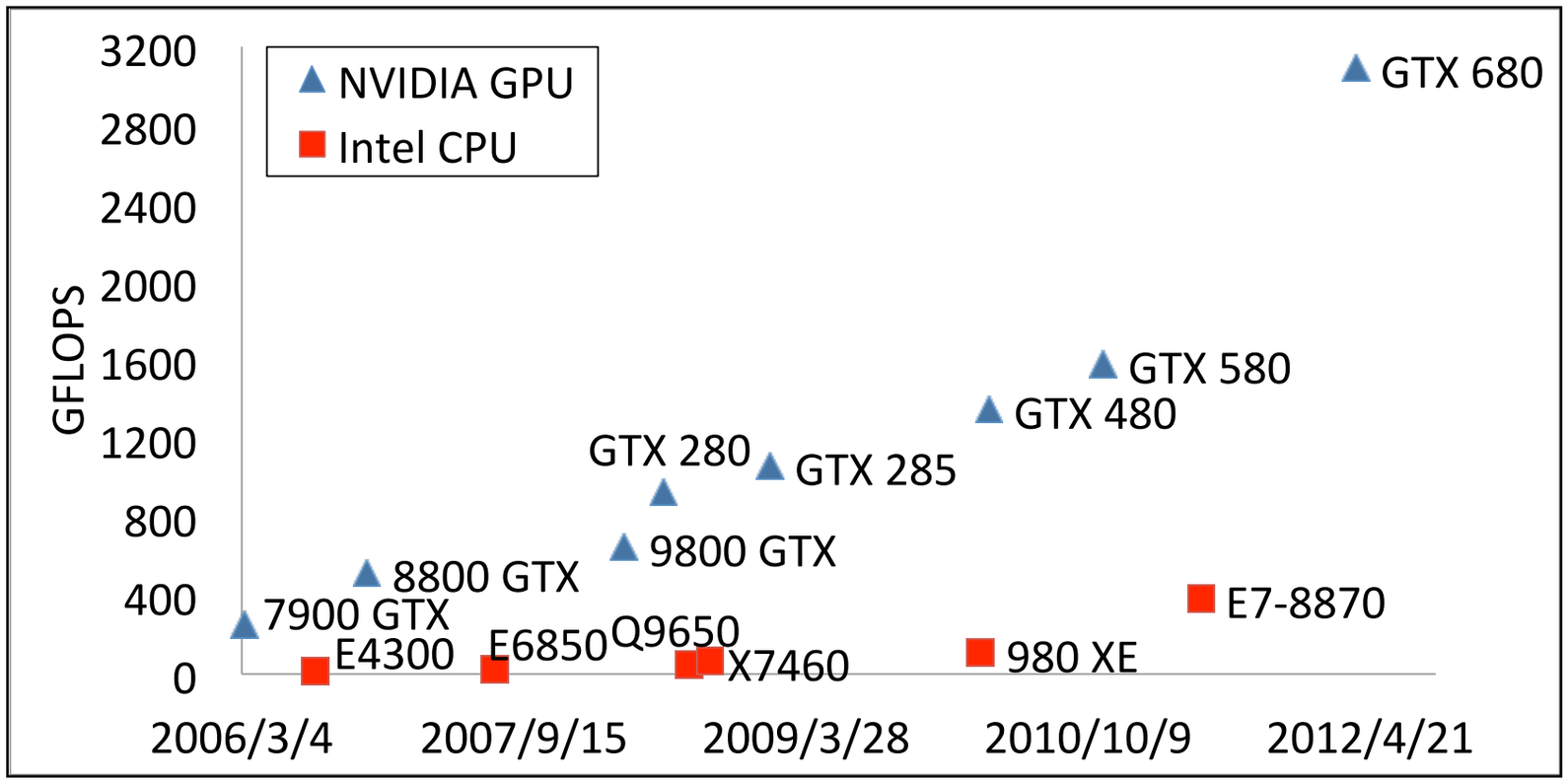
Power and Performance Analysis of GPU-Accelerated Systems

Yuki Abe*, Hiroshi Sasaki*, Martin Peres**,
Koji Inoue*, Kazuaki Murakami*, Shinpei Kato***

*Kyushu University
**Laboratoire Bordelais de Recherche en Informatique
***Nagoya University

