TECHNISCHE UNIVERSITÄT DRESDEN

ZIH
Center for Information Services & High Performance Computing

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

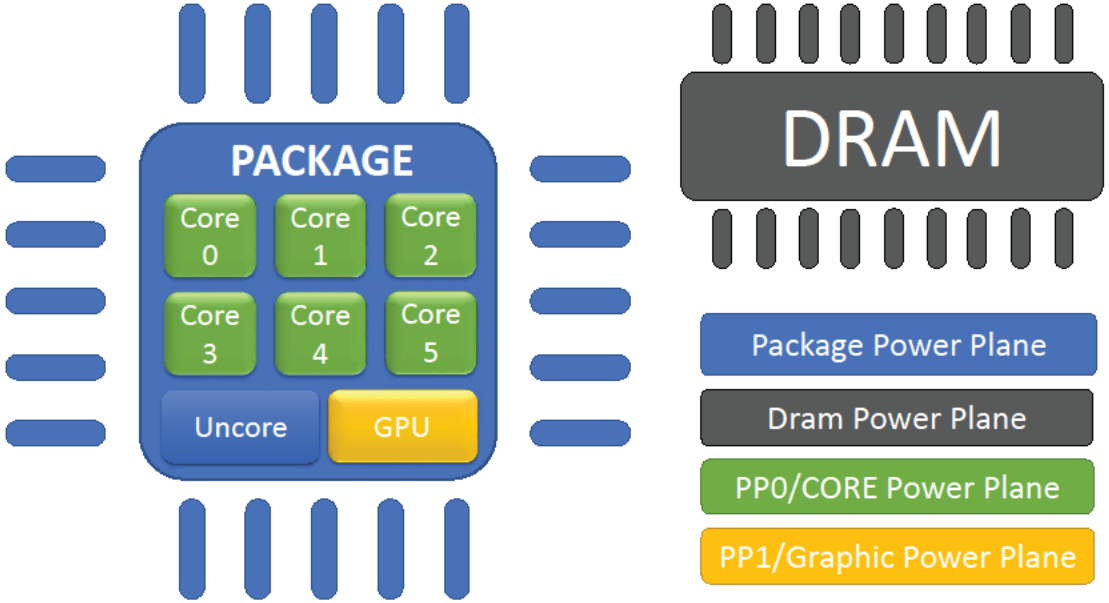
CINECA

RAPL

PRACE

HW Power Control - RAPL

Intel architectures implement a hardware power controller called Running Average Power Limit (RAPL).



Package Power Plane

Dram Power Plane

PP0/CORE Power Plane

PP1/Graphic Power Plane

Time limiter
(typically from few ms to seconds)

Power limiter
in Watts
(default TDP)

DVFS

Power Domains

Package Domain: limits the power consumption for the entire package of the CPU, this includes cores and uncore components.

DRAM Domain: is used to power cap the DRAM memory. It is available only for server architectures. (no client)

PP0/Core Domain: is used to restrict the power limit only to the cores of the CPU (no new server).

PP1/Graphic Domain: is used to power limit only the graphic component of the CPU (no server).

N.B. in the last slides of this presentation there is a complete description of RAPL registers

RAPL – Usage scenarios

- Hardware:
  - Measure power consumption for Turbo frequencies
  - Shift power budgets between cores and GPU
- Operating system/system vendors
  - Set given TDP
  - Monitor power consumption
- Users:
  - Monitor power consumption

RAPL – Mechanism

- Two power ranges:
  - Short term (typically 1 second?), typically exceeds TDP
  - Long term (typically 60 seconds?), typically TDP
- Processor must stay within both limits
- Internally sampled with typically (at least) 1 ms
- Processor can use free budgets for Turbo mechanisms
- Other facts:
  - Initially modelled, now measured
  - Also on AMD since Zen

Sysfs for intel_idle (and acpi_idle)

Sysfs Interface: /sys/devices/virtual/powercap/intel-rapl/intel-rapl:X/intel-rapl:0:Y

