PeaPaw: Performance and Energy Aware Workload Partitioning on Heterogeneous Platforms

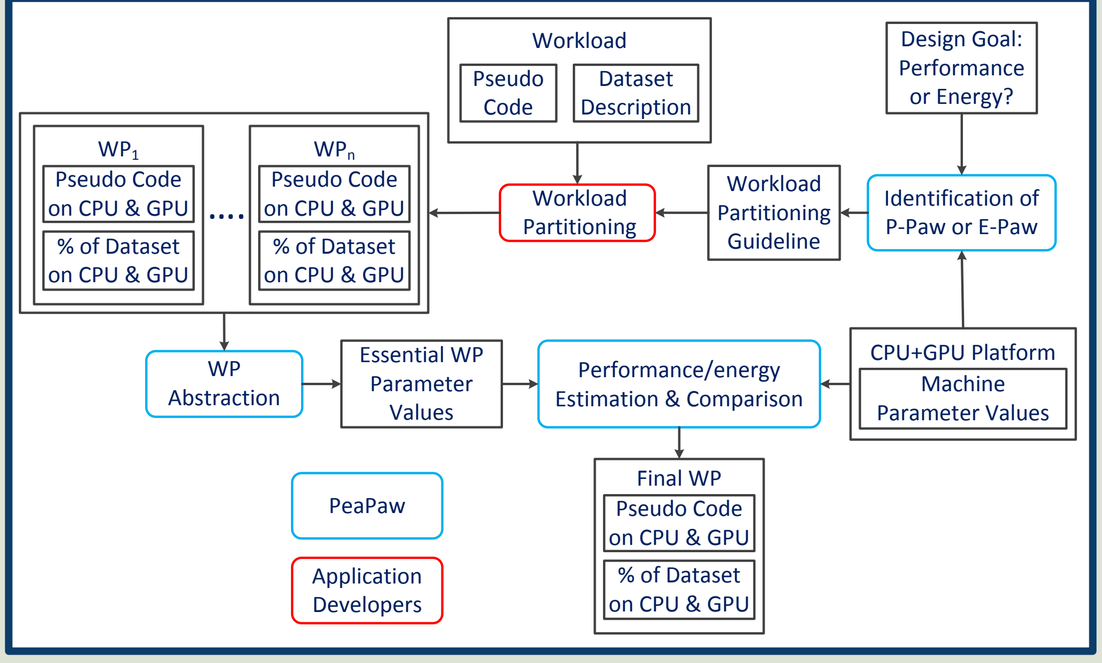
Li Tang
Advisor: X. Sharon Hu
University of Notre Dame

UNIVERSITY OF NOTRE DAME
College of Engineering

Introduction

- PeaPaw's goal
  - Help developers partition workload between CPU and GPU for high performance or energy efficiency with low cost
- Contributions
  - High-level performance/energy models for heterogeneous platforms and essential WP parameters
  - Performance/energy oriented workload partitioning guidelines
- Motivation example
  - Partition workload by using code partitioning (CP) or data partitioning (DP)?