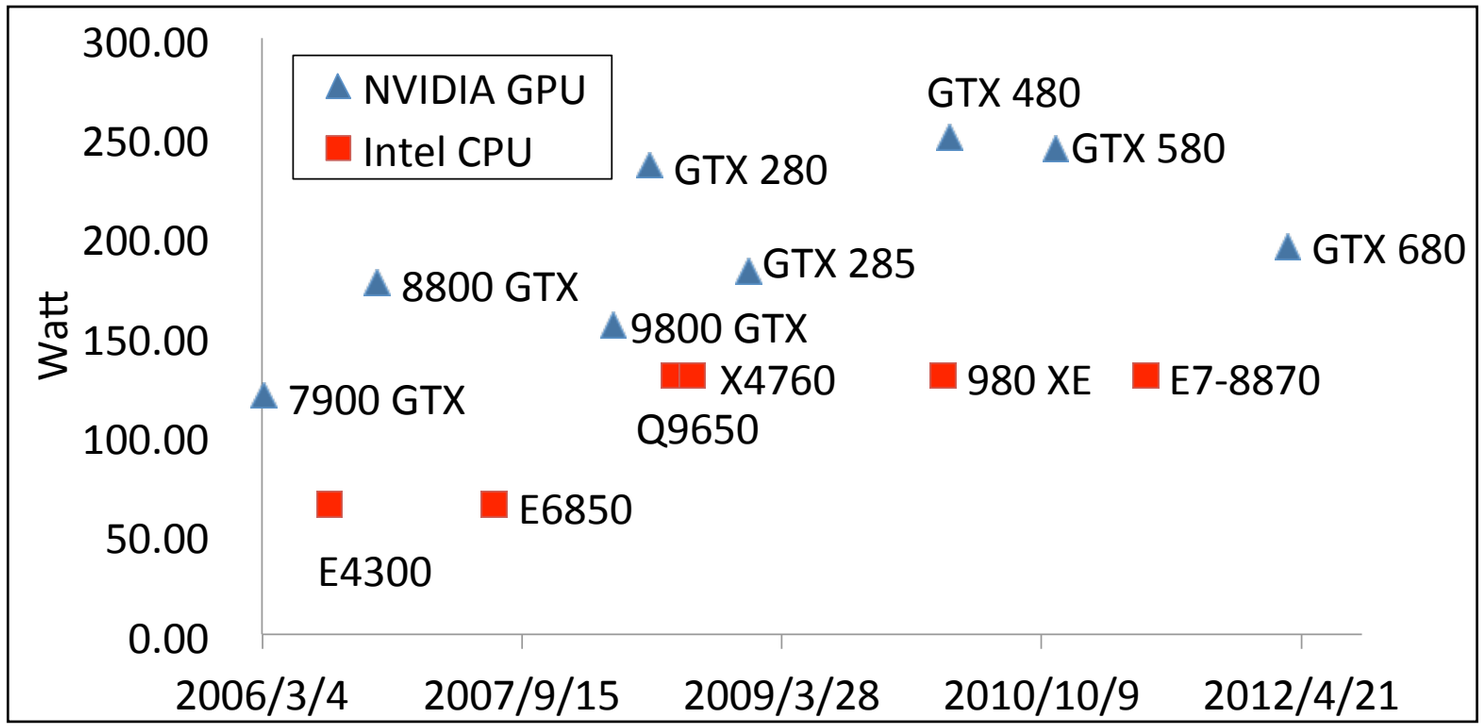
Graphics Processing Units (GPUs)

- GPUs have become popular
  - *Significant performance (peak performance of 3 TFLOPS for the latest Kepler GPUs)*
  - *Running general applications (GPGPU)*



