Reverse engineering power management on NVIDIA GPUs

Anatomy of an autonomic-ready system

Martin Peres

Ph.D. student at LaBRI, Bordeaux

July 9, 2013

Summary

Introduction – Motivation

Power management in computers, why?

- To lower the power consumption of Data Centers;
- To increase the battery life of mobile computers;
- To have quieter and slimmer devices.

Reverse engineering power management, why?

Power management is:

- at least partially-assisted by software;
- almost entirely non-documented;
- often considered to be a manufacturer secret;
- thus poorly studied by independent researchers;
- this is especially true in the GPU world.

Introduction – Nouveau

Nouveau : An open-source driver for NVIDIA cards

- Developed through clean-room reverse engineering;
- Supports all cards starting from the TNT2 (released in 1998);
- Provides video, 2D and 3D acceleration when possible;
- Plans on fully supporting GPGPU and power management.

Reverse engineering NVIDIA cards, how?

- By spying communications on the PCIE bus (MMIO trace);
- By tracing commands sent to the card (valgrind-mmt);
- By developing tools to help us (envytools);
- By poking and peeking registers.

Summary

Origin of the power consumption

Power consumption of a logic gate

$P = P_{static} + P_{dynamic}$

P_{static} : Base of the power consumption (leakage)

Depends on the etching process of the transistors.

$P_{dynamic}$: Power consumption related to the clock

- $P_{dynamic} = CfV^2$;
- C: Capacitance of the gate (fixed);
- f: Frequency at which the gate is clocked;
- V: Voltage at which the gate is powered.

Usual ways of saving power

Usual ways of saving power

- Clock gating: Cuts the dynamic-power cost;
- Power gating: Cuts all the power cost;
- Reclocking: Adjusts the clock frequency and voltage.

Clock gating: Stopping the clock of un-used gates

- Update rate: Every clock cycle;
- Effectiveness: Cuts the dynamic-power cost entirely;
- Drawbacks: Increase of the complexity of the clock tree;
- Executed by: Hardware.

Usual ways of saving power

Power gating: Shutting down the power of un-used gates

- Update rate: Around a microsecond;
- Effectiveness: Cuts the power cost entirely;
- Drawbacks: Need to save the context before shutdown;
- Executed by: Hardware and/or software.

Reclocking: Dynamic Voltage/Frequency Scaling (DVFS)

- Update rate: Around a millisecond;
- Effectiveness: Impacts the dynamic-power cost;
- Drawbacks: Affects performance;
- Executed by: Software.

Summary

PCOUNTER – Overview

Performance counters

- are blocks in modern processors that monitor their activity;
- count hardware events such as cache hit/misses;
- are tied to a clock domain;
- provide load information needed for DVFS's decision making.