- X = number of package but this numbering does not reflect the package ID, this level contain the package domain information
- Y = 0 core domain, 1 gpu/uncore domain, 2 dram domain (the numbering can be different check the name)

- **name:** name of the domain and of the package ID
- **max_energy_range_uj:** range of the above energy counter in micro-joules
- **energy_uj:** current energy counter in micro joules
- **enabled:** enable/disable controls at domain level
- **constraint_0_name:** the name of the constraint 0 (usually long term window -> seconds)
- **constraint_0_time_window_us:** time window in micro seconds for the constraint 0
- **constraint_0_power_limit_uw:** power limit in micro watts for the constraint time 0
- **constraint_0_max_power_uw:** maximum allowed power in micro watts for the constraint 0
- **constraint_1_name:** the name of the constraint 1 (usually short term window -> milli seconds)
- **constraint_1_time_window_us:** time window in micro seconds for the constraint 0
- **constraint_1_power_limit_uw:** power limit in micro watts for the constraint time 1
- **constraint_1_max_power_uw:** maximum allowed power in micro watts for the constraint 0

TECHNISCHE UNIVERSITÄT DRESDEN

ZIH
Center for Information Services & High Performance Computing

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

CINECA

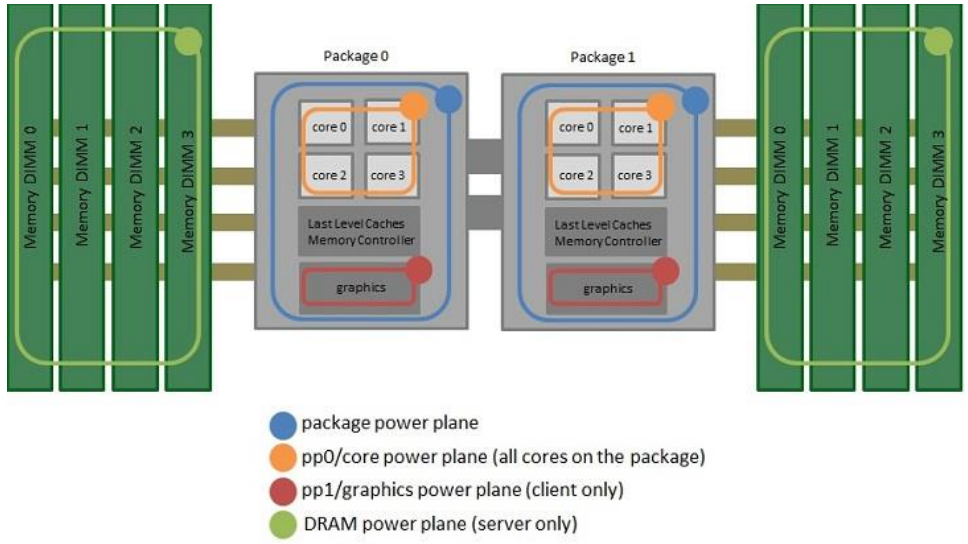
RAPL registers

PRACE

Running Average Power Limit RAPL

RAPL interfaces provide mechanisms to enforce power consumption limit. RAPL interfaces consist of non-architectural MSRs. RAPL expose multiple domains (power planes) of power rationing within each processor socket. Each RAPL domain supports the following set of capabilities:

