A Survey on Energy-Efficiency in GPUs

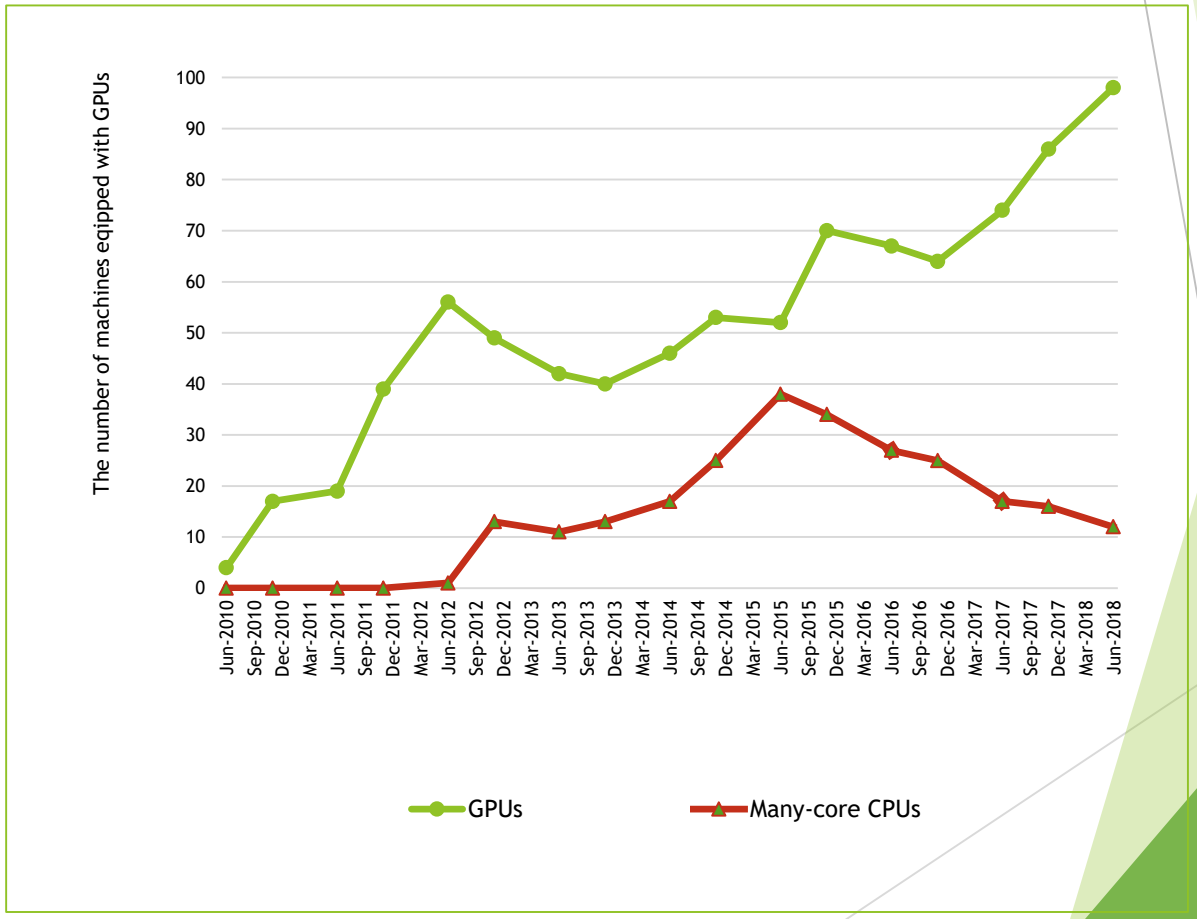
By: Ehsan Sharifi Esfahani

Outlines

- Upward trend of using accelerator in supercomputers
- An Argument about TOP500 website
- Motivations
- Challenges
- Source of energy consumption in GPUs
- Energy efficiency metrics
- Generalization of energy proportionality curve
- Energy Measuring
- Our taxonomies and classifications
- DVFS technique features in GPUs
- Other proposed solutions
- Conclusion and Future work

Upward trend of using GPUs in supercomputers

- Accelerators, such as GPUs, and coprocessors, such as Intel Xeon Phi are two combinations with CPU to build supercomputers.
- There is more interest to GPU instead of many-cores
- The combinations of CPU-GPU is more efficient than traditional many-core systems

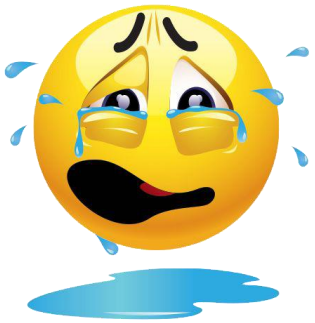


An Argument about TOP500 website

- Is really the numbers in TOP500 website precise and is it a proper referenceable source for academic papers?
  - Maybe !!! 
- Why?
  - We could find contradictions between available numbers
  - The available numbers are being revised !!!
- Why researchers refer to this numbers in the majority of academic and high-citation papers?
  - There is no other alternative !!! 

