Dynamic Inference for Efficient Inference on Mobile and Embedded Systems

Geoff Merrett

Huawei Sweden Future of Wireless Workshop on AI (Session on Efficient AI)

11 March 2025 | Stockholm, Sweden

spatialml.net

UNIVERSITY OF Southampton

UNIVERSITY OF SOUTHAMPTON

University of Southampton

- ~30,000 students
- Top 100 universities worldwide (#80 QS‘25)
- Founding member of UK’s Russell Group

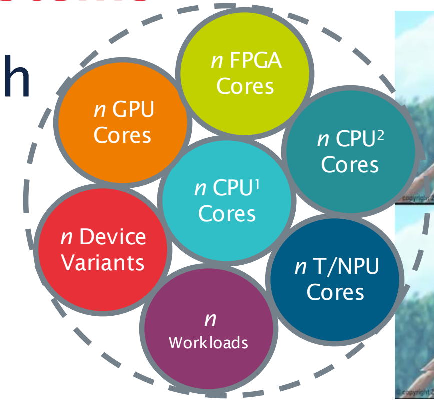
School of Electronics and Computer Science

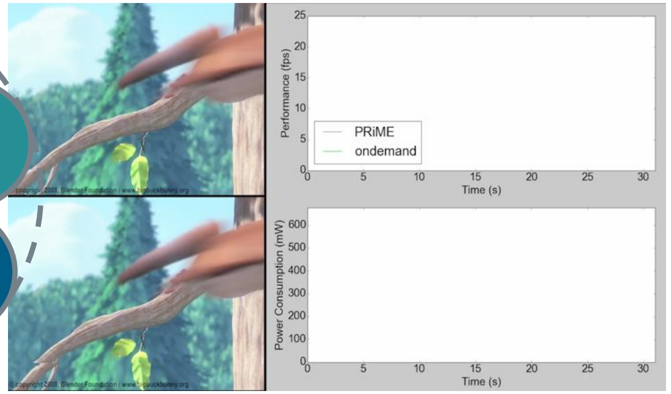
- ~2,500 students
- ~300 PhD research students
- ~150 academics/faculty
- Top 3 in UK for Electronic Engineering
- 16 research groups/centres

RESEARCH INTERESTS

Resource management in mobile/embedded systems

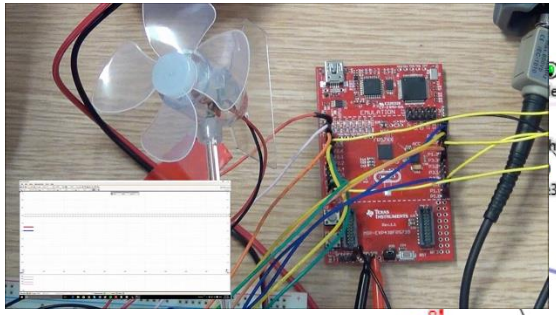
- Typically heterogeneous multi-core systems with numerous operating points/configurations
- Matching to application/user QoE and/or QoS metric

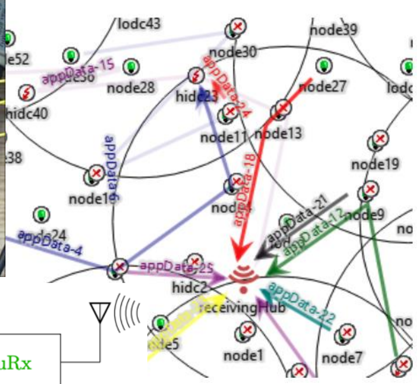


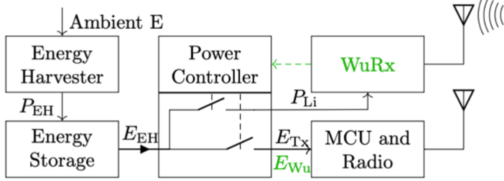


Self-powered embedded sensing systems

- Typically ultra-constrained MCU systems, with variable power harvesting and and limited storage
- Matching to system and application requirements



