Evaluating the energy impact of device and workload parameters for DNN inference on edge

Anurag Dutt, Sri Pramodh Rachuri, Ashley Lobo, Nazeer Shaik, Anshul Gandhi, Zhenhua Liu

Stony Brook University

Motivation

- Deployment of large DNN models
- Edge Computing
  - Examples - Jetson lineup
  - Scarcity of resources and energy
- Large parameter space to optimize

Introduction

- Sustainable DNN workload deployments on the Edge
- Study the impact of **hardware** parameters
  - CPU frequency
  - GPU frequency

Device Specifications and Workloads

Specification	Jetson Nano	Jetson Xavier NX
CPU	4-core ARM A57	8-core Nvidia Carmel
CPU Freq. range	102 MHz – 1.48 GHz	115 MHz – 1.9 GHz
CPU Freq. step	100 MHz (15 steps)	77 MHz (25 steps)
GPU	Nvidia Maxwell	NVIDIA Volta
CUDA Cores	128	384
Tensor Cores		48
Memory	4 GB LPDDR4	8 GB LPDDR4
GPU Freq. range	76 MHz – 921 MHz	114 MHz – 1.1 GHz
GPU Freq. steps	77 MHz (count 12)	90 MHz (count 15)
Throughput	472 GFLOPs	21 TOPs
Power Modes	5W, 10W	10W, 15W
Libraries	CUDA 10.2 + cuDNN 8.2.1	CUDA 10.2 + cuDNN 8.0.0

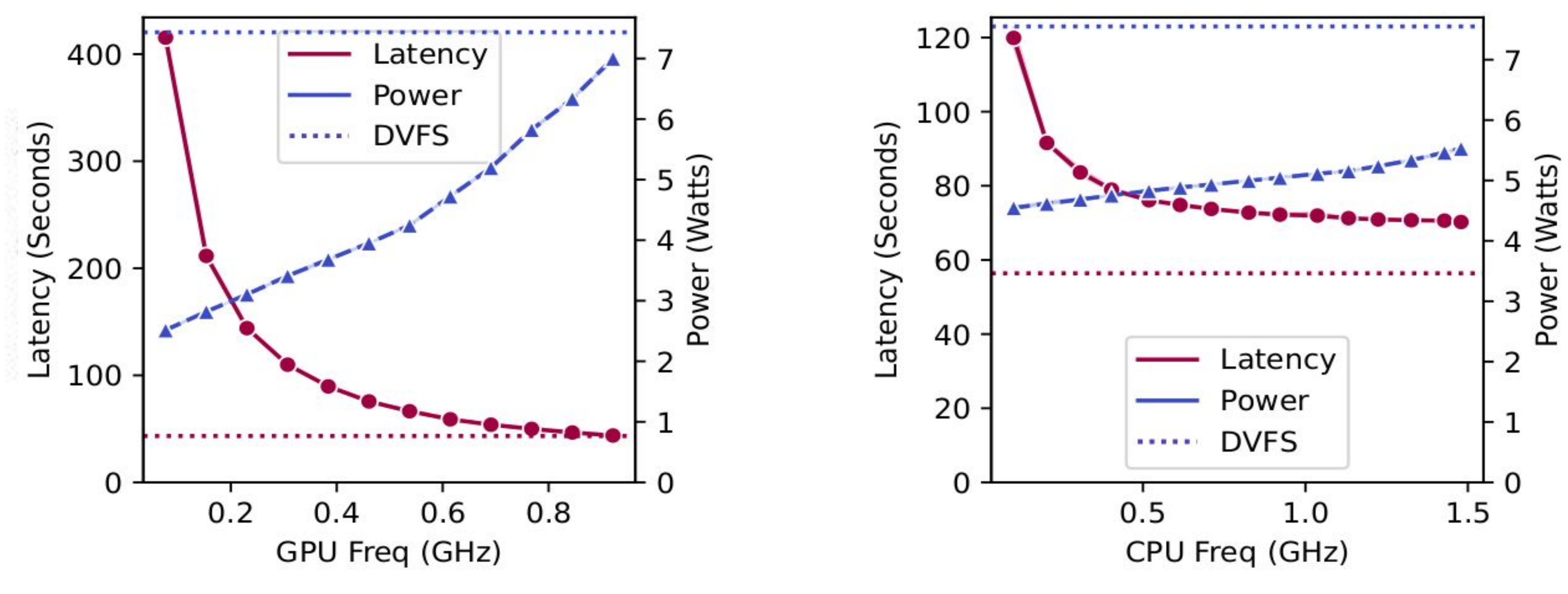
Model	Layers	Params	Ops (GFLOPs)	Batch Size	Input
AlexNet	8	61M	0.727	4, 8, 16, 32, 64	Tensor (3,224,224)
ResNet-18	18	11M	2	4, 8, 16, 32, 64	Tensor (3,224,224)
MobileNet-V2	53	3.4M	0.57	4, 8, 12	Tensor (3,224,224)
YOLOv4-Tiny	29	6.1M	6.9	4, 8, 16, 32, 64	Tensor (3,416,416)
BERT-Tiny	4	4.4M	0.0353	4, 8, 16, 32, 64	String (512 words, 1.1kb)
DistilBERT	6	43.2M	4.3	4, 8, 16	String (512 words, 1.1kb)

