Comprehensive Comparison of Power-Performance Efficiency on Accelerators

Keitaro Oka, Yuichi Inadomi, Takatsugu Ono, and Koji Inoue
Kyushu University

Motivation

- High-throughput accelerates are commonly used

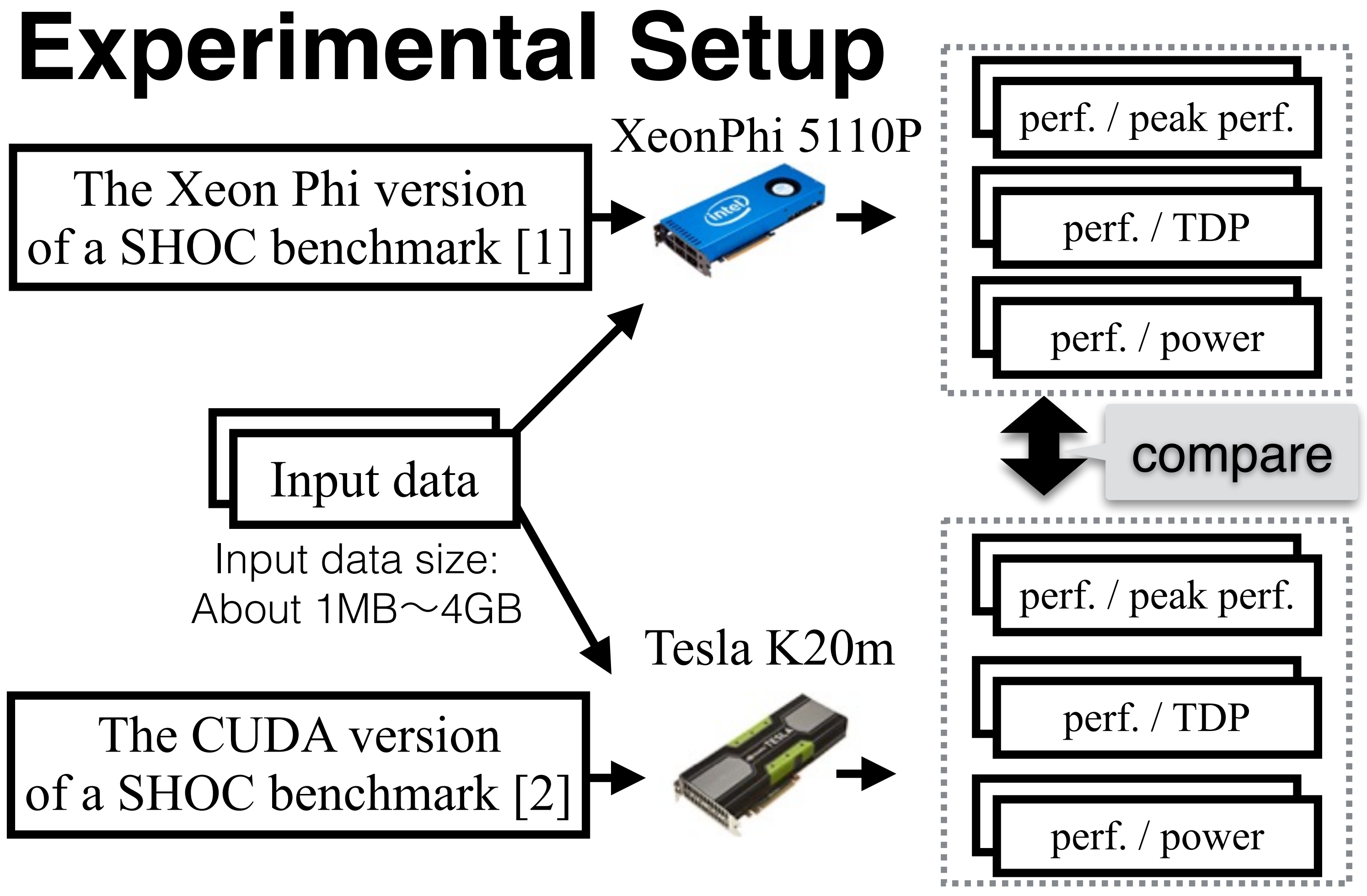
Intel Xeon Phi

NVIDIA Tesla GPU

The 2 platforms of attention

- Both platforms show the difference power-performance efficiency among workloads
- Which metric must be used on system design?
  - Performance / Peak performance
  - Performance / TDP
  - Power-performance efficiency, etc.

Experimental Setup



The Xeon Phi version of a SHOC benchmark [1]

XeonPhi 5110P

Input data

Input data size: About 1MB〜4GB

Tesla K20m

The CUDA version of a SHOC benchmark [2]

perf. / peak perf.

perf. / TDP

perf. / power

compare

perf. / peak perf.

perf. / TDP

perf. / power

[1] K. Spafford and R. Rahman, “Pre-release of SHOC for Intel Xeon Phi,” https://github.com/vetter/shoc-mic.