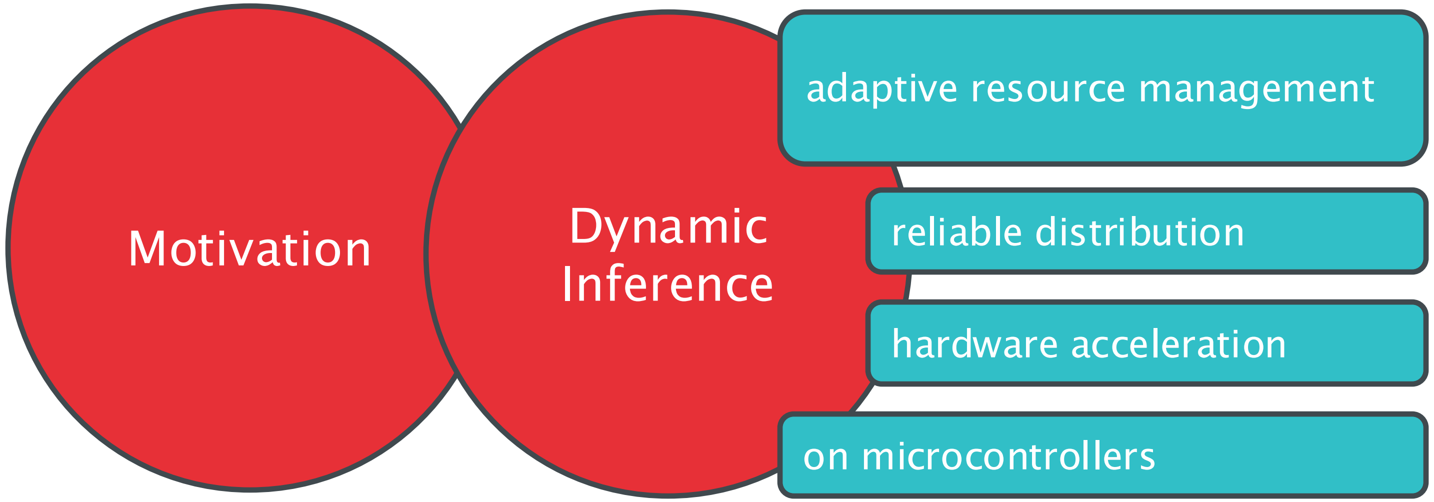




Increasingly, efficient AI as a workload in these domains

DYNAMIC INFERENCE FOR EFFICIENT EDGE INFERENCE

"Broad brush strokes…"



Acknowledgements

- **Lei Xun**, Anastasios Dimitriou, Sulaiman Sadiq, Hengrui Zhao, Mingyu Hu, …
- EPSRC Funded Centre for Spatial Computational Learning https://spatialml.net

AI AT THE EDGE

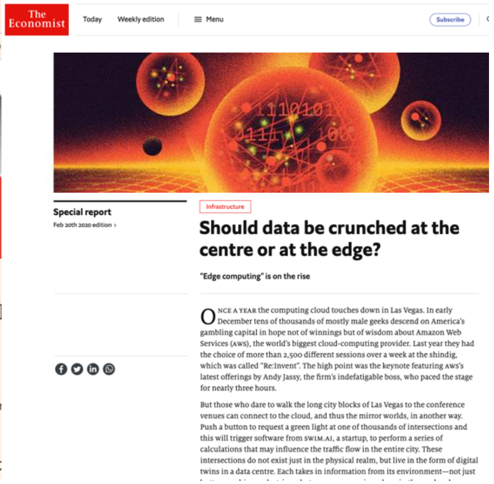
Inference at the Edge

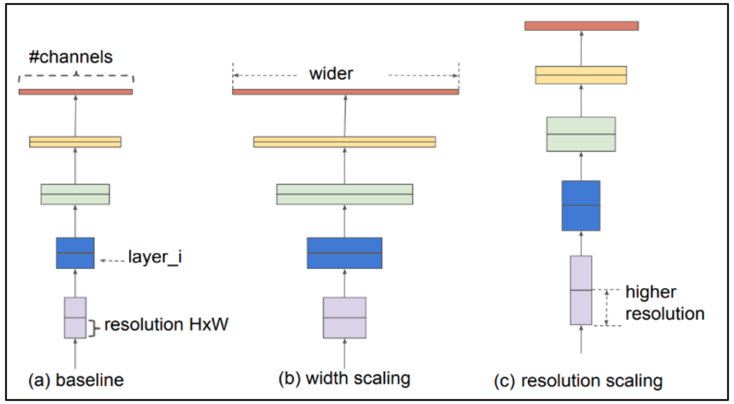
- Increased privacy
- Reliance on network connectivity/latency/bandwidth
- Reduced power/energy