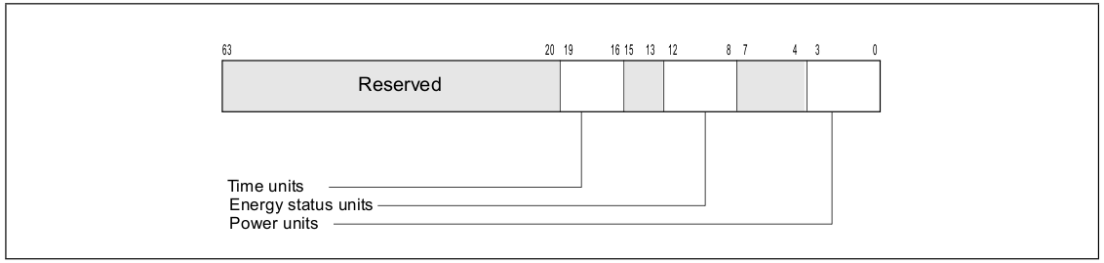
- **Power limit** - MSR interfaces to specify power limit, time window; lock bit, clamp bit etc.
- **Energy Status** - Power metering interface providing energy consumption information.
- **Policy** (Optional) - 4-bit priority information that is a hint to hardware for dividing budget between sub-domains in a parent domain.
- **Perf Status** (Optional) - Interface providing information on the performance effects (regression) due to power limits. It is defined as a duration metric that measures the power limit effect in the respective domain. The meaning of duration is domain specific.
- **Power Info** (Optional) - Interface providing information on the range of parameters for a given domain, minimum power, maximum power etc.



Each of the above capabilities requires specific units in order to describe them. Power is expressed in Watts, Time is expressed in Seconds, and Energy is expressed in Joules. Scaling factors are supplied to each unit to make the information presented meaningful in a finite number of bits.

RAPL – Time Units and Domains

Units for power, energy, and time are exposed in the read-only **MSR_RAPL_POWER_UNIT** (0x606) MSR.



- **Power Units** (bits 3:0): Power related information (in Watts) is based on the multiplier, **1/2^PU**; where PU is an unsigned integer represented by bits 3:0. Default value is 0011b, indicating power unit is in 1/8 Watts increment.
- **Energy Status Units** (bits 12:8): Energy related information (in Joules) is based on the multiplier, **1/2^ESU**; where ESU is an unsigned integer represented by bits 12:8. Default value is 10000b, indicating energy status unit is in 15.3 micro-Joules increment.
- **Time Units** (bits 19:16): Time related information (in Seconds) is based on the multiplier, **1/2^TU**; where TU is an unsigned integer represented by bits 19:16. Default value is 1010b, indicating time unit is in 976 microseconds increment.

RAPL support the following RAPL domain hierarchy: entire package (PKG), DRAM, power plane for cores (PP0) and power plane for uncore graphic device (PP1). Each level of the RAPL hierarchy provides a respective set of RAPL interface MSRs.

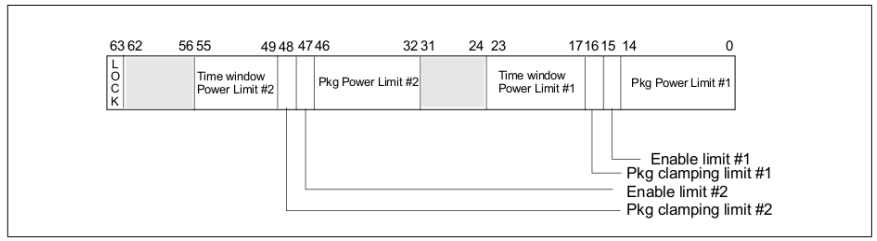
Domain	Power Limit (Offset 0)	Energy Status (Offset 1)	Policy (Offset 2)	Perf Status (Offset 3)	Power Info (Offset 4)
PKG	MSR_PKG_POWER_LIMIT	MSR_PKG_ENERGY_STATUS	RESERVED	MSR_PKG_PERF_STATUS	MSR_PKG_POWER_INFO
DRAM	MSR_DRAM_POWER_LIMIT	MSR_DRAM_ENERGY_STATUS	RESERVED	MSR_DRAM_PERF_STATUS	MSR_DRAM_POWER_INFO
PP0	MSR_PP0_POWER_LIMIT	MSR_PP0_ENERGY_STATUS	MSR_PP0_POLICY	MSR_PP0_PERF_STATUS	RESERVED
PP1	MSR_PP1_POWER_LIMIT	MSR_PP1_ENERGY_STATUS	MSR_PP1_POLICY	RESERVED	RESERVED

NOTE: PP1 not present in server architectures
DRAM present only in server architectures