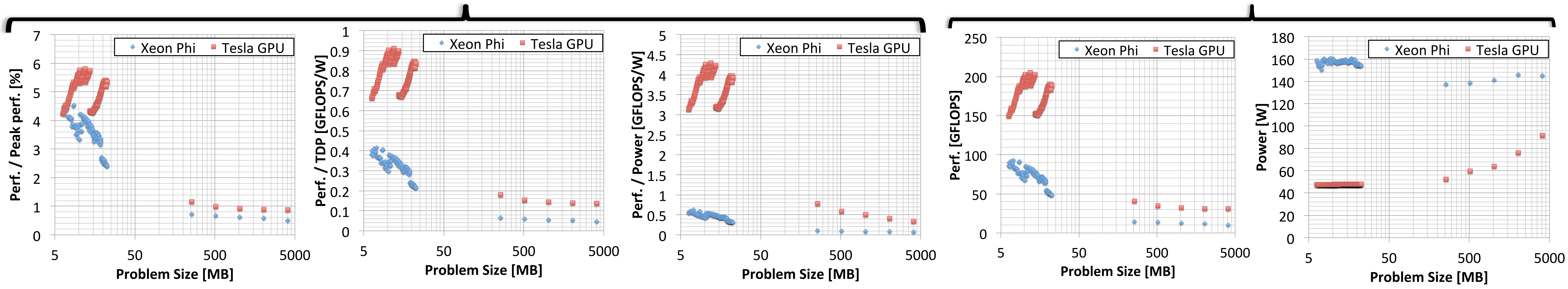
[2] A. Danalis et al., “The Scalable HeterOgeneous Computing (SHOC) benchmark suite,” in Proc. 3rd Workshop on GPGPU, 2010.

Comparison results

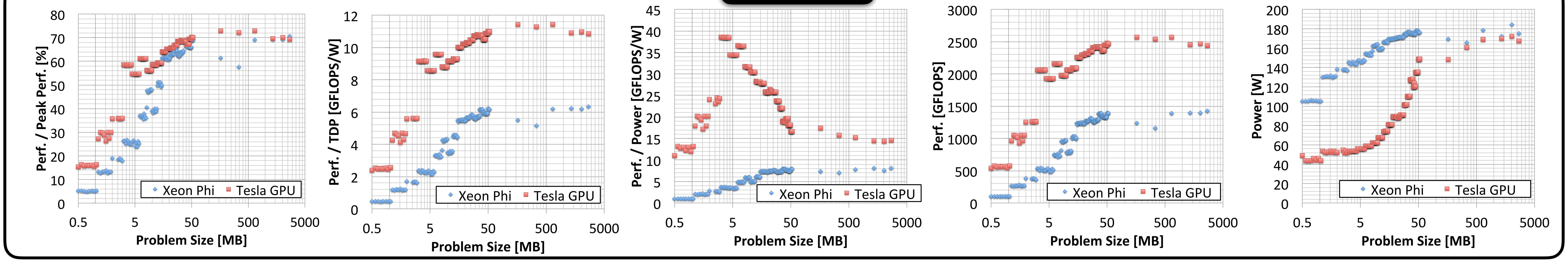
MD

The main metric for comparison

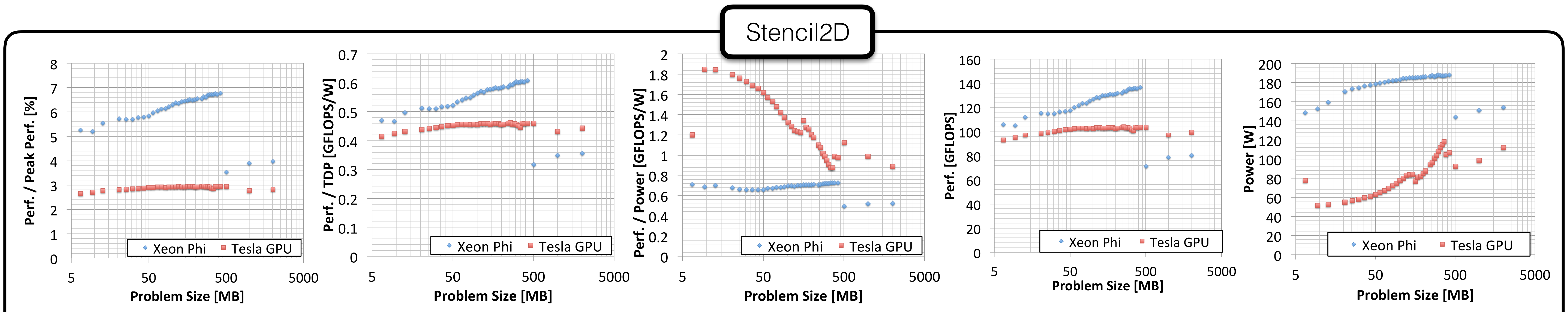
The metric for discussion



GEMM



Stencil2D



Discusion

- Upper and lower relationship between Tesla and Phi across the three main metrics
  - The same in MD and GEMM
  - Different in Stencil2D between Performance/Power and the other metrics
  - ➡ In some case, Performance/Power should be considered on system design
- Similarity of the trend for varying input sizes across the three main metrics
  - Similar in MD
  - Different in GEMM and Stencil2D. Performance/Power on Tesla rapidly decreases in larger input sizes
  - ➡ Because the power consumption of Tesla rapidly increases for varying input sizes while the performance doesn’t