

# AMD PowerTune Technology

December 2010

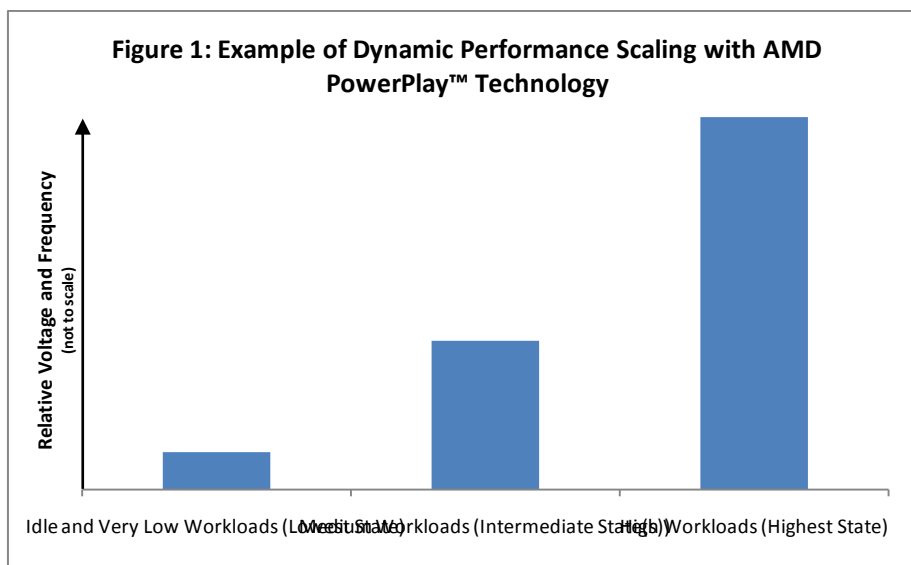
# Table of Contents

<b>BACKGROUND .....</b>	<b>3</b>
AMD PowerPlay™ Technology .....	3
Thermal Design Power and GPU Design .....	4
<b>DYNAMICALLY OPTIMIZED PERFORMANCE WITH AMD POWERTUNE TECHNOLOGY .....</b>	<b>6</b>
Introduction.....	6
Dynamic TDP Management .....	7
Improved Performance on Critical Applications .....	8
Minimized Performance Impact on Applications which would Otherwise Exceed TDP .....	8
Programmable, Deterministic and Application Profile Independent Behaviour .....	9
<b>SUMMARY .....</b>	<b>10</b>
AMD PowerTune Technology Maximizes TDP Constrained Performance by Enabling Higher GPU Clock Speeds.....	10

## BACKGROUND

### AMD PowerPlay™ Technology

For many years, nearly all ATI Radeon™ and AMD Radeon™ desktop and notebook discrete graphics products have been equipped with AMD PowerPlay™ (formerly ATI PowerPlay™). AMD PowerPlay™ is a set of technologies which includes a highly advanced form of Dynamic Power Management (DPM) which assess relative workloads to aggressively conserve power (and battery life in the case of notebooks) when the demand on the graphics processor is low. With this technology the GPU can minimize power during light workloads such as idle mode by enabling reductions in voltages, engine and memory clock speeds. In such cases, the GPU is in the lowest power state of voltage and frequency. When demanding workloads are placed on the GPU, AMD PowerPlay™ technology increases voltages and clock speeds to a significantly higher state for maximum performance. High workloads tend to push the GPU into the highest power state. ATI PowerPlay™ technology also supports intermediate power state(s) for tasks of light-to-moderate demand – such as light 3D, video and compute tasks. An example of an AMD PowerPlay™ enabled GPU with 3 primary power states is shown in Figure 1.



Traditionally, the highest power state has been a fixed setting. When the discrete graphics device is in this state, the voltage and clock speeds are the same regardless of the type of application that is being processed. In practice, the manner in which applications load the GPU can vary greatly based on how they are coded and how they specifically interact with the GPU architecture. As a result, GPU power draw in the highest power state can vary to a large degree based on the specific application that is running.