RAPL – Package Domain

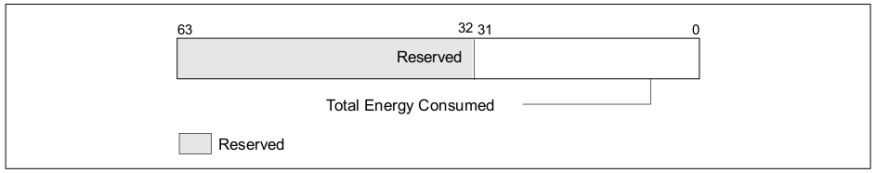
MSR_PKG_POWER_LIMIT allows a software agent to define power limitation for the package domain. Power limitation is defined in terms of average power usage (Watts) over a time window. Two power limits can be specified, corresponding to time windows of different sizes. Each power limit provides independent clamping control that would permit the processor cores to go below OS-requested state to meet the power limits. A lock mechanism allow the software agent to enforce power limit settings. Once the lock bit is set, the power limit settings are static and un-modifiable until next RESET.

- **Package Power Limit**: Sets the average power usage limit of the package domain corresponding to related time window. The unit of this field is specified by the "Power Units" field of MSR_RAPL_POWER_UNIT
- **Enable Power Limit**: 0 = disabled; 1 = enabled
- **Package Clamping Limitation**: Allow going below OS-requested P/T state setting during time window.
- **Time Window for Power Limit**: Indicates the time window for power limit.
  Time limit = 2^Y * (1.0 + Z/4.0) * Time Unit.
  Here "Y" is the unsigned integer value represented by bits 21:17– 53:49, "Z" is an unsigned integer represented by bits 23:22 – 55:54. "Time Unit" is specified by the "Time Units" field of MSR_RAPL_POWER_UNIT.
- **Lock**: If set, all write attempts to this MSR are ignored until next RESET.



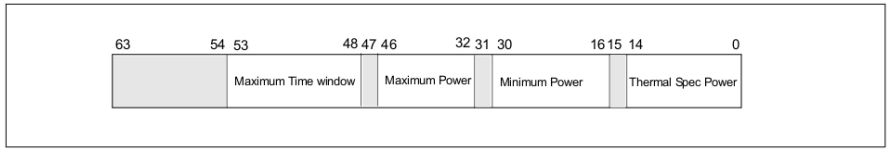
MSR_PKG_ENERGY_STATUS is a read-only MSR. It reports the actual energy use for the package domain. This MSR is updated every ~1msec. It has a wraparound time of around 60 secs when power consumption is high, and may be longer otherwise.

- **Total Energy Consumed**: The unsigned integer value represents the total amount of energy consumed since that last time this register is cleared. The unit of this field is specified by the "Energy Status Units" field of MSR_RAPL_POWER_UNIT.



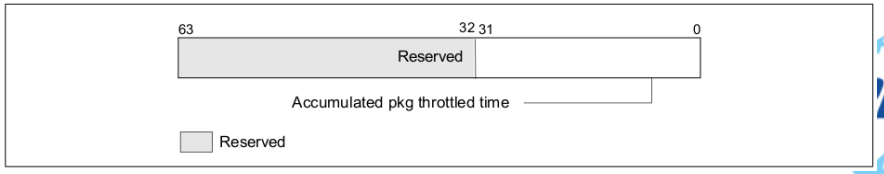
MSR_PKG_POWER_INFO is a read-only MSR. It reports the package power range information for RAPL usage. This MSR provides maximum/minimum values (derived from electrical specification) thermal specification power of the package domain. It also provides the largest possible time window for software to program the RAPL interface.

- **Thermal Spec Power**: The unsigned integer value is the equivalent of thermal specification power of the package domain. The unit of this field is specified by the "Power Units" field of MSR_RAPL_POWER_UNIT.
- **Minimum Power**: The unsigned integer value is the equivalent of minimum power derived from electrical spec of the package domain. The unit of this field is specified by the "Power Units" field of MSR_RAPL_POWER_UNIT.
- **Maximum Power**: The unsigned integer value is the equivalent of maximum power derived from the electrical spec of the package domain. The unit of this field is specified by the "Power Units" field of MSR_RAPL_POWER_UNIT.
- **Maximum Time Window**: The unsigned integer value is the equivalent of largest acceptable value to program the time window of MSR_PKG_POWER_LIMIT. The unit of this field is specified by the "Time Units" field of MSR_RAPL_POWER_UNIT.