The Edge* is Resource Constrained

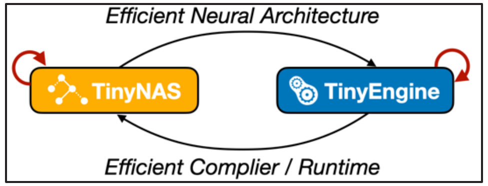
- DNN models are computationally and memory-access intensive.
- Model compression (e.g. pruning, quantization, knowledge distillation), architecture search, distributed networks, frameworks, kernels, etc).

** Referring to the mobile/embedded edge in this presentation*





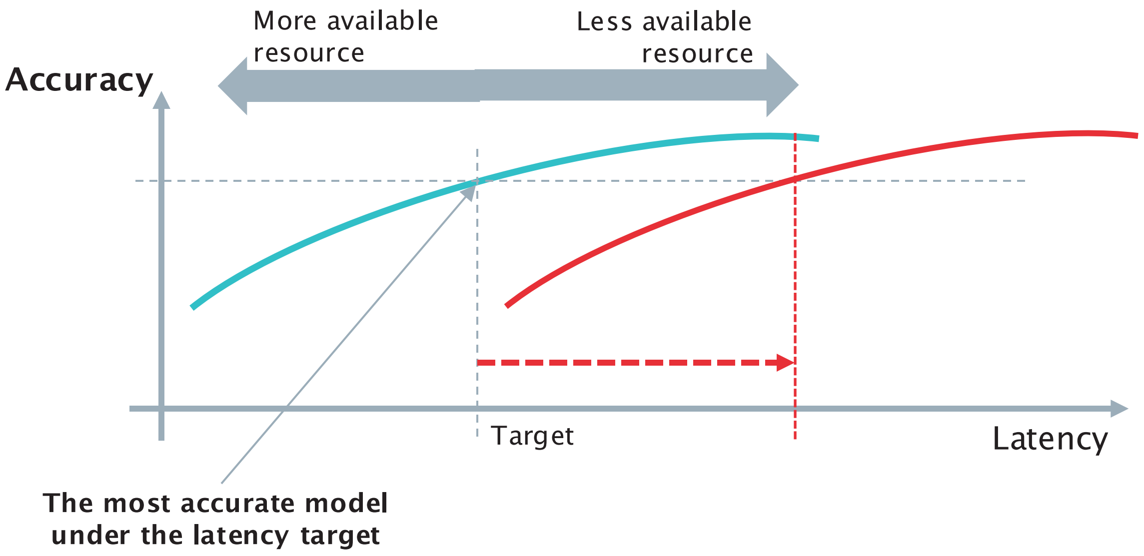
Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks." *International conference on machine learning*. PMLR, 2019.