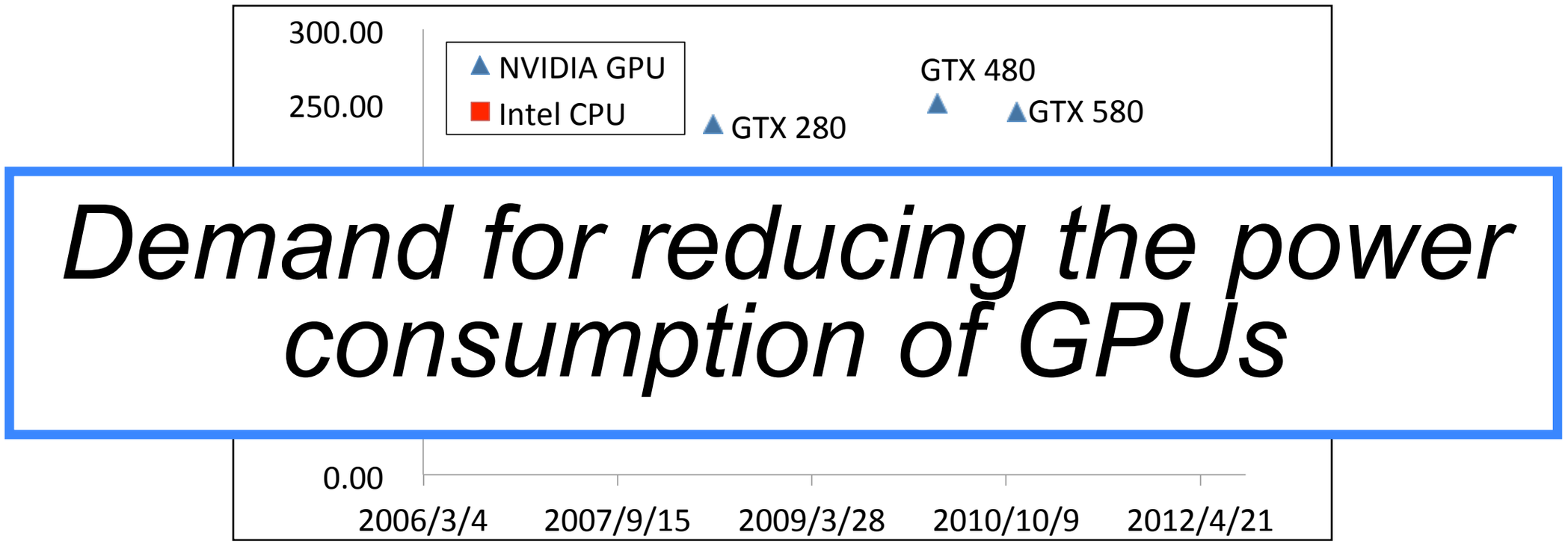
Power Consumption of GPUs

- Power consumption of most GPUs is higher than that of CPUs



Power Consumption of GPUs

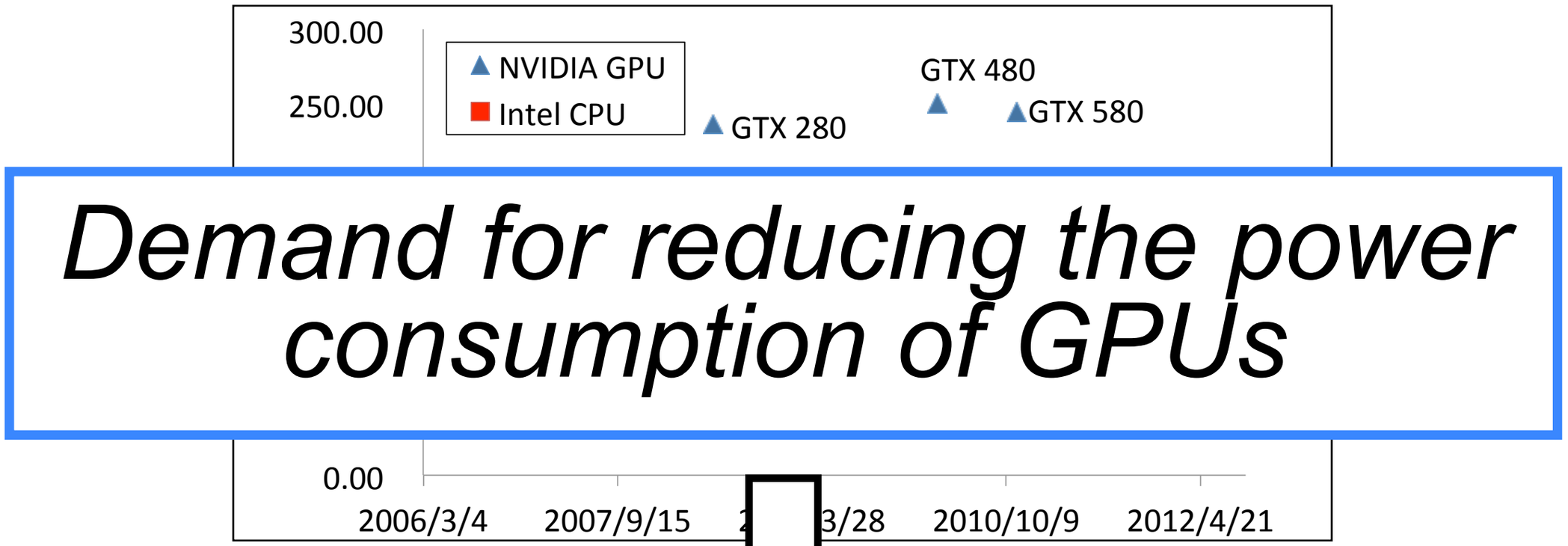
- Power consumption of most GPUs is higher than that of CPUs



Demand for reducing the power consumption of GPUs

Power Consumption of GPUs

- Power consumption of most GPUs is higher than that of CPUs



Demand for reducing the power consumption of GPUs

DVFS on GPUs

DVFS on GPU-Accelerated Systems

- DVFS is a popular way to reduce the power consumption of CPUs
- We answer to two questions through this study:
  - *Is **CPU** frequency scaling effective?*
  - *Is **GPU** frequency scaling effective?*

Experimental Setup

- GPU: NVIDIA GeForce GTX480
- CPU: Intel Core i5-2400
- OS: Linux Kernel : 3.3.0+

- Benchmark programs
  - *3 benchmark programs from Rodinia Benchmarks*
  - *Micro benchmark (Matrix Multiplication)*

Available Frequencies

- GPU frequencies

Clock Domain	Low [MHz]	High [MHz]
Core	405	700
Memory	324	1848

- CPU frequencies

Clock Domain	Low [MHz]	High [MHz]
Core	2700	3300.1

GPU Runtime and Driver

- NVIDIA proprietary software
  - *Change GPU's frequency by modifying BIOS file*
  - *Require to reload the driver when changing GPU's frequency*
- Gdev [Kato et al, USENIX ATC'12]
  - *Open-source runtime and driver*
  - *Allows the system to change GPU's frequency dynamically at runtime through the Linux "/proc" file system interface*
  - *The GPU memory frequency is fixed at 135MHz*