MSR_PKG_PERF_STATUS is a read-only MSR. It reports the total time for which the package was throttled due to the RAPL power limits. Throttling in this context is defined as going below the OS-requested P-state or T-state. It has a wrap-around time of many hours.

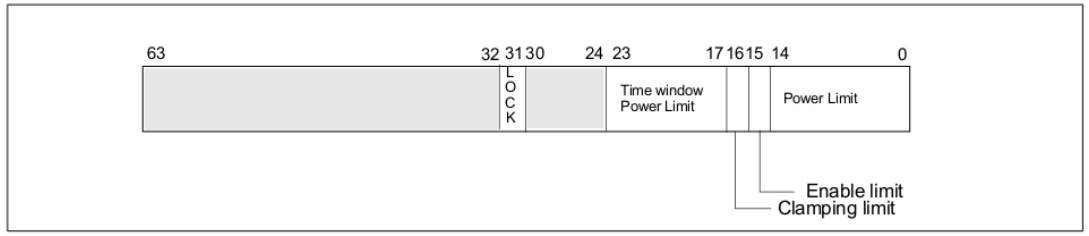
- **Accumulated Package Throttled Time**: The unsigned integer value represents the cumulative time (since the last time this register is cleared) that the package has throttled. The unit of this field is specified by the "Time Units" field of MSR_RAPL_POWER_UNIT.



RAPL – PP0/PP1 Domain

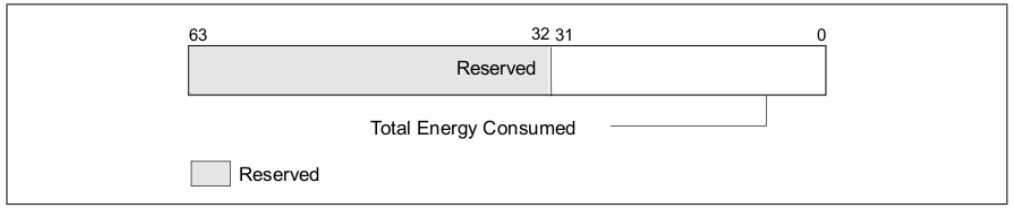
MSR_PP0_POWER_LIMIT/MSR_PP1_POWER_LIMIT allow a software agent to define power limitation for the respective power plane domain. A lock mechanism in each power plane domain allows the software agent to enforce power limit settings independently. Once a lock bit is set, the power limit settings in that power plane are static and un-modifiable until next RESET.

- **Power Limit**: Sets the average power usage limit of the respective power plane domain. The unit of this field is specified by the "Power Units" field of MSR_RAPL_POWER_UNIT.
- **Enable Power Limit**: 0 = disabled; 1 = enabled.
- **Clamping Limitation**: Allow going below OS-requested P/T state setting during time window.
- **Time Window for Power Limit**: Indicates the length of time window over which the power limit will be used by the processor. The numeric value encoded by bits 23:17 is represented by the product of 2^Y *F; where F is a single-digit decimal floating-point value between 1.0 and 1.3 with the fraction digit represented by bits 23:22, Y is an unsigned integer represented by bits 21:17. The unit of this field is specified by the "Time Units" field of MSR_RAPL_POWER_UNIT.
- **Lock**: If set, all write attempts to the MSR and corresponding policy MSR_PP0_POLICY/MSR_PP1_POLICY are ignored until next RESET.