Experimental Setup

- Power readings for each device are polled at 100ms intervals
  - Overhead for 100 ms < 0.5%; Overhead for more frequent polling (10 ms or 1 ms) > 2%
- PyTorch for all the workloads except for YOLOv4 (OpenCV)
- Implemented a separate thread to poll the I2C interface for continuous power readings
- Each experiment on a given model
  - One out of x CPU+GPU Freq combinations
  - Fixed workload - 3200 inferences inputs
  - 10 reruns; variance was less than 5%

Evaluation

Frequency Sweeps - Jetson Nano

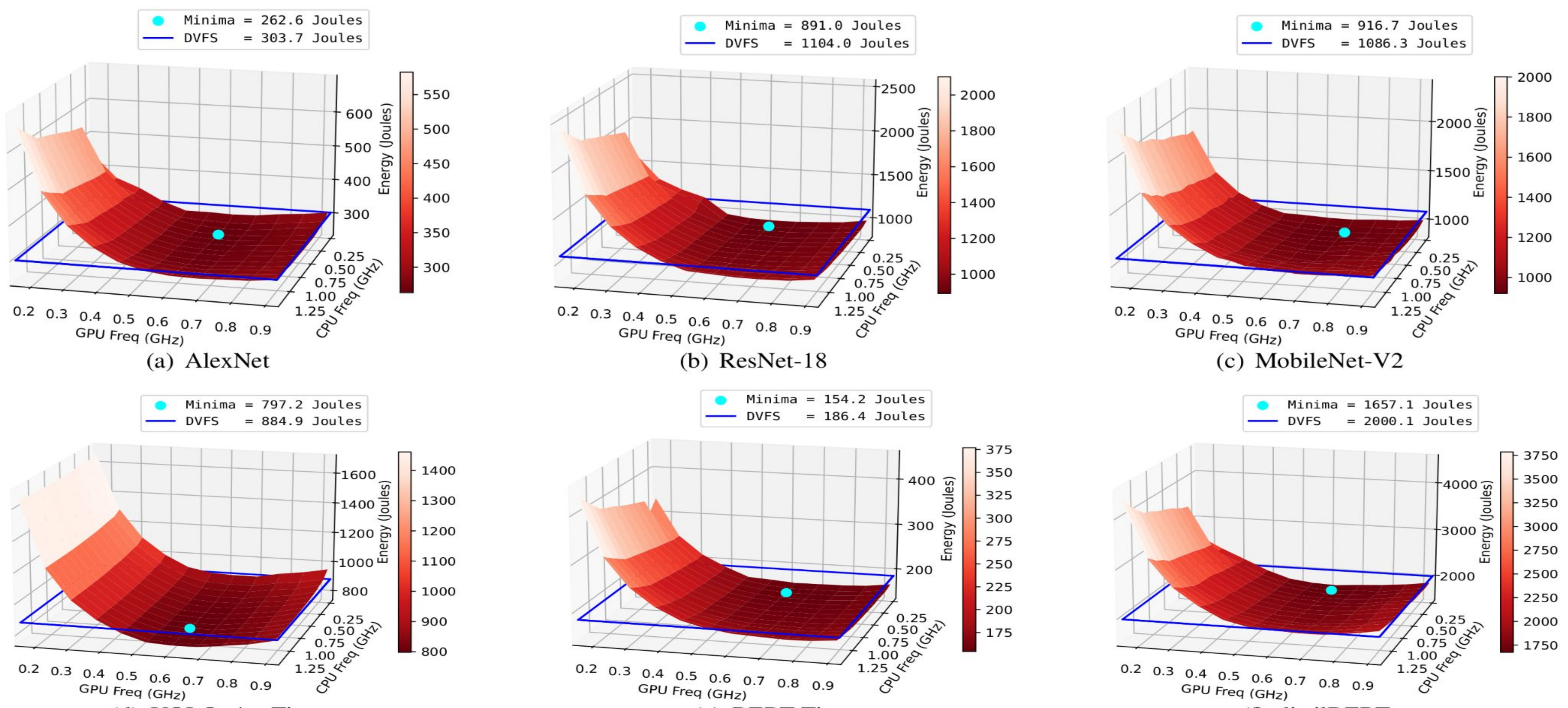


(a) Changing GPU frequency (b) Changing CPU frequency

- DVFS Governor
  - CPU Default - "schedutil"
  - GPU Default - "nvhost_podgov"
  - Highest freq - CPU 89%; GPU 83%
  - Other governors < 1% variation

Monotonic relation with freq | **Impact of CPU Freq < GPU Freq**

Energy usage trends on Jetson Nano

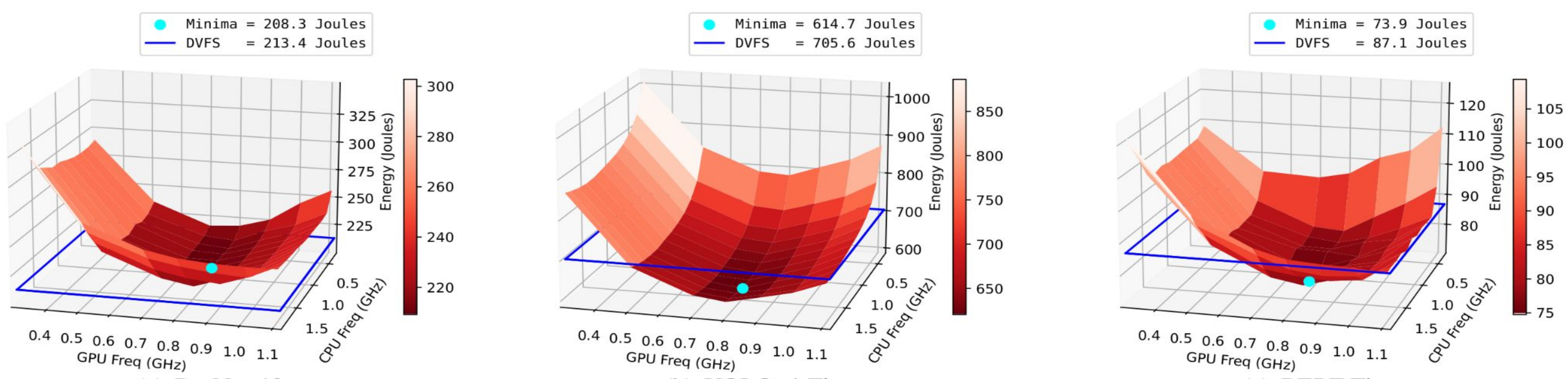


(a) AlexNet (b) ResNet-18 (c) MobileNet-V2 (d) YOLOv4 - Tiny (e) BERT-Tiny (f) distilBERT

Minima consumes 13%, 19%, 15%, 9%, 17% lower energy than DVFS

GPU Freq substantially impacts Energy but non-monotonic

Energy Usage Trends on Xavier NX