## Thermal Design Power and GPU Design

Like all microprocessors, GPUs consume electrical energy while in operation, and convert it to heat energy which must be dissipated. The rate of energy consumption is therefore limited by a system's ability to both deliver power to the device and cool it by removing the heat it generates. GPU manufacturers provide system builders with a Thermal Design Power (TDP) figure for their products to allow them to design their systems appropriately. This represents the maximum power draw for reliable operation. There are many factors which can affect TDP, including:

- Voltages and Clock speeds – higher transistor voltages and switching speeds mean more power consumption
- Workload – applications that keep a larger percentage of the chip busy most of the time will draw more power
- Leakage – transistors consume some amount of power even when they are not switching; the amount of leakage for a particular GPU can vary significantly as a result of the manufacturing process and changes in operating temperatures
- Ambient temperature – GPUs operating in a hot environment, or in enclosures with restricted airflow, are more difficult to cool; temperature can also increase if a device is kept heavily loaded for significant periods of time

TDP figures are typically provided for the entire graphics card including the GPU ASIC, voltage regulators, memory devices, interconnects, and other board components. However, assumptions are made when generating such figures regarding the various factors mentioned above. For maximum reliability, TDP figures should assume a 'worst case' scenario. In the case of discrete graphics cards, this usually means running at the maximum supported clock frequency, on a device with the highest allowable leakage, with one or more known stressful workloads running trouble-free for several minutes in a closed system, and with multiple displays connected (as many as the card can support simultaneously).

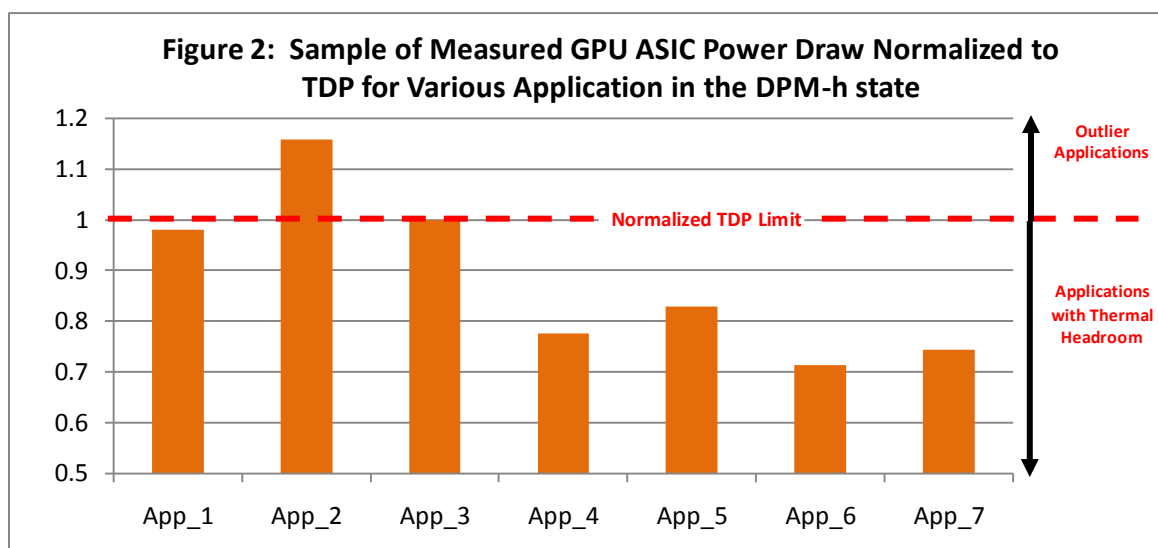
Some allowances can still be made to keep the TDP reasonable, such as limiting ambient temperature to 45 C, so long as these can be assumed to fall well outside typical usage scenarios. However, this can still result in a TDP much higher than what most users are likely to encounter in normal operation. For this reason, a "typical maximum board power" figure can also be useful. This still represents a board running at maximum frequency with stressful workloads, but assumes a device with average leakage rather than worst case, fewer connected displays, and a well-ventilated system.

Furthermore, voltage and clock speeds that are selected for the highest power state of a GPU include consideration of the following factors:

1. The TDP constraints of the overall design using a 'worst case' approach
2. The upper limit of frequency of the GPU for a given voltage

3. Applications used to determine the power characteristics of the GPU in the high power state

These three factors will largely determine the upper limit of a given GPU design. In many cases, GPUs can exceed the TDP limits of their designs well before reaching their clock speed limits. This is particularly common with GPUs in power constrained notebook platforms as well as very high performance desktop platforms. For example, a notebook GPU may be able to reach a clock of 900 MHz, but thermal design constraints may limit its clock in the high state to 550 MHz to ensure that it does not go over an assigned power budget of 15 Watts under any circumstances. Similarly, a high performance desktop GPU may need to be limited in frequency to fit within a given board power envelope such as 75, 150, 225 or 300 Watts. Such requirements need to be strictly enforced to ensure that the desktop or notebook system design considerations are not compromised.