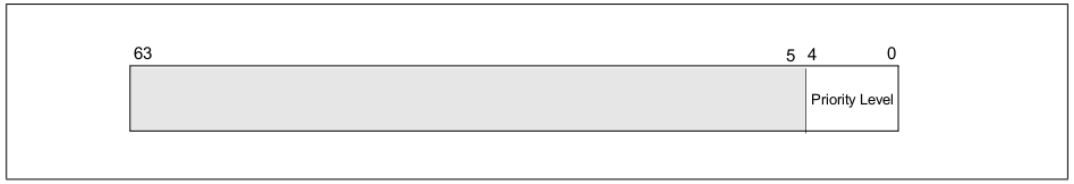
MSR_PP0_ENERGY_STATUS/MSR_PP1_ENERGY_STATUS are read-only MSRs. They report the actual energy use for the respective power plane domains. These MSRs are updated every ~1msec.

- **Total Energy Consumed**: The unsigned integer value represents the total amount of energy consumed since the last time this register was cleared. The unit of this field is specified by the "Energy Status Units" field of MSR_RAPL_POWER_UNIT.



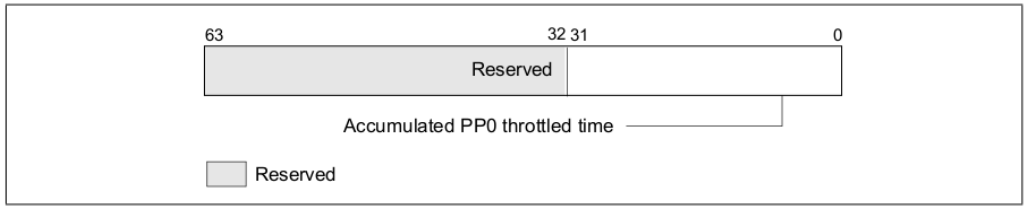
MSR_PP0_POLICY/MSR_PP1_POLICY provide balance power policy control for each power plane by providing inputs to the power budgeting management algorithm. On platforms that support PP0 (IA cores) and PP1 (uncore graphic device), the default values give priority to the non-IA power plane. These MSRs enable the PCU to balance power consumption between the IA cores and uncore graphic device.

- **Priority Level**: Priority level input to the PCU for respective power plane. PP0 covers the IA processor cores, PP1 covers the uncore graphic device. The value 31 is considered highest priority.



MSR_PP0_PERF_STATUS is a read-only MSR. It reports the total time for which the PP0 domain was throttled due to the power limits. This MSR is supported only in server platform. Throttling in this context is defined as going below the OS-requested P-state or T-state.

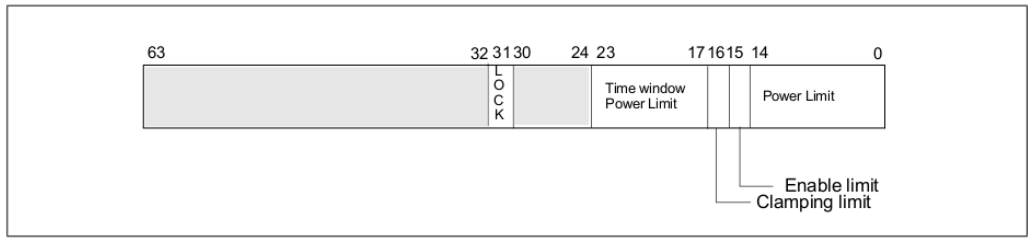
- **Accumulated PP0 Throttled Time**: The unsigned integer value represents the cumulative time (since the last time this register is cleared) that the PP0 domain has throttled. The unit of this field is specified by the "Time Units" field of MSR_RAPL_POWER_UNIT.