(a) ResNet-18 (b) YOLOv4-Tiny (c) BERT-Tiny

Minima consumes 2%, 13%, 15% lower energy than DVFS

Non-monotonic behaviour of CPU Freq is more prominent

Conclusion

- Selecting optimal freqs gives upto 19% saving in energy for Jetson Nano
- Selecting optimal freqs gives upto 15% savings in energy for Xavier NX
- Energy Consumption of Xavier NX is significantly lower between 2x and 4x as compared to Nano

Future Work

- Study the impact of **workload** parameters
  - Batch Size
  - Number of layers
- Develop a joint workload parameter optimization strategy for optimal energy configuration

PACE
Stony Brook University

References

- You, J., Chung, J.-W., & Chowdhury, M. (2023). *Zeus: Understanding and Optimizing {GPU} Energy Consumption of {DNN} Training*
- Trainer: An Energy-Efficient Edge-Device Training Processor Supporting Dynamic Weight Pruning. (n.d.). Ieeexplore.ieee.org
- S.K, P., Kesanapalli, S. A., & Simmhan, Y. (2022). Characterizing the Performance of Accelerated Jetson Edge Devices for Training Deep Learning Models.
- S. Holly, A. Wendt and M. Lechner, "Profiling Energy Consumption of Deep Neural Networks on NVIDIA Jetson Nano,"