). The ZeroCore Power state also manages the power sequencing of the GPU to ensure that the power up/down mechanism is self-contained and independent of the rest-of-the system.

At the system level, the ZeroCore Power state is controlled by the driver. When the GPU driver determines that the system meets the condition that applications are not updating display contents or using background GPU resources, the GPU is put into the ZeroCore Power state once the system is in long idle. If any applications update the screen contents in the long idle state, the driver can periodically wake the GPU from the ZeroCore Power state to update the contents of the frame buffer and put the GPU back into the ZeroCore Power state. While the AMD graphics driver can handle applications which may wake the GPU from the ZeroCore Power state, many applications are ‘power state’ aware to minimize system activity during long idle. One such example is gadget applications for the Windows 7 operating system. These gadgets are known to suspend updates to the display in the long idle state and resume updating their dynamic contents (weather, RSS feeds, stock symbols, slideshows, etc.) once the system exits long idle. These applications will not wake the GPU from the ZeroCore Power state in long idle. Figure 6 highlights the power down condition for AMD ZeroCore power from the traditional static screen idle state to the long idle state.
Scalable Energy Efficiency with AMD CrossFire™ Technology

AMD ZeroCore Power technology scales to enable exceptional power efficiency with platforms employing AMD CrossFire™ technology. Traditionally, multi-GPU platforms have had to keep all GPU cores powered on to ensure that their context is readily visible to the system (including the OS, SBiOS and applications) which required the non-primary GPUs to be in an idle or near-idle state. With AMD ZeroCore Power technology, this context can be maintained in hardware while the core graphics engine is completely powered. The end result is an AMD CrossFire system which moves beyond the traditional power limitations of multi-GPU configurations. Additional GPUs in the system consume the absolute minimum of power by virtue of the graphics engine core being completely powered down. Similarly, AMD ZeroCore Power technology enables AMD CrossFire systems to scale to 4 total GPUs without an increase in idle noise. The GPU driver intelligently wakes the secondary GPUs from the ZeroCore Power state when needed to ensure that the full performance potential is realized during active workloads. Meanwhile the primary GPU in the system is enabled to leverage the ZeroCore Power state while in long idle similar as explained in the previous section. Figure 7 shows how AMD ZeroCore Power technology powers down all GPUs in the system at every opportunity.
AMD ZeroCore Power enables tremendous energy efficiency and end user benefits. By completely powering down the GPU core in the long idle state, users can still enjoy non-graphics activities such as file serving/sharing/streaming, motherboard audio and music and remote access without worrying about the traditional GPU power costs and impact. With systems employing AMD CrossFire technology, AMD ZeroCore Power technology powers down all GPUs cores in the system at every opportunity to enable incredible power efficiency alongside AMD CrossFire’s traditional value proposition of incredible performance. The unique ability of AMD ZeroCore Power technology to tremendously improve energy efficiency in single and multi-GPU configurations creates a technology that is highly relevant across the spectrum benefiting everyday consumers, enthusiasts and professionals.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE.

IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

This document contains forward-looking statements, which are made pursuant to the safe harbor provisions of the U.S. Private Securities Litigation Reform Act of 1995. Forward-looking statements are generally preceded by words such as “plans,” “expects,” “believes,” “anticipates” or “intends.” Investors are cautioned that all forward-looking statements in this release involve risks and uncertainty that could cause actual results to differ materially from current expectations. We urge investors to review in detail the risks and uncertainties in the Company’s filings with the United States Securities Exchange Commission.

SUBSTANTIATION
1 AMD PowerPlay™, AMD PowerTune and AMD ZeroCore Power are technologies offered by certain AMD Radeon™ products, which are designed to intelligently manage GPU power consumption in response to certain GPU load conditions. Not all products feature all technologies – check with your component or system manufacturer for specific model capabilities.

©2012 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, Catalyst, CrossFire, PowerPlay, Radeon and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other names are for informational purposes only and may be trademarks of their respective owners. PID 51587A