RAPL – DRAM Domain

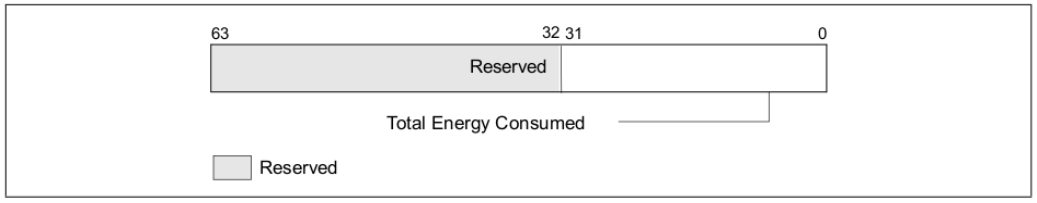
MSR_DRAM_POWER_LIMIT allows a software agent to define power limitation for the DRAM domain. Power limitation is defined in terms of average power usage (Watts). A power limit can be specified along with a time window. A lock mechanism allow the software agent to enforce power limit settings. Once the lock bit is set, the power limit settings are static and un modifiable until next RESET.

- **DRAM Power Limit**: Sets the average power usage limit of the DRAM domain corresponding to time window. The unit of this field is specified by the "Power Units" field of MSR_RAPL_POWER_UNIT.
- **Enable Power Limit**: 0 = disabled; 1 = enabled.
- **Time Window for Power Limit**: Indicates the length of time window over which the power limit will be used by the processor. The numeric value encoded by bits 23:17 is represented by the product of 2^Y *F; where F is a single-digit decimal floating-point value between 1.0 and 1.3 with the fraction digit represented by bits 23:22, Y is an unsigned integer represented by bits 21:17. The unit of this field is specified by the "Time Units" field of MSR_RAPL_POWER_UNIT.
- **Lock**: If set, all write attempts to this MSR are ignored until next RESET.



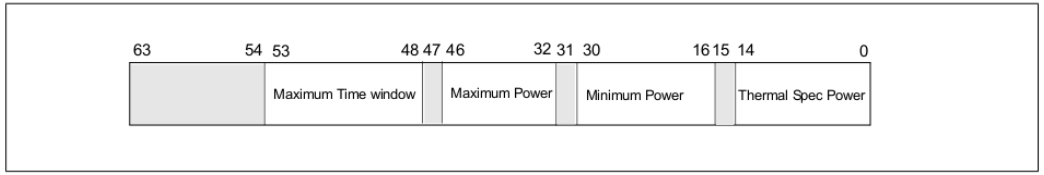
MSR_DRAM_ENERGY_STATUS is a read-only MSR. It reports the actual energy use for the DRAM domain. This MSR is updated every ~1msec.

- **Total Energy Consumed**: The unsigned integer value represents the total amount of energy consumed since that last time this register is cleared. The unit of this field is specified by the "Energy Status Units" field of MSR_RAPL_POWER_UNIT.



MSR_DRAM_POWER_INFO is a read-only MSR. It reports the DRAM power range information for RAPL usage. This MSR provides maximum/minimum values (derived from electrical specification) thermal specification power of the DRAM domain. It also provides the largest possible time window for software to program the RAPL interface.

- **Thermal Spec Power**: The unsigned integer value is the equivalent of thermal specification power of the DRAM domain. The unit of this field is specified by the "Power Units" field of MSR_RAPL_POWER_UNIT.
- **Minimum Power**: The unsigned integer value is the equivalent of minimum power derived from electrical spec of the DRAM domain. The unit of this field is specified by the "Power Units" field of MSR_RAPL_POWER_UNIT.
- **Maximum Power**: The unsigned integer value is the equivalent of maximum power derived from the electrical spec of the DRAM domain. The unit of this field is specified by the "Power Units" field of MSR_RAPL_POWER_UNIT.
- **Maximum Time Window**: The unsigned integer value is the equivalent of largest acceptable value to program the time window of MSR_DRAM_POWER_LIMIT. The unit of this field is specified by the "Time Units" field of MSR_RAPL_POWER_UNIT.



MSR_DRAM_PERF_STATUS is a read-only MSR. It reports the total time for which the package was throttled due to the RAPL power limits. Throttling in this context is defined as going below the OS-requested P-state or T-state. It has a wrap-around time of many hours.

- **Accumulated Package Throttled Time**: The unsigned integer value represents the cumulative time (since the last time this register is cleared) that the DRAM domain has throttled. The unit of this field is specified by the "Time Units" field of MSR_RAPL_POWER_UNIT.