Measuring Power Consumption

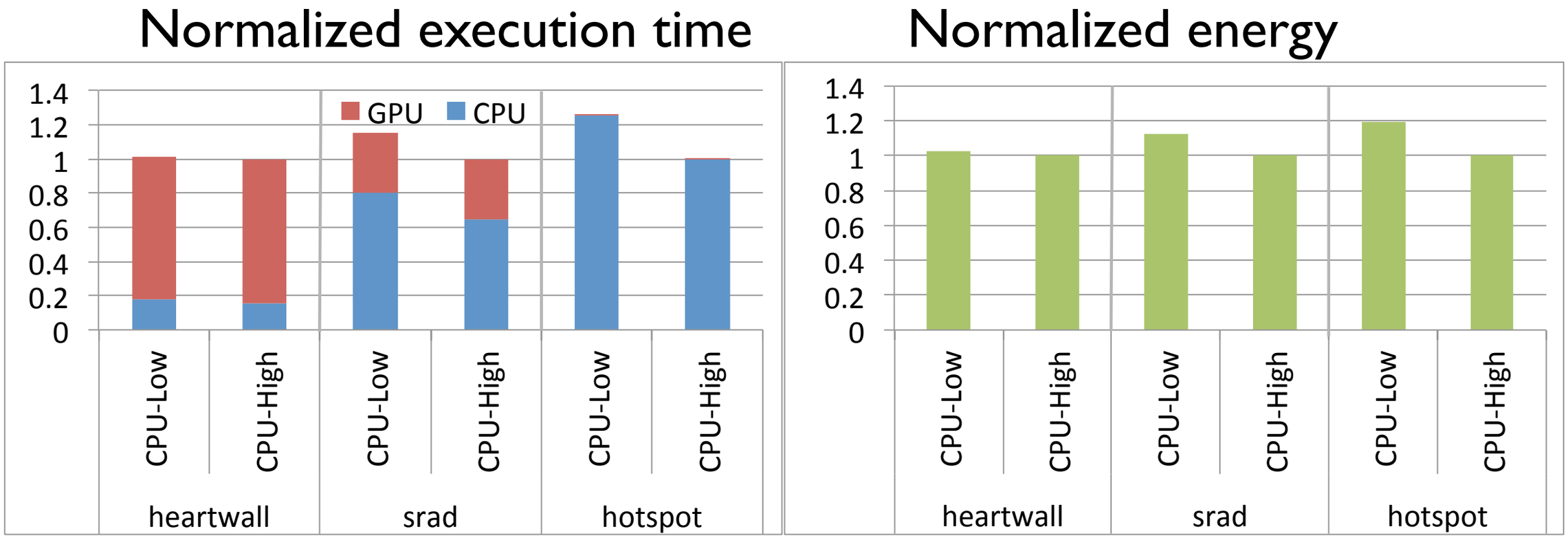
- Power meter: YOKOGAWA WT1600 Digital Power Meter
- Obtain the voltage and electric current from power plug of the machine
  - *Measure every 50 ms*

Plug in the power plug of the machine

Impact of CPU frequency scaling

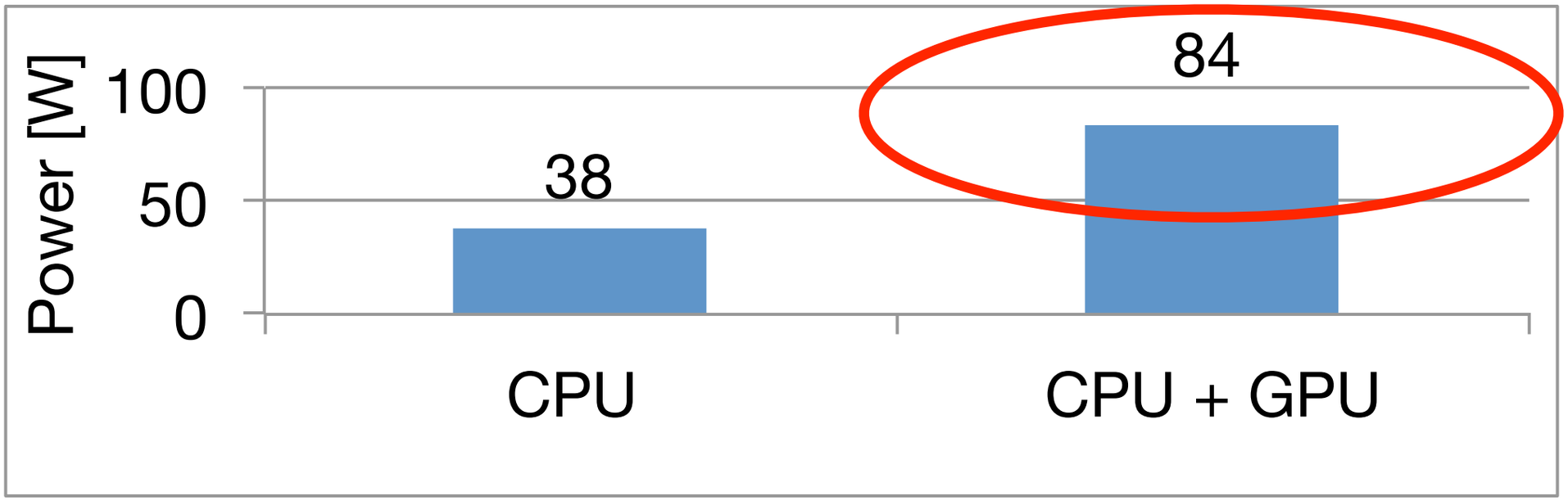
- Compare 2 frequency settings: (1) **CPU-High** and (2) **CPU-Low**
  - *CPU's clock is set to Low when idle*
  - *GPU's core clock is set to High when executing a CUDA kernel; otherwise Low*
- 3 benchmarks (heartwall, srad and hotspot) from Rodinia benchmarks
  - *CPU and GPU intensive workloads*