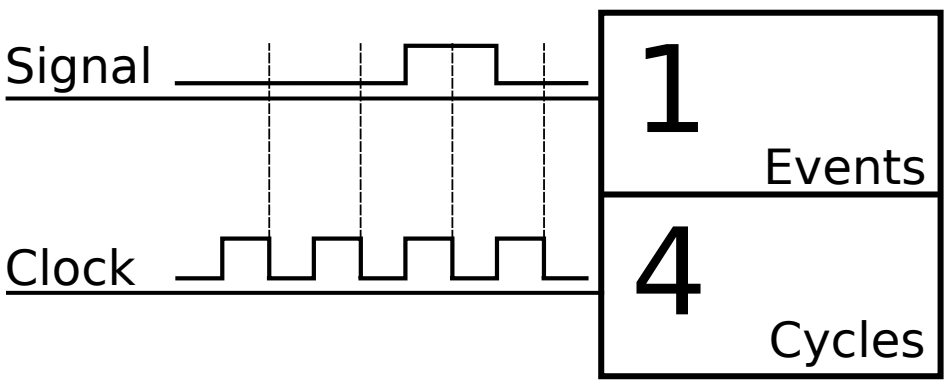

Figure : Example of a simple performance counter

PCOUNTER – Overview of a domain

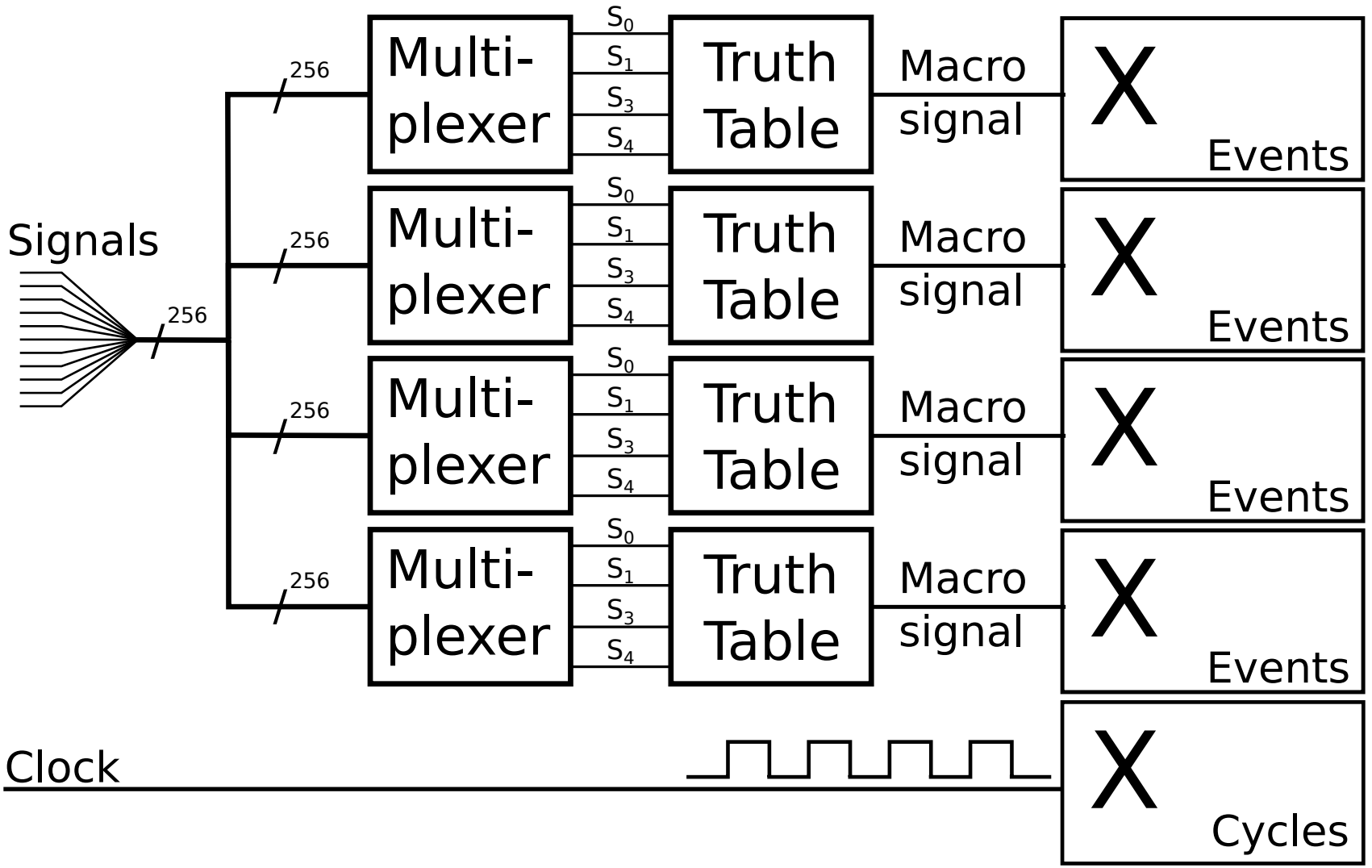

Figure : Schematic view of a domain from PCOUNTER

Summary

PTHERM – Thermal management

PTHERM's thermal management

- sends IRQs to the host when reaching temperature thresholds;
- can cut the power of the card through a GPIO;
- can force the fan to the maximum speed;
- can lower the frequency of the main engine of the GPU.

PTHERM – Frequency-Switching Ratio Modulation

Frequency-Switching Ratio Modulation (FSRM)

- is used to lower the frequency of the main engine of the GPU;
- is useful to lower the temperature or the power consumption;
- is triggered automatically when reaching thresholds.

How can the FSRM lower power consumption?

- A divided clock is generated from the main engine's clock;
- The clock must be divided by a power-of-two (2 to 16);
- It can generate any clock frequency between these two clocks;
- With a lower clock, an engine consumes less power.

