Power management on NVIDIA GPUs
Anatomy of an autonomic-ready processor

Martin Peres
LaBRI, Université de Bordeaux - France (martin.peres@labri.fr)

Power consumption of a CMOS gate

Power consumption:
- \( P = P_{\text{dynamic}} + P_{\text{static}} \)
- sum of the dynamic and static power consumption

Static power consumption: leakage current of the gate
- \( P_{\text{static}} = V I_{\text{leak}} \)
- influenced by the voltage at which the gate is powered

Dynamic power consumption: fighting the capacitance of the gate
- \( P_{\text{dynamic}} = C F_{\text{V}} \)
- square impact by the switching frequency (\( \sim \) clock)

Power consumption of GPU clients:
- could be calculated because GPUs are executing one thing at a time
- requires detecting the hardware context switch (easy)
- requires polling the power sensor: can be done by the RTOS

Power scheduler:
- using the above solution to implement power consumption quotas
- quotas could be instantaneous or averaged
- the RTOS needs to reclock / context switch when the quotas is met

Changing the voltage and frequency

PCIe Port 75W
6 pins
8 pins

Voltage Controller
Chipset

Power adapter

VID (/5)
GPIOs
Therm Shutdown

PCIE power supply connected

Power consumption:
- can be read using Ohm’s law, as seen on the above figure
- can be calculated by counting active blocks and using a hw model

Reading the GPU’s usage: Performance counters

Performance counters:
- are counting hardware events (engine-idle, cache hit/miss, ld32, ...)
- are tied to a clock domain
- can be read, configured and reset by the driver

Power gating:
- stops the power supply of un-used blocks / engines
- cuts the entire power consumption of the block / engine
- requires saving / reloading the context
- can be executed hundreds of times per second

Dynamic Voltage/Frequency Scaling (DVFS):
- change the performance level depending on the load
- change also if the card overheats or is using too much power
- is good to save power when the GPU is used
- can be executed tens of times per seconds

Autonomic-power management on NVIDIA GPUs:
- metrics are power consumption, perf. counters and temperature
- temperature can be regulated using the temperature sensor
- metrics are tied to a clock domain
- can be read, configured and reset by the driver

Future work

An autonomic power management, IBM’s vision:
- self-configuration: find a configuration to fill the user’s request
- self-optimization: save power while still meeting the QoS
- self-healing: power consumption when overheating
- self-protection: isolation between users and killing long-running jobs
- self-optimization: save power while still meeting the QoS
- self-configuration: find a configuration to fill the user’s request
- self-healing: power consumption when overheating
- self-protection: isolation between users and killing long-running jobs

Power-saving techniques

Clock gating:
- stops the clock of un-used blocks / engines
- cuts entirely the dynamic power consumption
- can be executed millions of times per second

Dynamic Voltage/Frequency Scaling (DVFS):