Motivations

- Energy-efficiency in GPUs has not been studied enough
- A lot of energy inefficient applications
- In some applications, high energy consumption is a bottleneck, not the absolute performance.
- High energy consumption →more heat dissipation → increasing hardware temperature → increasing cooling costs, decreasing reliability and scalability
- Make possible to build exascale future machines
  - High energy consumption and running costs are two of the main challenges
- Environmental consequences.
  - CO2 emission form data centers worldwide is estimated to increase from 80 Megatons (MT) in 2007 to 340 MT in 2020, more than double the amount of current CO2 emission in the Netherlands (145 MT).

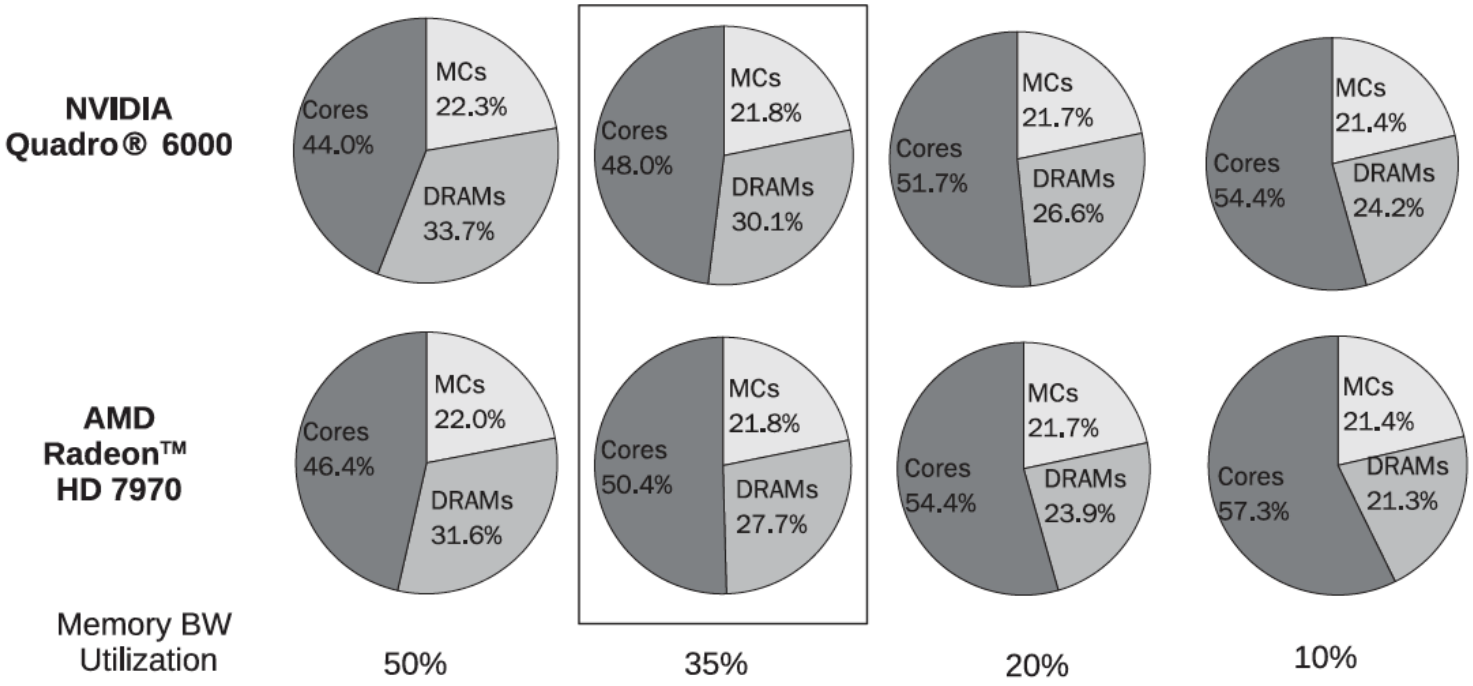


Challenges

- We cannot apply the energy consumption reduction methods in CPU to GPU
- Diverse and progressing quickly of GPU technologies and architectures
  - We cannot apply the same methodology on different generation of GPUs.
- Lack of accurate estimation and simulation tools for performance/energy.
- Complication of defining an accurate energy model
- In some cases, trade-off between performance and energy-efficiency.
  - So, in a multi-objective environments put more complexities in the proposed solutions since we should make a balance between these two conflicting goals.
- Lack of information about GPU hardware and its power management

Source of energy consumption in GPUs

- The most significant energy usage in GPU is caused by processing units and caches, and memory.