PTHERM – Frequency-Switching Ratio Modulation

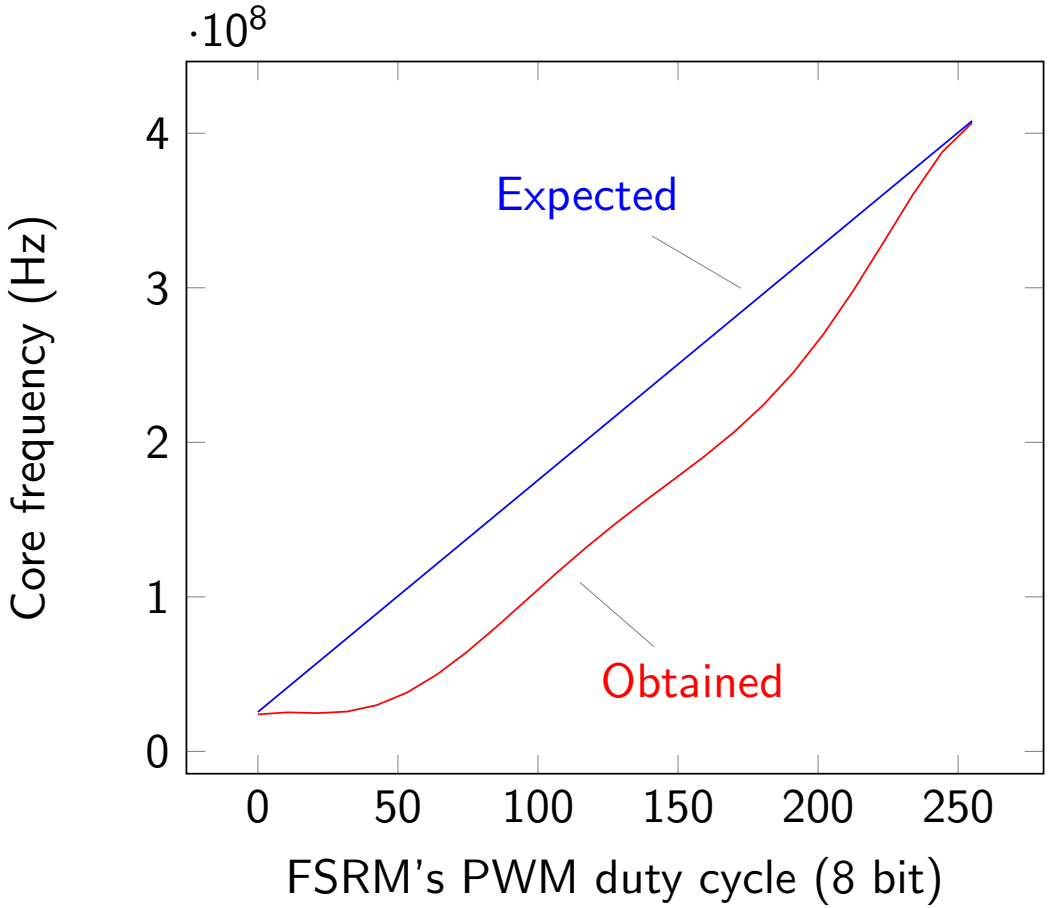

Figure : Frequency of the core clock (original @ 408MHz) when using a 16-divider and varying the FSRM

PTHERM – Power limitation

PTHERM's power limitation can

- read the power consumption by counting the active blocks;
- update the FSRM ratio to stay in the power budget;
- use two hysteresis windows for altering the FSRM ratio;
- do all that automatically.

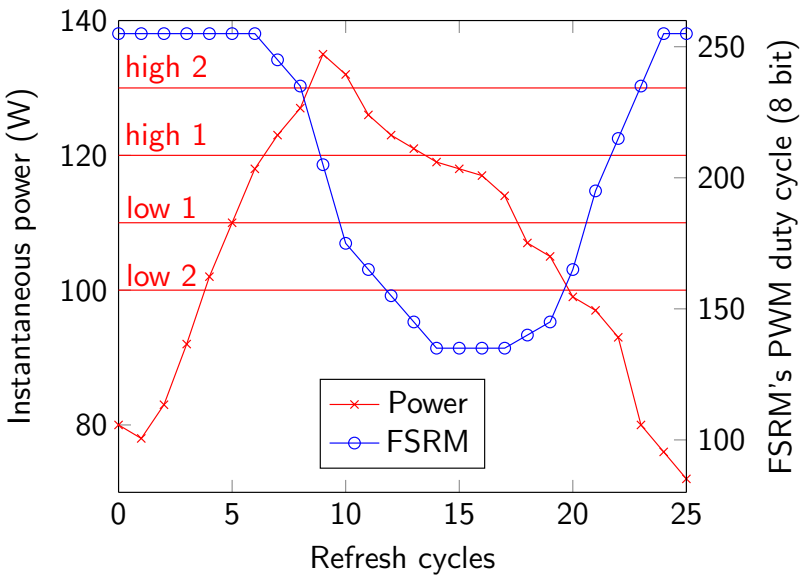

Figure : Example of the power limiter in the dual window mode

Summary

PDAEMON – An embedded RTOS in your GPU

PDAEMON

- is an RTOS embedded in every new NVIDIA GPU (Fermi+);
- clocked at 200MHz and is programmed in the FμC ISA;
- has access to all the registers of the card;
- can catch all the interrupts from the GPU to the Host;
- features internal performance counters.

NVIDIA's usage of PDAEMON

- Fan management;
- Hardware scheduling (for memory reclocking);
- Power gating and power budget enforcement;
- Performance and system monitoring.