A sample of measured graphics applications and corresponding levels of normalized power draw for a GPU ASIC is provided in Figure 2. The determination of the final voltage and clock speeds for the TDP allowance in the high state is generally based on a set of applications that are known to be particularly power intensive. These include applications that are known to put an exceptionally high demand on the GPU. Some applications are written for the specific purpose of pushing the GPU past its thermal limits and are generally referred to as outlier applications. Outlier applications tend to generate much more activity within the GPU silicon than the vast majority of applications and consequently generate the largest dynamic power requirements. However some well-known 3D applications that are not written for such specific purposes are known to push some GPUs beyond their TDP limits. Some outlier applications load the GPU in a transient fashion such that they may only approach or exceed TDP limits occasionally. For example, power dynamic power for a 3D application can vary based on the content of the rendered scene.

Despite these factors, the majority of applications do not necessarily approach the TDP of the GPU in the highest power state.

The existing method of dealing with applications that may exceed TDP includes thermal monitoring which may lead to a thermal event flag. A thermal event occurs when the GPU is loaded to the point where the junction temperature exceeds a pre-determined warning value and forces the GPU into either an intermediate power state or the lowest power state. Hence an application that triggers a thermal event will be forced to a much lower level of performance to ensure that TDP limits are not being exceeded. While this helps to ensure that TDP limits are being enforced, it is not an ideal situation from an application performance perspective. A more ideal scenario would be to precisely curtail the power and manage it gradually so that it is slightly below an absolute limit while the outlier application is running.

For applications that do not approach the limit of a TDP constrained GPU during the highest power state, a situation arises where more performance could be made available if the GPU could be cognizant of the available power headroom. However since the GPU's highest power state clock speeds are fixed settings, the potential of added performance is not realized. In this situation, performance is left on the table since the GPU has additional headroom for more performance, but lacks a mechanism to exploit it.

## **DYNAMICALLY OPTIMIZED PERFORMANCE WITH AMD POWERTUNE TECHNOLOGY**

---

### **Introduction**

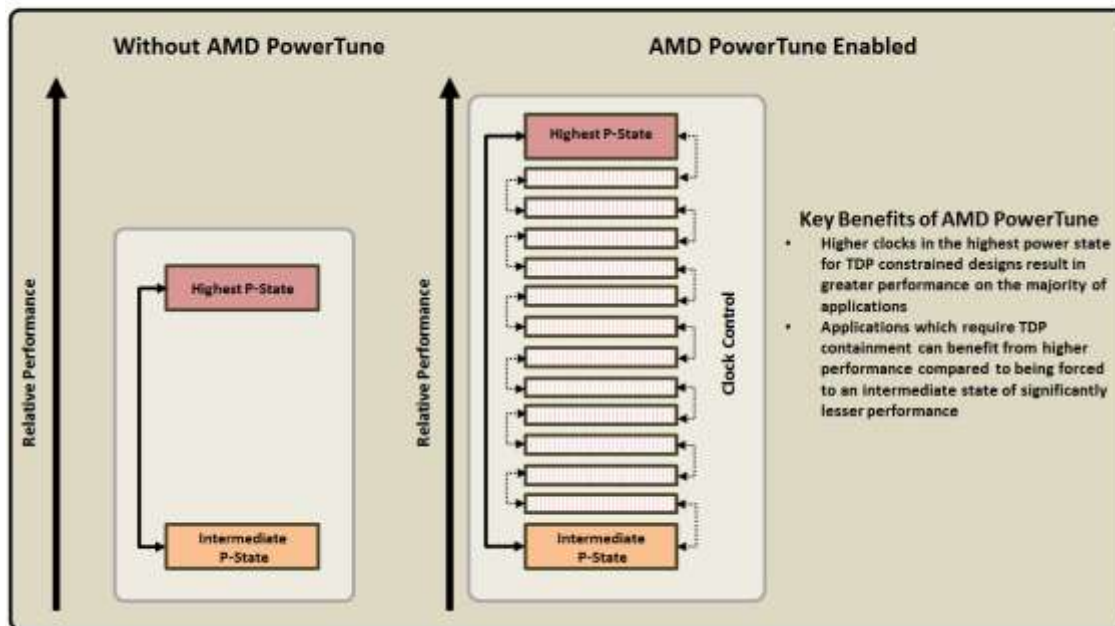
AMD PowerTune technology is a very significant leap forward to better ensure that performance is optimized for TDP constrained GPUs. AMD PowerTune technology helps deliver higher performance that is optimized to the thermal limits of the GPU by dynamically adjusting the clock during runtime based on an internally calculated GPU power assessment. AMD PowerTune technology also improves the mechanism to deal with applications that would otherwise exceed the GPU's TDP. By dynamically managing the engine clock speeds based on calculations which determine the proximity of the GPU to its TDP limit, AMD PowerTune allows for the GPU to run at higher nominal clock speeds in the high state than otherwise possible. AMD PowerTune technology is very different from existing methods; rather than setting highest state GPU clock speeds based on a worst case TDP approach that can compromise performance in a majority of applications, AMD PowerTune technology can dynamically adjust the performance profile in real time to fit within the TDP envelope.

