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

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

freedesktop.org

Power consumption of a CMOS gate

Power consumption:
- $P = P_{dynamic} + P_{static}$
- sum of the dynamic and static power consumption

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

Dynamic power consumption: fighting the capacitance of the gate
- $P_{dynamic} = C f V^2$
- squarely affected by the voltage at which the gate is powered
- linearly affected by the (fixed) capacitive charge of the gate
- linearly affected by the switching frequency ($\precsim$ clock)

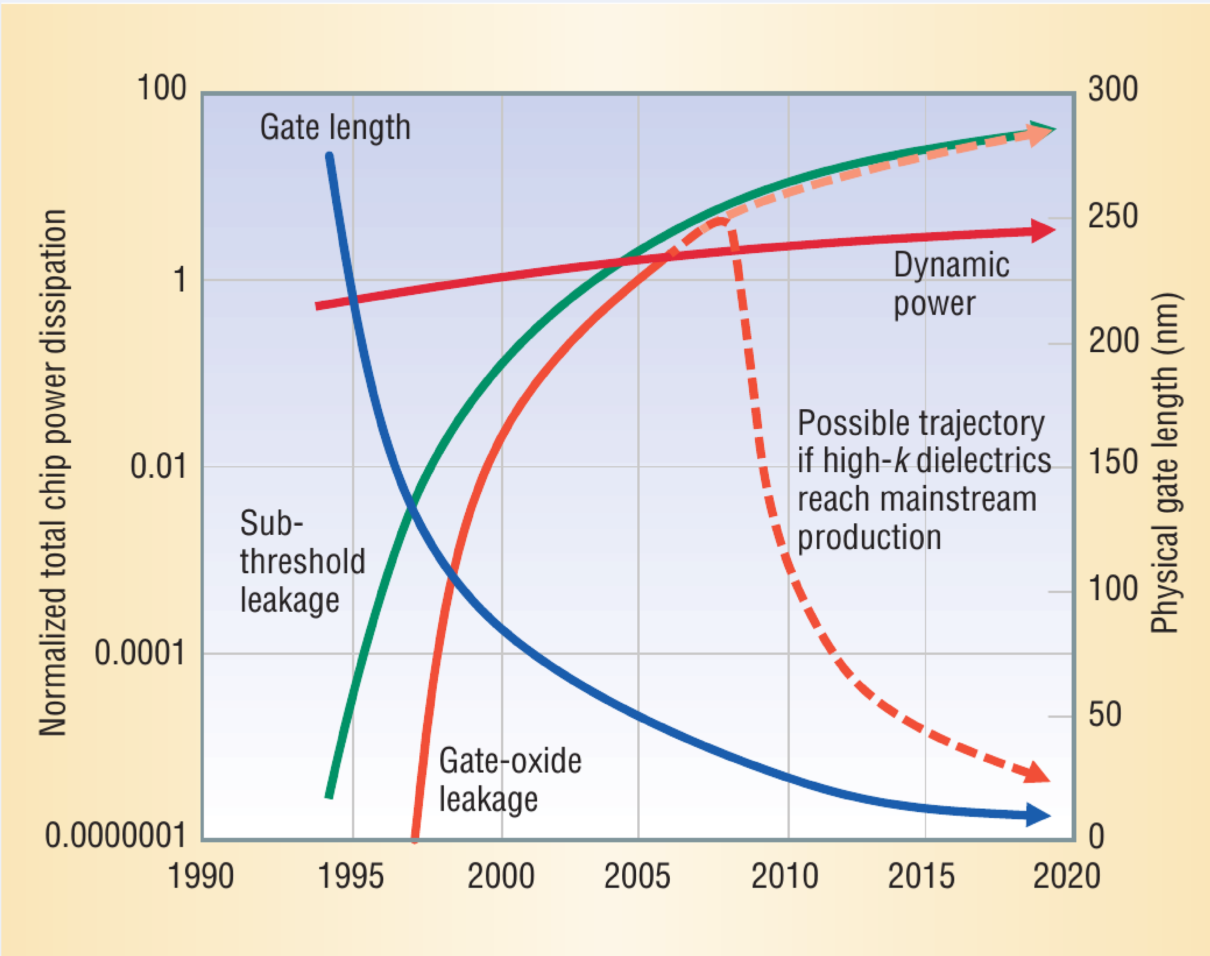