Evaluation Result



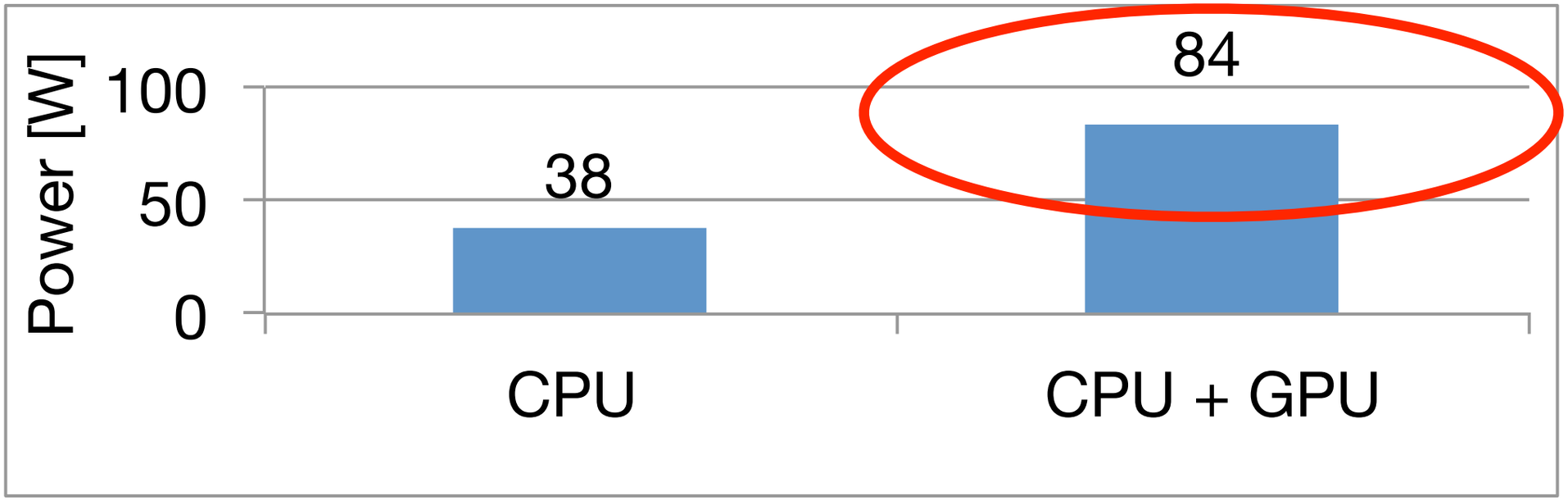
- Energy consumption can't be reduced with CPU-Low
- This is counter-intuitive considering CPU-only system

Idle Power



- Idle power consumption of GPU is larger than that of CPU
- Increased execution time in GPU-accelerated system wastes power

Idle Power



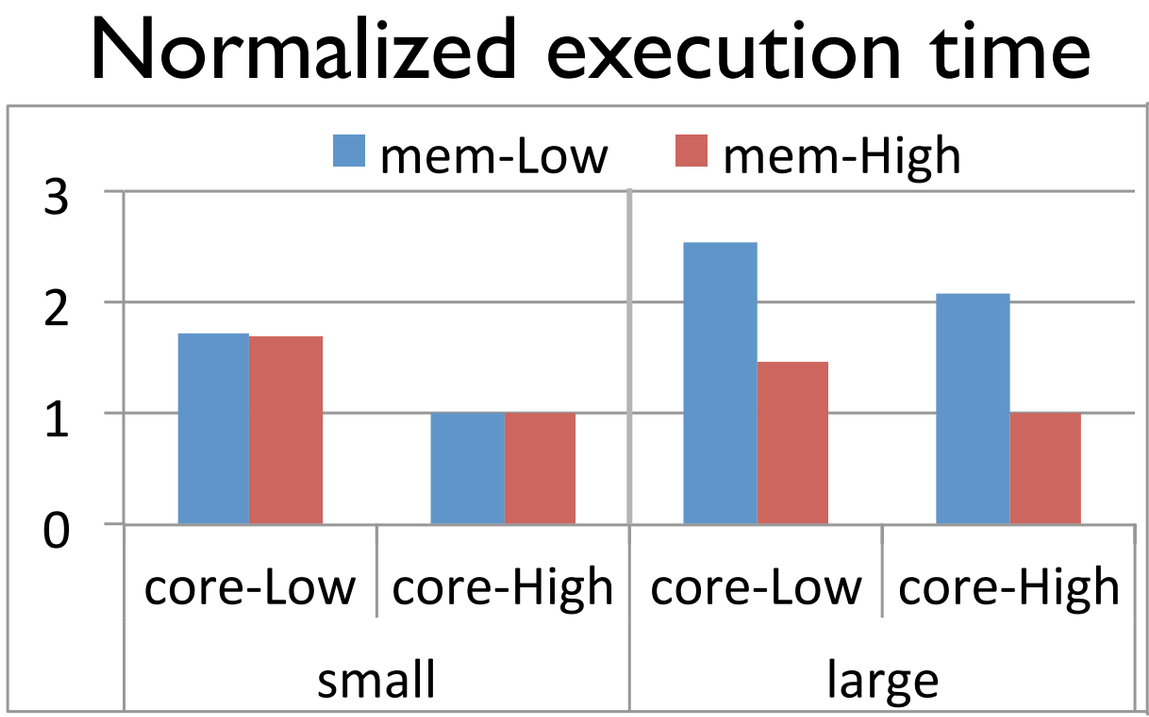
- Idle power consumption of GPU is lar
- Ind -accelerated system wastes power