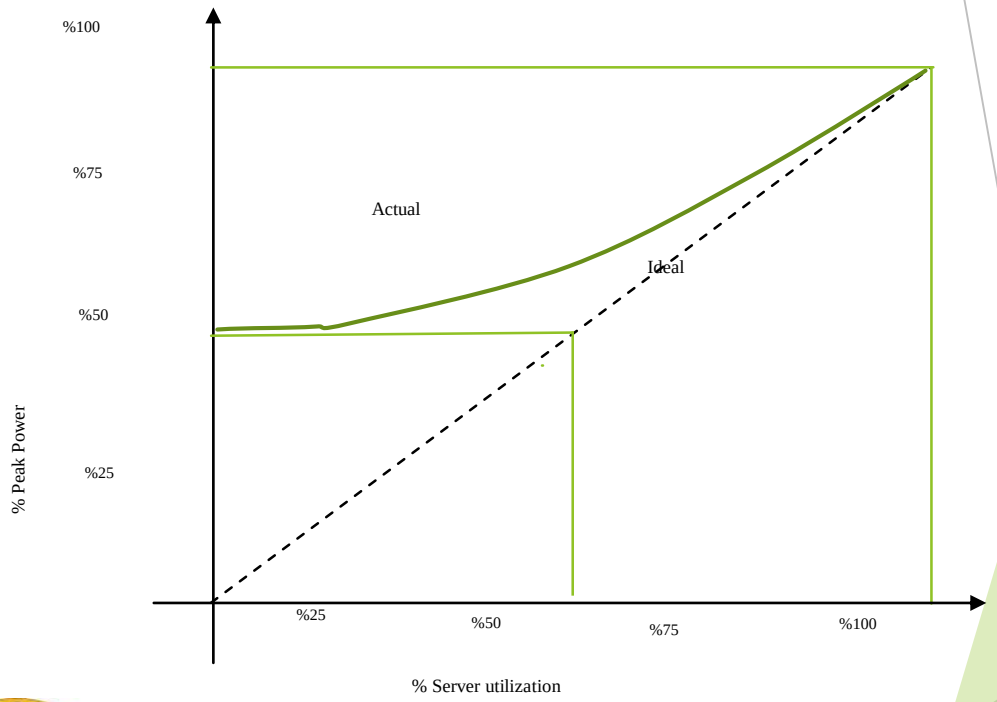


Energy efficiency metrics

- **Performance/watt**, number of operations per each watt
  - To compare the energy efficiency of different machines, or algorithms.
- **Power** is the rate of consuming energy while **energy** is summation of power consumed during a period.
- **Energy Delay product (EDP)** and **Energy Delay squared product (E2DP)**
  - They used to take into account both of these metrics together when there is trade-off.

Generalization of energy proportionality curve

- The main source of energy usage has been trending to GPU
  - Summit, each node has 6 GPUs with totally 1800 watt.
- There is a range of energy consumption for GPU
  - For instance, NVIDIA Tesla V100 (96, 300)
- $EP = 1 - \frac{Area_{actual} - Area_{ideal}}{Area_{ideal}}$



An argument

- It is generally believed that there is a trade-off between energy-efficiency and performance in parallel applications
  - Is this really correct in GPU environments? Not always
- They can support each other as well.
  - such as using less barriers

Power measuring method : 1 - Energy Models

- Empirical
  - A bottom-top method based on the underlying hardware

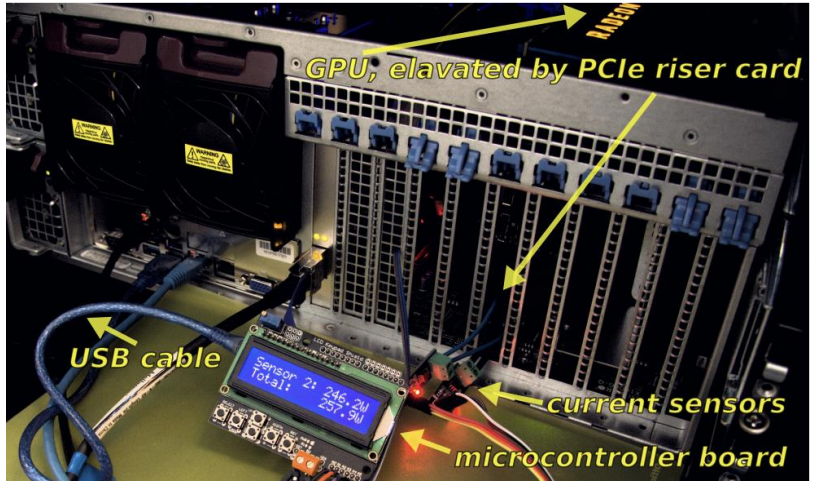
$$P_{GPU} = \sum_{i=1}^{n} P_i$$

$$E_{application} = \int_{t1}^{t2} P_{GPU} d_t$$

- Statistical
  - Machine learning and analytical techniques used to find a relationship between GPU power consumption and performance independent of the underlying hardware

Power measuring method : 2- External sensor power

- Needs physical access to the system
- Low sampling rate
- Less scalable and portable since it needs extra hardware
- Coarse-grain power profiling
- Lack of available tools in the market for some specific HPC systems.