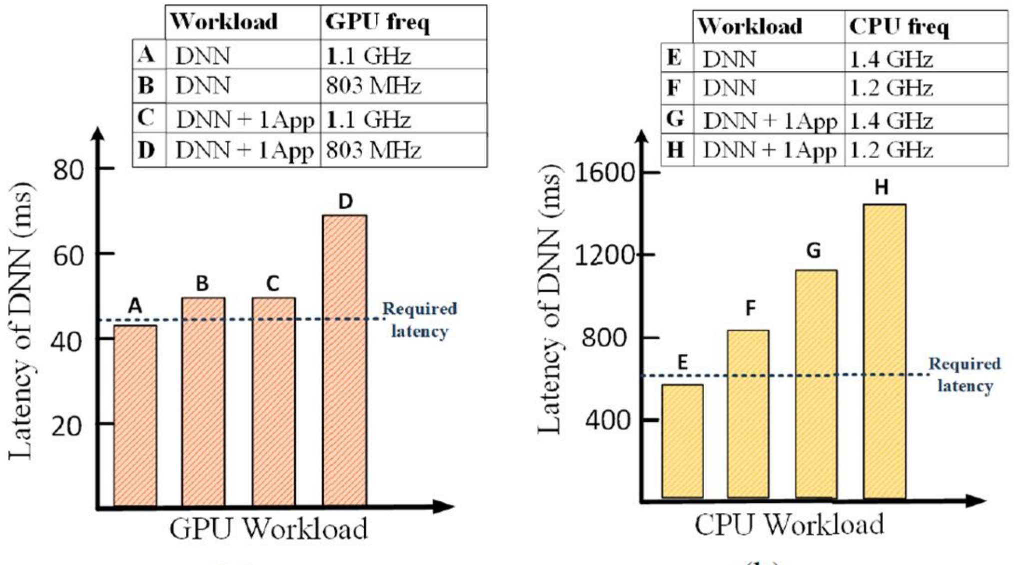


Lin, Ji, et al. "Mcunet: Tiny Deep Learning on IoT Devices." *Advances in Neural Information Processing Systems* 33, 2020

DYNAMIC RESOURCE AVAILABILITY

- Model compression trades-off accuracy and latency (hardware-dependent)
- Modern heterogeneous platforms are dynamic:
  - Dynamic Hardware and Runtime Conditions (moving trade-off curve)
  - Dynamic Application Requirements (moving performance targets)





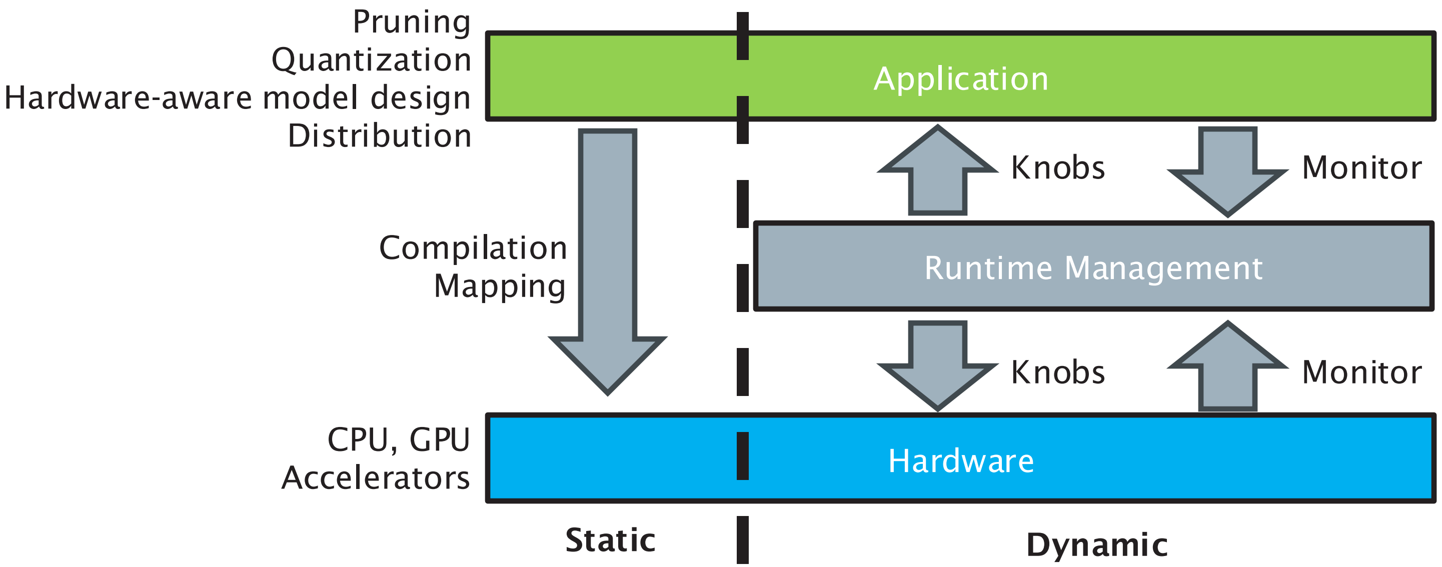
(a) (b)