## Dynamic TDP Management

Traditionally with AMD PowerPlay™ technology, modern AMD GPUs were equipped to transition between fixed power states with the upper states having increasingly higher clock speeds and voltages to increase performance when needed and minimize power when performance is not needed.

AMD PowerTune technology expands on this by removing the constraint that a given power state must be fixed. The AMD PowerTune algorithm embedded in the GPU hardware calculates the engine clock based on an internal assessment of the runtime power draw. When the GPU is in highest activity or power state and not exceeding TDP, it will remain in the highest power state for maximum performance. In the case where AMD PowerTune calculates that the GPU is exceeding TDP, the power is dynamically reduced in a gradual manner by reducing the clock while still maintaining the high power state. The amount of clock reduction is variable and depends on the GPU's assessment of the power draw. A representation of how the GPU engine clock speeds in the highest state can be managed is shown in Figure 3.

**Figure 3: GPU Power State Comparison with AMD PowerTune**



This approach has multiple advantages. First, it allows TDP constrained GPUs to ship with engine clock speeds in the highest state that would otherwise have been lower without AMD PowerTune technology. This subsequently provides greater performance on the majority of applications which do not exceed the TDP constraints on the GPU. Second, it helps to avoid

throttling of the GPU for extreme outlier applications by managing down the GPU clock speeds before a thermal event is flagged. This results in outlier applications running at significantly higher levels of performance than would otherwise be possible since the GPU is not necessarily forced into an intermediate or low power state through a thermal event. When an application is running that would otherwise exceed the TDP limit, AMD PowerTune technology can adjust the clock to contain the runtime power at a safe level that is slightly below the TDP limit. Also, outlier applications tend to vary in their runtime workloads. AMD PowerTune can manage this while the application is running by re-calculating power draw many times within a given frame interval. By keeping the outlier application in the realm of the highest state (albeit an inferred state with a reduction in engine clock), the fast transient response of the AMD PowerTune algorithm is able to quickly raise clock speeds back to the nominal highest power state levels if the near term demands of the application create additional headroom.