Summary

Conclusion

The GPU as an autonomic system

The GPU can:

- self-configure: thanks to PDAEMON that can act as a driver;
- self-optimise: using the performance counters;
- self-heal: recovering from over-temperature/current;
- self-protect: GPU users are isolated in separate VM.

Future works

- Implement stable reclocking across all GPUs;
- Write a test-bed for DVFS algorithms implementations;
- Document clock- and power-gating details;
- Reverse engineer more performance-counter signals.

Questions & Discussions

Questions & Discussions

FSRM – How does it work?

FSRM, how does it work?

The FSRM:

- is set in-between the source clock and the engine;
- generates a power-of-two divided clock (2, 4, 8 or 16);
- mixes both frequencies by alternating between the two clocks;
- can thus generate any frequency between the two clocks;
- can thus linearly affect the dynamic power consumption.

PTHERM – Power limitation

Calculating the power consumption

PTHERM estimates power consumption by:

- reading every block's activity (in use or not);
- summing the weighted activity blocks signals;
- applying a low pass filter.

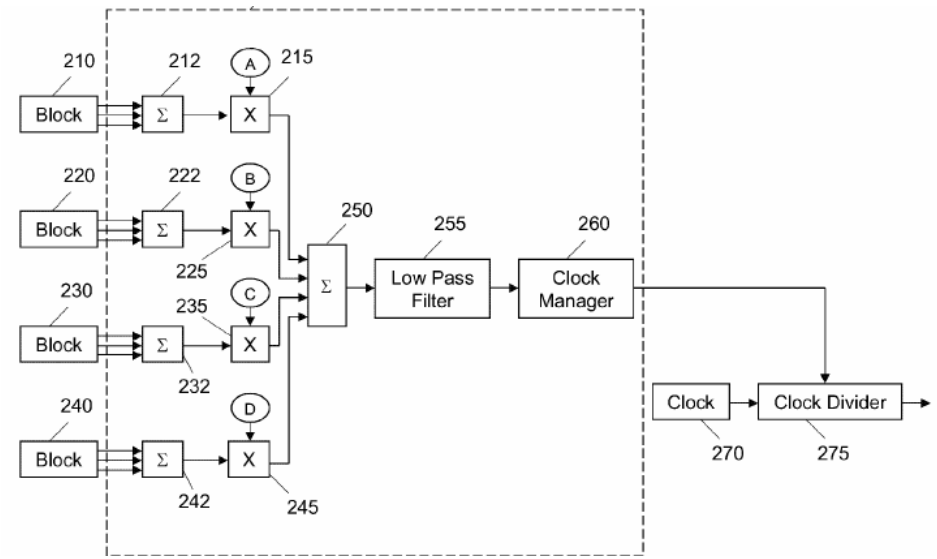

Figure : Extract of NVIDIA's patent on power estimation (US8060765)

Power limitation – Actual implementation of NVIDIA

Power limitation – Actual implementation

- NVIDIA doesn't use PTHERM to implement power limitation;
- It may read power consumption from the voltage controller;
- and downclock the card when exceeding performance.

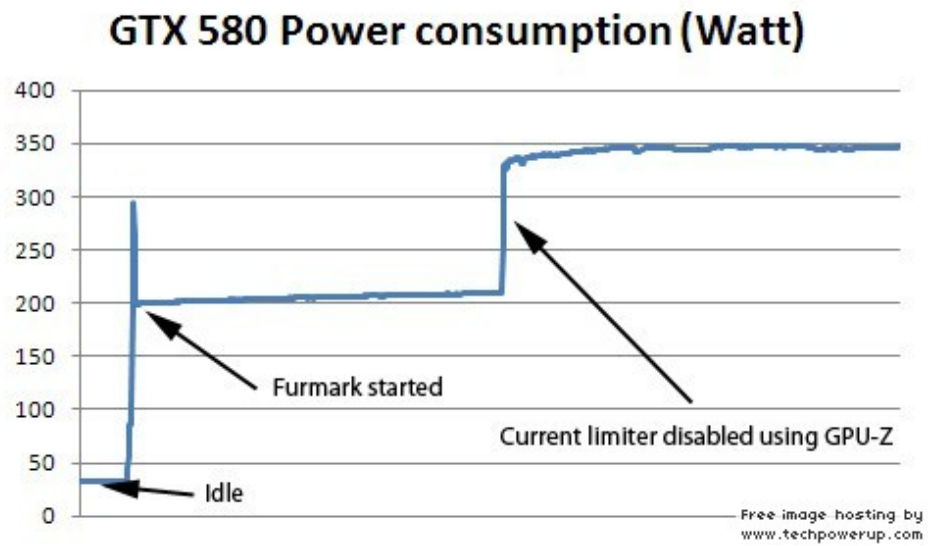

Figure : Effect of disabling the power limiter on the Geforce GTX 580. Copyrights to W1zzard from techpowerup.com.