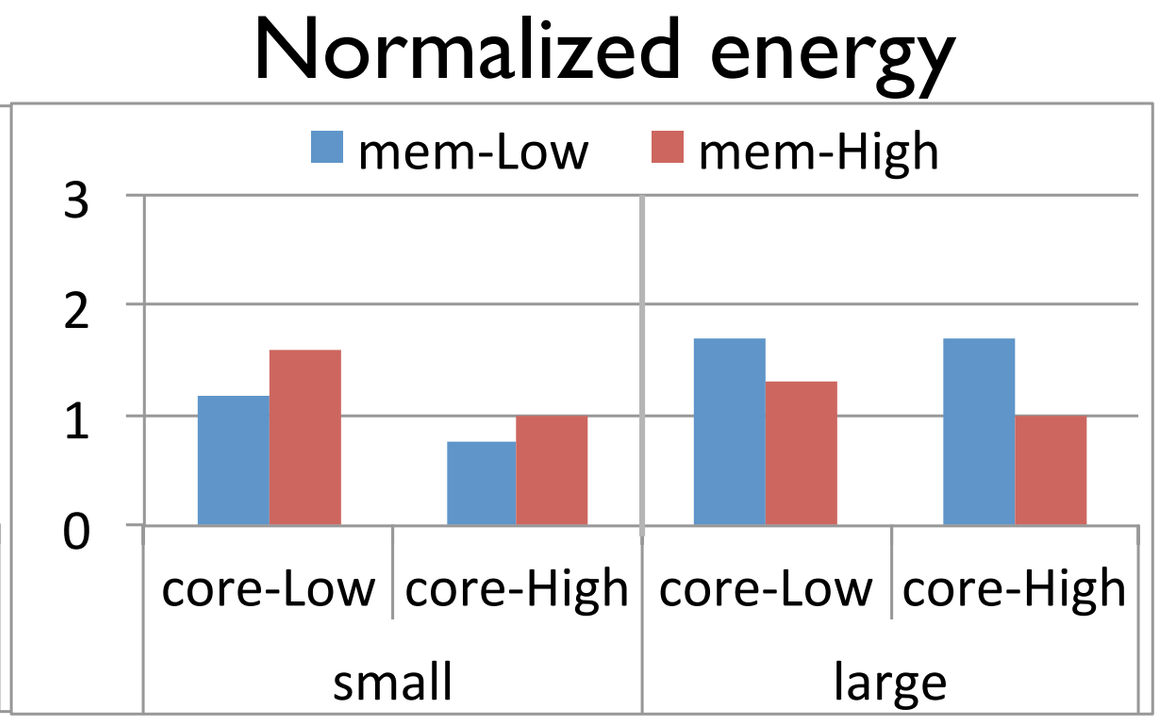
CPU is a weak factor

Impact of GPU frequency scaling

- Compare 4 frequency settings:
  (1) **Mem-High** and **Core-High**
  (2) **Mem-High** and **Core-Low**
  (3) **Mem-Low** and **Core-High**
  (4) **Mem-Low** and **Core-Low**
  - *CPU clock is always set to Low*
- Matrix Multiplication (small and large inputs)
  - *GPU intensive workloads*