L. Xun, L. Tran-Thanh, B. M. Al-Hashimi, and G. V. Merrett, "Optimising Resource Management for Embedded Machine Learning" in Design, Automation & Test in Europe Conference (DATE), 2020.
W. Lou, L. Xun, A. Sabet, J. Bi, J. Hare, and G. V. Merrett, "Dynamic-OFA: Runtime DNN Architecture Switching for Performance Scaling on Heterogeneous Embedded Platforms" in CVPRW, 2021.

DYNAMIC/ADAPTIVE INFERENCE

- We need models that can adapt to platform/resource and workload diversity, to:
  - adapt to available system resources
  - adapt to application requirements
  - improve model reuse on similar platforms

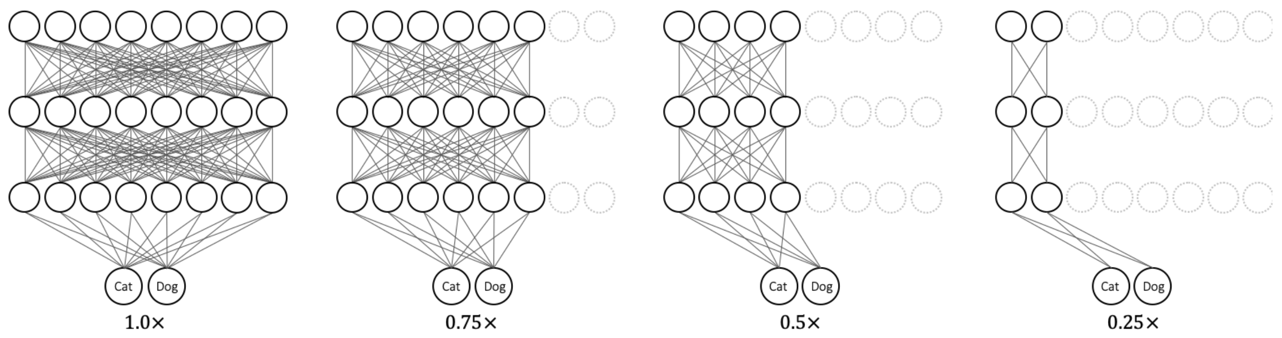
meeting latency / power/energy / accuracy requirements



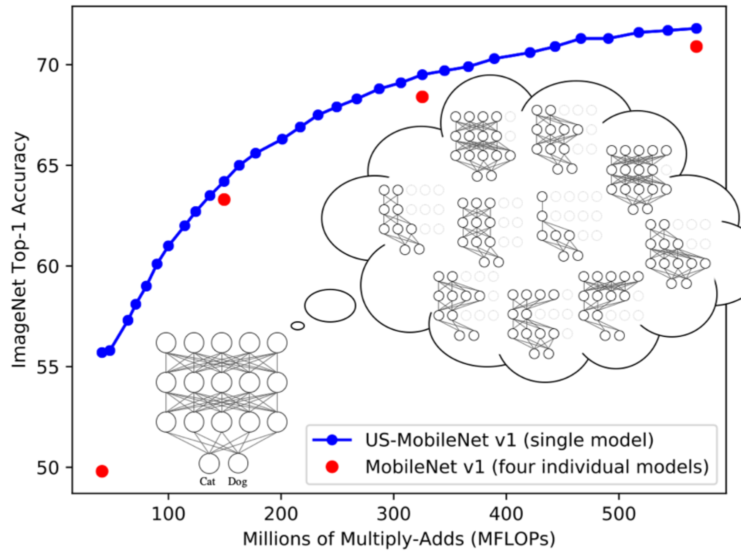
L. Xun, L. Tran-Thanh, B. M. Al-Hashimi, and G. V. Merrett, "Optimising Resource Management for Embedded Machine Learning" in Design, Automation & Test in Europe Conference (DATE), 2020.