Power measuring method : 3- Internal sensor power

- Current area of research
- Disadvantages
  - The way of obtaining power is unknown for us due to lack of documentations about them.
  - Low sampling frequency.
  - Inaccurate measurement
- Advantages
  - Available
  - Easy to use
  - No extra expenditures

Our taxonomies and classifications

- Hardware-based and Software-based
- Thermal-aware and energy-aware
  - Thermal-aware solutions take temperature as a core component when building an energy model
  - The temperature depends on the power consumption of GPU, dimension of GPU card, and relative location of the GPU and so forth.
- Single and composite
- Online and offline
    - Every online proposed approach put an overload on our computing system, thereby increasing energy consumption. The energy saving gained by our solution must outweigh the added energy consumption caused by it.

DVFS technique features in GPUs

- DVFS was the most common studied method
- GPU provide better environment to apply DVFS technique
  - The peak power consumption of a modern GPU is almost double that of the common modern CPU.
  - The frequencies of GPUs do not only have a larger range than CPUs, they are also more granular
- Applying DVFS in GPU is more complicated
  - We can scale working frequency of processing component and memory.
- DVFS definition voltage and frequency can vary, mostly frequency scaling is accessible to be changed by software.
  - There is no tool for scaling voltage, especially in Linux platform !!!

A few results

- **Theoretically:**
  - Compute-bounded
    - Increasing core frequency and decreasing memory frequency
  - Memory-bounded
    - Decreasing core frequency and Increasing memory frequency
  - Hybrid
    - Increasing both memory and core frequency
- **Practically:**
  - Predicting the best frequency and voltage in GPU is really complicated, it depends on the application type, underling hardware and measuring energy consumption method, problem size and input data.

Other proposed solutions and studies

- Energy Strong scaling
  - Total energy consumption remains constant for a fixed problem size when the number of processing unit increases.
    - matrix multiplication and n-body problem
- Energy consumption in GPU was influenced by two factors: how much the application is compute-bounded and how much the application is memory-bounded.
- Memory access pattern and the number of blocks in CUDA framework can impact energy efficiency
  - more memory access can increase energy consumption
- Increasing warp occupancy
  - Number of blocks and threads per blocks in CUDA environment can impact energy consumption.

Other proposed solutions and studies

- Warp scheduler can impact energy consumptions
- Hard-ware based Code compression in the communications links with less toggle
- Neighboring concurrent thread arrays usually use a large amount of shared data.
  - The GPU scheduler distributed these threads in a round-robin fashion among the SMs to achieve better load balancing, thereby increasing data replication in L1 cache.
  - To synchronize, we need more data movements and it causes less power-efficiency and performance.
  - A new scheduler can improve performance and energy-efficiency.

Classifications of the studied proposed solutions

Classifications / Proposed Solutions	Thermal-aware	Energy-aware	Single	Composite	Online	Offline	Hardware-based	Software-based
Luk at al [36]	×	✓	×	✓	✓	×	×	✓
NVIDIA Co [32]	✓	×	✓	×	✓	×	×	✓
ElTantawy et al [41]	×	✓	×	✓	✓	×	✓	×
Wang et al [42]	×	✓	✓	×	×	✓	×	✓
Li et al [43]	×	✓	×	✓	✓	×	✓	×
Guerreiro et al [44]	×	✓	✓	×	×	✓	✓	×
Zhang el al [46]	×	✓	✓	×	✓	×	✓	×
Tabbakh et al [47]	×	✓	×	✓	✓	×	✓	×
Prakash et al [48]	✓	×	✓	×	✓	×	✓	×
Pekhimenko et al [49]	×	✓	×	✓	✓	×	✓	×

Conclusion and Future possible work

- **Conclusion**
  - There is an upward trend to equipped supercomputers with GPUs
  - GPUs are the main component of energy consumption in servers
  - Energy-efficiency in GPU is challenging

- **Future possible works**
  - Multi-GPU environment
  - Thermal-aware energy model in HPC context
  - Auto-tuning for energy-efficiency in GPUs
  - Generalizations of energy proportionally curve in GPUs

Thank you for your attentions

Any questions?