Figure: Evolution of the static and dynamic power consumption

Changing the voltage and frequency

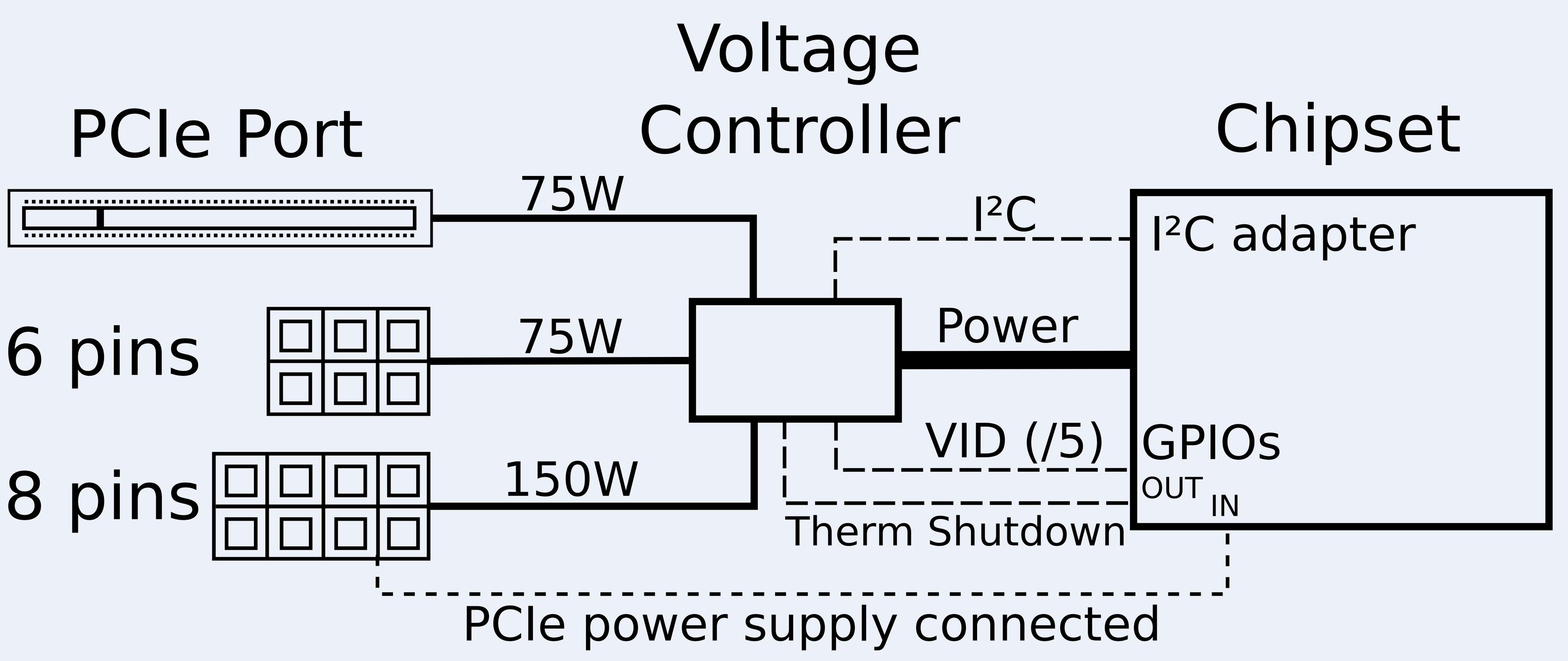

Figure: General overview of the power sources and voltage controller

Voltage:
- can be adjusted by selecting the right voltage ID (VID)
- should always be sufficient for the current clock of every engine
- is usually selected using a vbios-table mapping frequency to voltage

Frequency:
- can be adjusted by using Phase-Locked Loops
- is adjusted by a multiplier and a divisor factor: $F_{out} = F_{in} * \frac{N}{M}$
- is generally generated from a complex clock tree