DYNAMIC DNNS

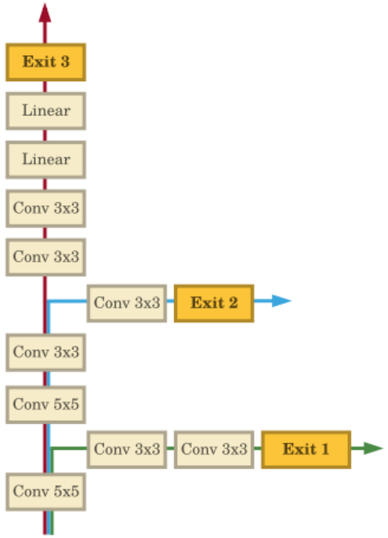
- Width scaling
- Dynamic bit-width/quantisation
- Channel scaling
- Resolution scaling



J. Yu, L. Yang, N. Xu, J. Yang, and T. Huang, "Slimmable Neural Networks" in ICLR, 2019.



J. Yu and T. Huang, "Universally slimmable networks and improved training techniques" ICCV, 2019.

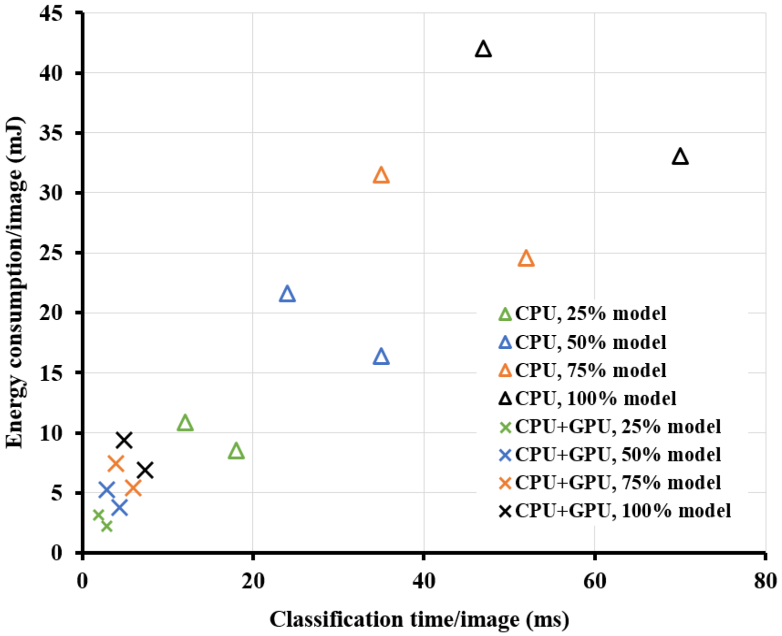


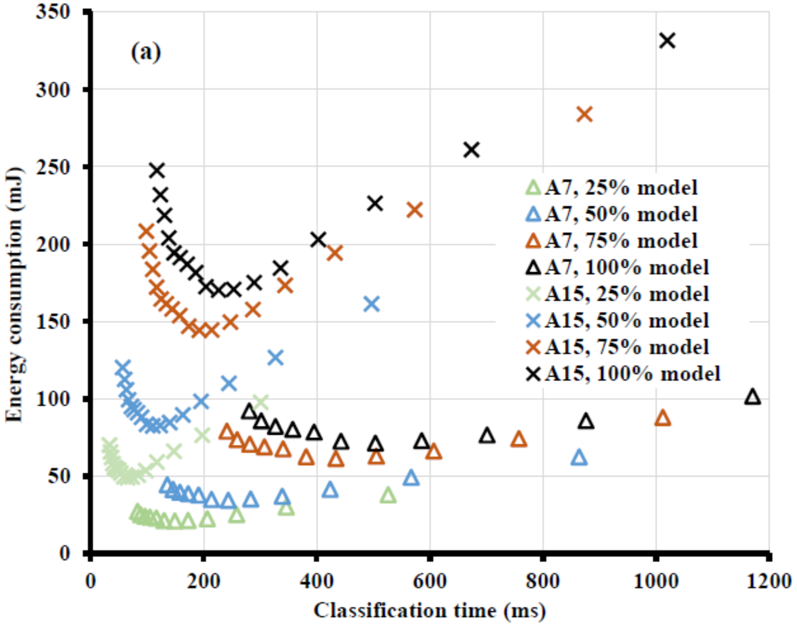
S. Teerapittayanon, B. McDanel, and H.T. Kung, "BranchyNet: Fast inference via early exiting from deep neural networks" in ICPR, 2016.