PeaPaw Framework



PeaPaw Model

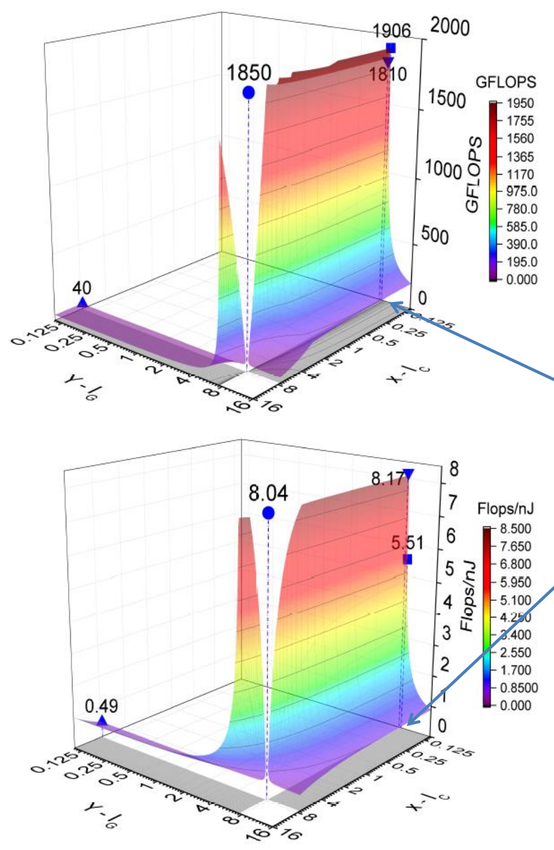
- Operation intensity: $I = \frac{N_{f-c} + N_{f-g}}{N_{b-c} + N_{b-g}}$, $I_c = \frac{N_{f-c}}{N_{b-c}}$, $I_g = \frac{N_{f-g}}{N_{b-g}}$
- PerPaw:
  - $GFlops = \frac{N_{f-c} + N_{f-g}}{\max\{t_{f-c}N_{f-c}, t_{b-c}N_{b-c}, t_{f-g}N_{f-g}, t_{b-g}N_{b-g}\}}$
    $= \max\{t_{f-c}\frac{I_c(I-I_g)}{I(I_c-I_g)}, t_{b-c}\frac{I-I_g}{I(I_c-I_g)}, t_{f-g}\frac{I_g(I-I_c)}{I(I_g-I_c)}, t_{b-g}\frac{I-I_c}{I(I_g-I_c)}\}$
  - $Flops/J = (e_{f-c}\frac{I-I_g}{I_c-I_g} + e_{b-c}\frac{I-I_g}{I(I_c-I_g)} + e_{f-g}\frac{I-I_c}{I_g-I_c} + e_{b-g}\frac{I-I_c}{I(I_g-I_c)} + (p_{s-c} + p_{s-g})*\max\{t_{f-c}\frac{I_c(I-I_g)}{I(I_c-I_g)}, t_{b-c}\frac{I-I_g}{I(I_c-I_g)}, t_{f-g}\frac{I_g(I-I_c)}{I(I_g-I_c)}, t_{b-g}\frac{I-I_c}{I(I_g-I_c)}\})^{-1}$

Parameter	Description
N_{f-c}	Number of flops in the CPU workload
N_{b-c}	Number of bytes in the CPU workload
N_{f-g}	Number of flops in the GPU workload
N_{b-g}	Number of bytes in the GPU workload
t_{f-c}	Time per flop of CPU
t_{b-c}	Time per byte of CPU
t_{f-g}	Time per flop of GPU
t_{b-g}	Time per byte of GPU
e_{f-c}	Energy per flop of CPU
e_{b-c}	Energy per byte of CPU
e_{f-g}	Energy per flop of GPU
e_{b-g}	Energy per byte of GPU
p_{s-c}	Static power of CPU
p_{s-g}	Static power of GPU
I	Operational intensity of the given workload
I_c	Operational intensity of the CPU workload
I_g	Operational intensity of the GPU workload

Model Validation

Processor	# of cores	GHz	t_f (pS)	t_b (pS)	e_f (pJ)	e_b (pJ)	p_s (W)
Intel i7 2600K	4	3.4	9.5	65.9	118	462	26.8
Intel i3 2100T	2	2.5	25	73	135	581	9.7
NVIDIA GTX Titan	2688	0.837	0.4	4.2	57	187	64.1
NVIDIA GTX 750	512	1.02	1.9	14.8	78	169	16.4

Platforms	P-Paw category (guideline)	E-Paw category (guideline)
i7+Titan	CPU_MEM-GPU_COMP	CPU_MEM-GPU_COMP
i7+750	CPU_MEM-GPU_COMP	Race-to-halt
i3+Titan	CPU_MEM-GPU_COMP	GPU-only
i3+750	CPU_MEM-GPU_COMP	GPU-only