Reading the power consumption

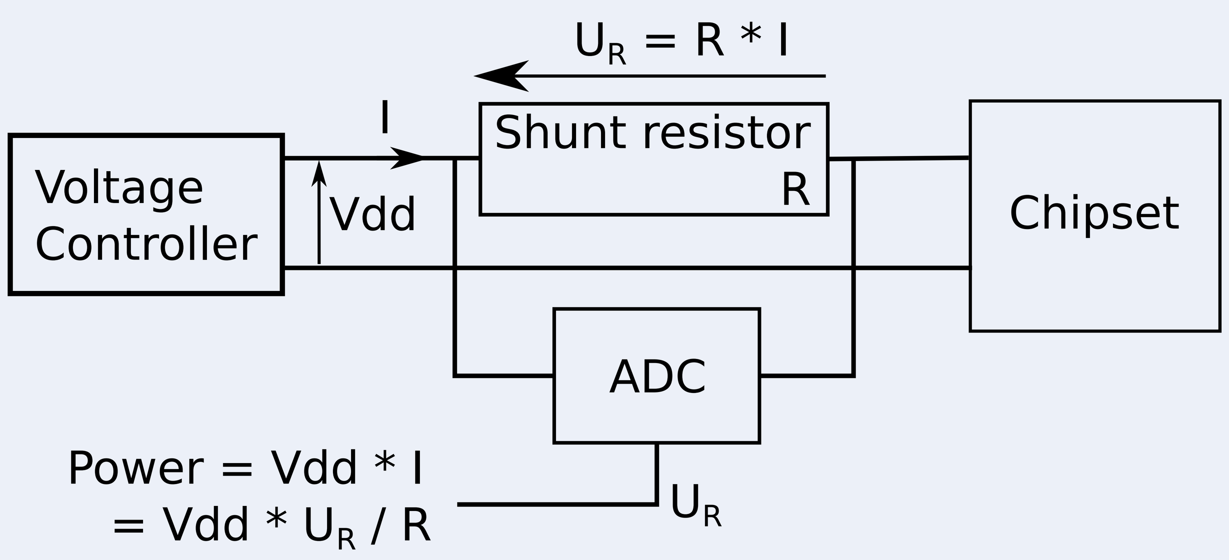

Figure: Reading the power consumption reading Ohm's law

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

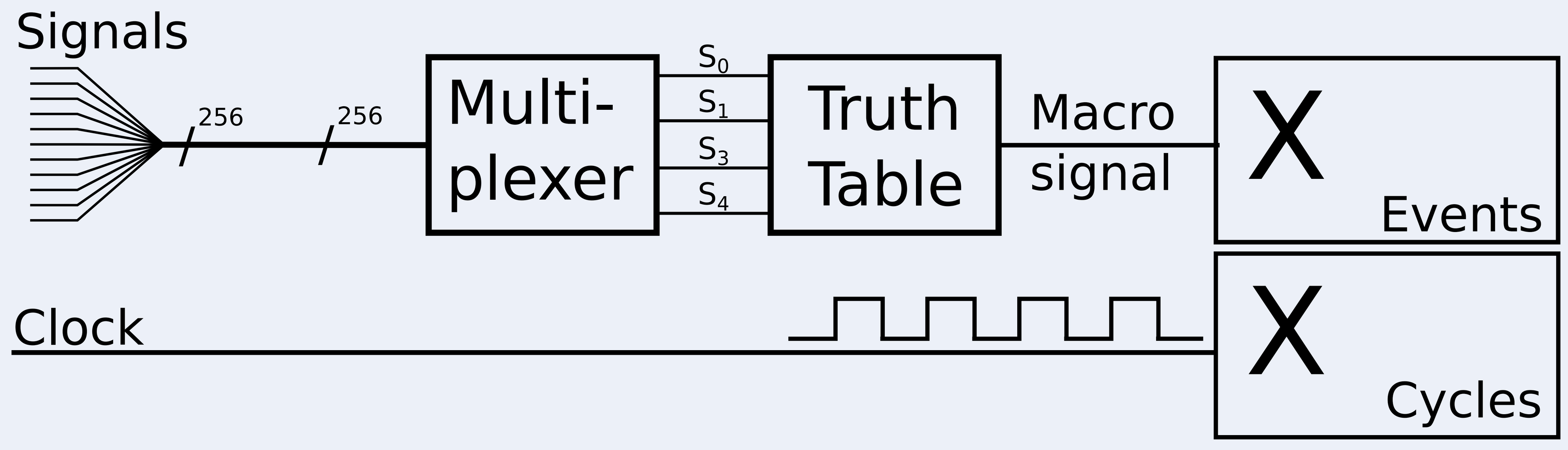

Figure: A simplified overview of a performance counter

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

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

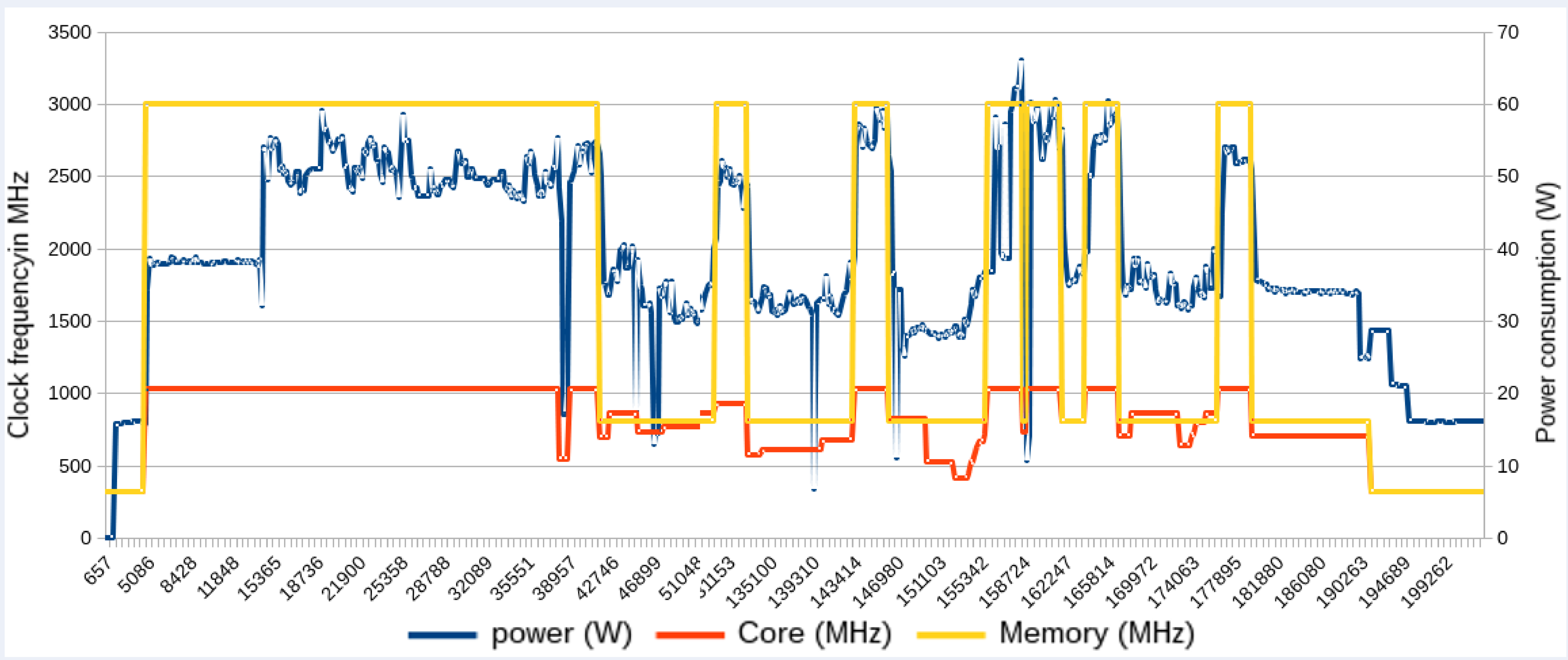

Figure: Evolution of the power consumption and clocks over time while playing a game (Xonotic) - wrong DVFS decisions

Autonomic-ready processors

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: lower power consumption when overheating
- self-protection: isolation between users and killing long-running jobs

Autonomic power management on NVIDIA GPUs:
- metrics are power consumption, perf. counters and temperature
- temperature can be regulated using the temperature sensor
- can be implemented in the RTOS embedded in newer GPUs

Future Work

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