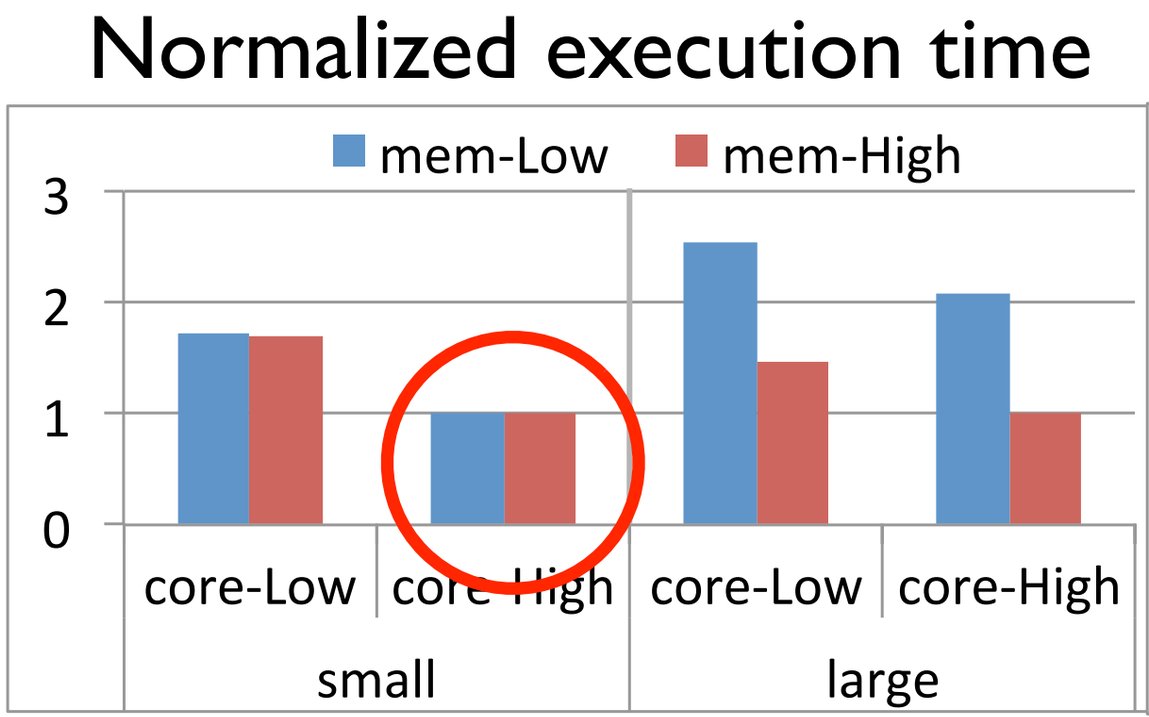
Evaluation Result

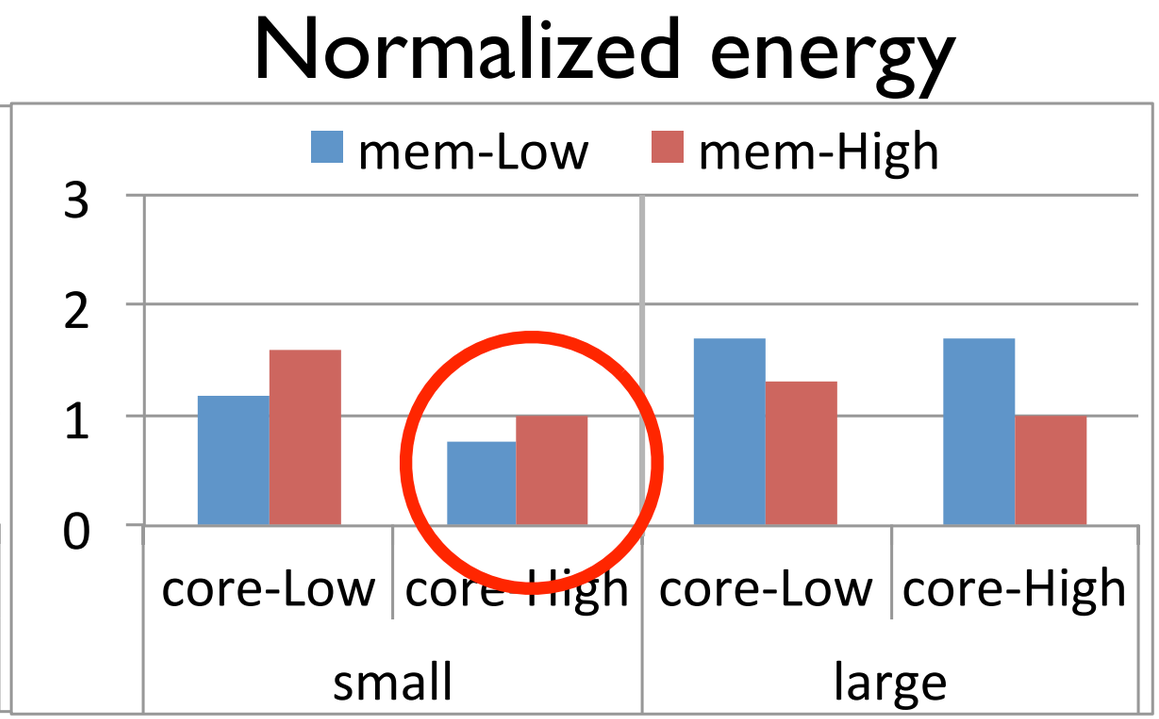




- When input size is small, the program is core bound
  - *Memory clock can be down-scaled retaining the performance*
- When input size is large, the program is core and memory bound
  - *GPU clocks cannot be down-scaled retaining the performance*