## Improved Performance on Critical Applications

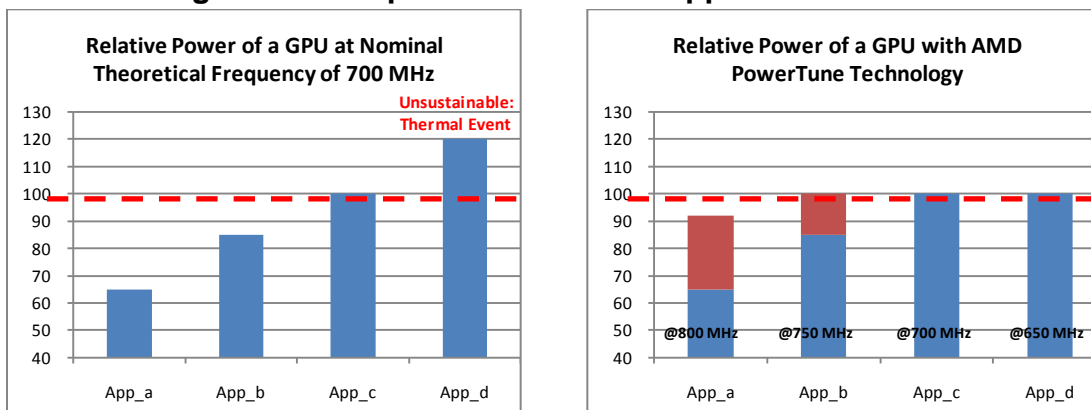
It is quite common for applications running in the highest power state to have TDP levels below the GPU allowance. The large majority of applications are not outliers. When the high power state clock speeds are fixed, it is not possible to take advantage of the remaining TDP headroom and increase clock speeds at runtime to further improve application performance. However, with AMD PowerTune technology the GPU is able to ship with engine clock speeds in the highest state that are greater than what could be achieved without the technology. As a result, AMD PowerTune technology directly improves performance on critical applications.

As outlined in the example in Figure 4, an application with a large amount of headroom (in this case, App\_a) has the greatest potential for performance improvement with AMD PowerTune technology by allowing the maximum engine clock available for the GPU. Similarly, an application with some headroom (App\_b) can still take advantage of the higher available clock speeds in the highest power state, but perhaps to a lesser degree than an application with more headroom.

## Minimized Performance Impact on Applications which would Otherwise Exceed TDP

AMD PowerTune technology allows TDP constrained GPUs to ship with greater nominal clock speeds in the highest power state due to the mechanism by which it handles applications which exceed TDP. Without AMD PowerTune technology, applications which exceed the GPU TDP are forced to lower power states (such as intermediate or lowest states) and pay a very steep performance penalty as a result of drastically reduced clock speeds and voltages. In the AMD PowerTune enabled GPU, the clock speeds in the highest state can be dynamically managed to hold the TDP budget in a way that was not otherwise achievable. The goal for applications that exceed the TDP budget is to maintain operation in the highest power state, but dial back on runtime power by modulating the high power state clock to keep the TDP range slightly below the absolute limit. This keeps the GPU away from the undesirable performance penalty of a forced state reduction arising from a thermal throttling event.

**Figure 4: Comparison of Outlier Application Behaviour**



## Programmable, Deterministic and Application Profile Independent Behaviour

AMD PowerTune power monitoring and management technology is integrated into the GPU silicon itself to essentially eliminate the unpredictable variability that would arise if it were implemented at the board or system level with analog sensors and feedback mechanisms. Activity is monitored in to infer real time power draw at the device level through integrated counters that are placed throughout the GPU. As a result, AMD PowerTune is transparent in its ability to contain applications in real time without a reliance on specific drivers or application profiles. AMD PowerTune technology is also programmable in a way that can allow GPU and system designers to tailor the power containment behaviour to the specific needs of the user. This can allow some systems to set a lower TDP threshold to enable a cooler overall system under heavy load, or set a higher TDP threshold in the case where there is known to be additional TDP headroom.

A much less desirable approach would be to contain GPU power through application profiles and application detection. This creates a dependence on driver software that simply cannot take into account all applications that can use GPU compute and rendering resources. Driver based application profiles can also become outdated very quickly and result in severe performance degradation. AMD PowerTune technology's hardware based digital monitoring solution contained within the GPU silicon avoids these pitfalls.

## SUMMARY

---

### **AMD PowerTune Technology Maximizes TDP Constrained Performance by Enabling Higher GPU Clock Speeds**

AMD PowerTune is a breakthrough technology that sets an entirely new direction for maximum performance at TDP. It allows the GPU to be designed with higher engine clock speeds which can be applied on the broad set of applications that have thermal headroom. It also improves how GPUs manage outlier applications by managing them down to power levels within the TDP limits with minimal performance impact. Without AMD PowerTune technology, a TDP limited GPU's final clock speeds would inherently be based on a compromise between severe performance loss on higher power applications and performance left on the table with lower power applications. With the intelligent monitoring and management capabilities introduced by AMD PowerTune technology, these compromises are removed to maximize performance across the board.

## Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

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, NON-INFRINGEMENT, 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.

## Attribution

© 2010 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD logo, ATI, Radeon, PowerPlay, 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.