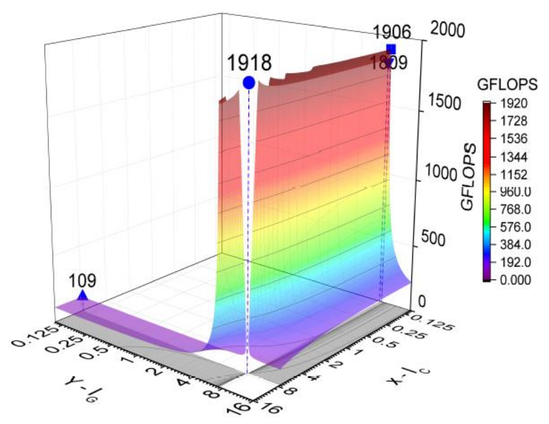


(a) Estimated performance and energy of SA's WPs on i7+750

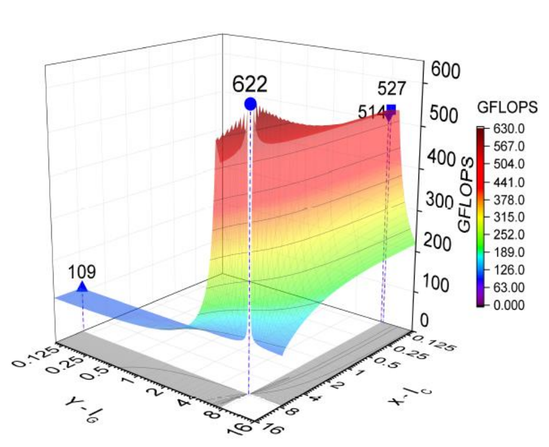
Case study: Synthetic application (SA)

WP	I	I_c	I_g
▲ CPU-only	7.6	7.6	0
▼ GPU-only	7.6	0	7.6
● p-DP	7.6	7.6	7.6
■ CP	7.6	0.1	8.0

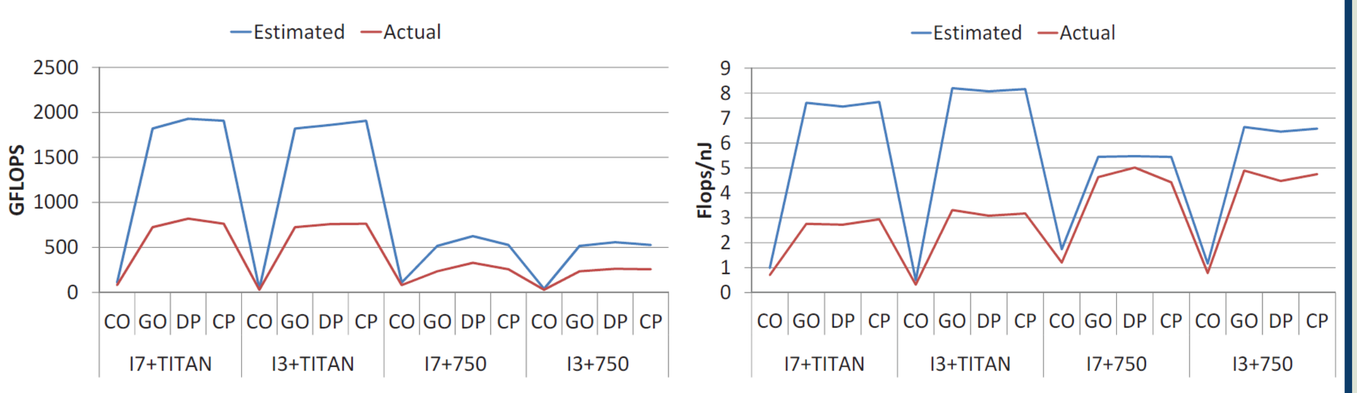
- Performance-oriented guideline on i7+750: CPU_MEM-GPU_COMP
  - CP: $I_c < I_g$
- Energy-oriented guideline on i7+750: Race-to-halt
  - CP: CP's performance is higher than the others



(b) Estimated performance and energy of SA's WPs on i7+750



(c) Estimated performance and energy of SA's WPs on i7+750



(d) Estimated and actual performance and energy of SA's WPs