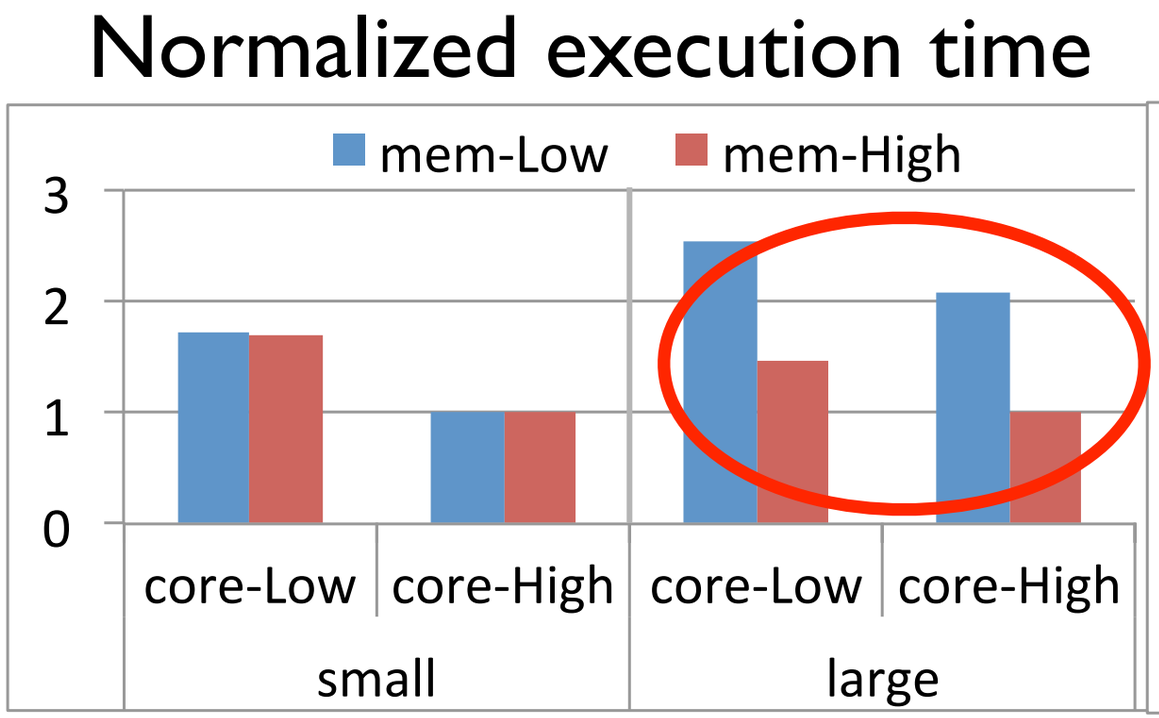
Evaluation Result

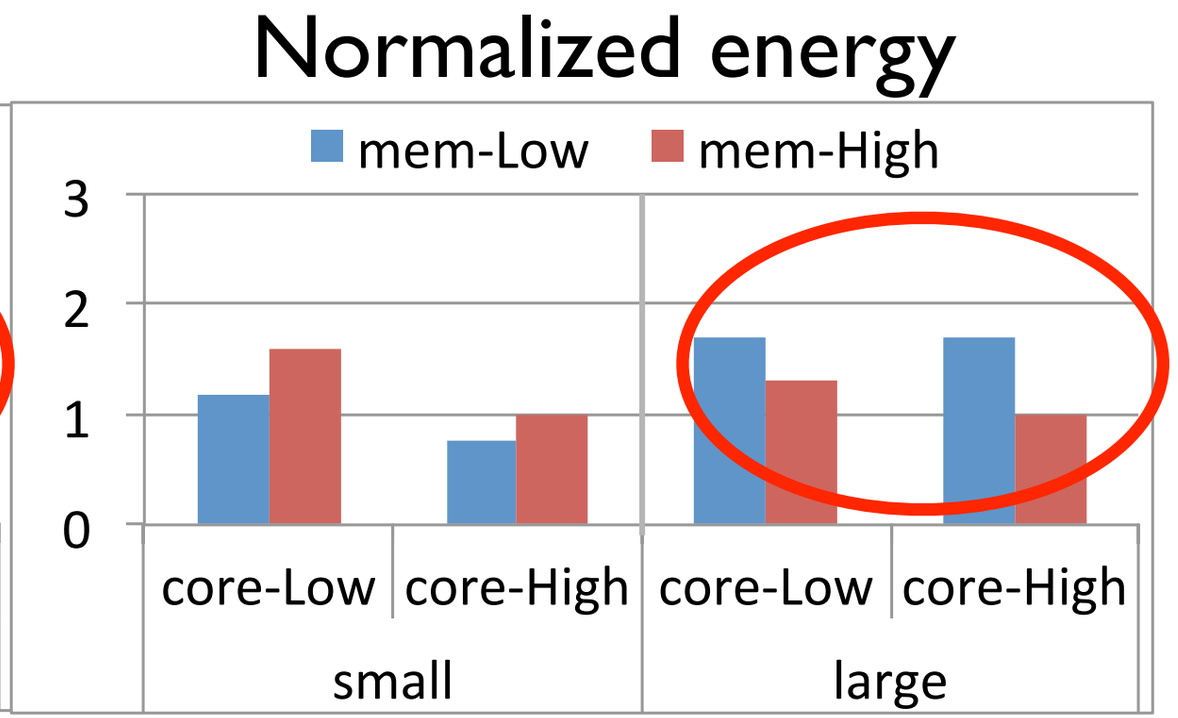




- When input size is small, the program is core bound
  - *Memory clock can be down-scaled retaining the performance*
- When input size is large, the program is core and memory bound
  - *GPU clocks cannot be down-scaled retaining the performance*

Evaluation Result





- When input size is small, the program is core bound
  - *Memory clock can be down-scaled retaining the performance*
- When input size is large, the program is core and memory bound
  - *GPU clocks cannot be down-scaled retaining the performance*

Conclusions

- CPU is a weak factor for energy savings of GPU-accelerated systems
- Effective voltage and frequency scaling of the GPU can reduce the power consumption retaining the performance

Thank you for your attention!