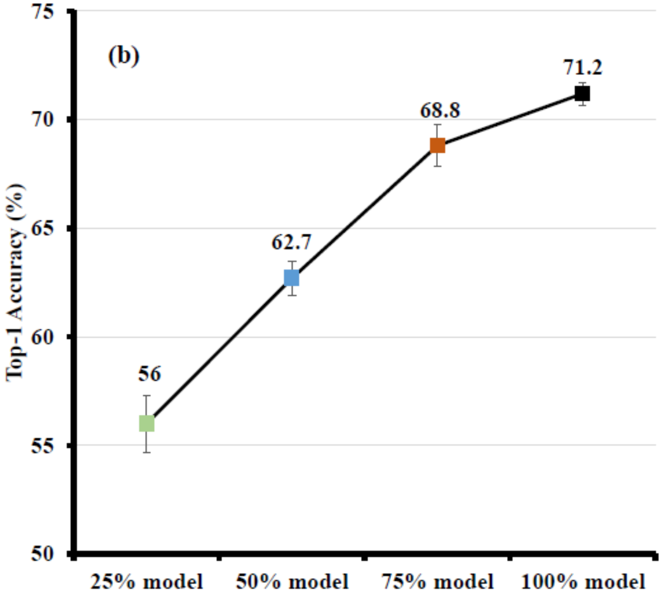
Max Sponner, Bernd Waschneck, and Akash Kumar. 2024. Adapting Neural Networks at Runtime: Current Trends in At-Runtime Optimizations for Deep Learning. ACM Comput. Surv. 56, 10, Article 248 (May 2024)

OUR EARLY WORK

- Incremental training with group convolution pruning
- Adapted AlexNet (~320kB) on CIFAR10
- Odroid XU3 (4x A15 + 4x A7) + Nvidia Jetson Nano (4x Arm A57 + 128x Maxwell CUDA cores)





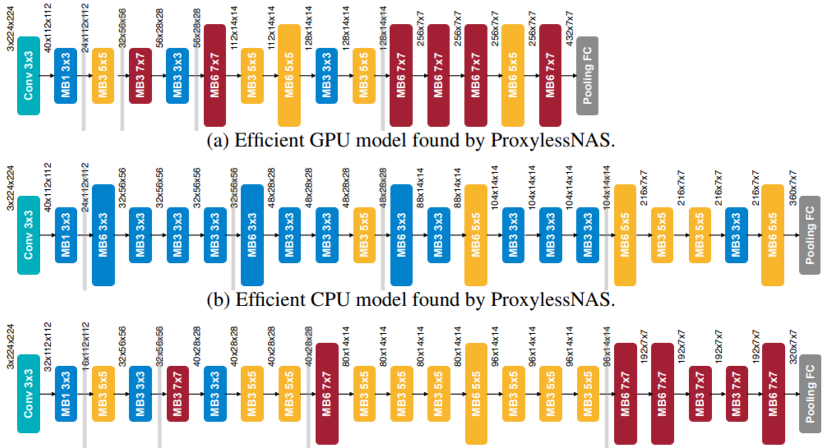


L. Xun, L. Tran-Thanh, B. M. Al-Hashimi and G. V. Merrett, "Incremental Training and Group Convolution Pruning for Runtime DNN Performance Scaling on Heterogeneous Embedded Platforms," 2019 ACM/IEEE 1st Workshop on Machine Learning for CAD (MLCAD), Canmore, AB, Canada, 2019, pp. 1-6

DYNAMIC OFA

Issues with dynamic networks

- Significant training time cost
- Conflict with the SOTA NAS model pipeline
- Inference inefficient on heterogeneous resources
  - **GPUs** prefer **shallow and wide** DNN architectures.
  - **CPUs** prefer **deep and narrow** DNN architectures.



(a) Efficient GPU model found by ProxylessNAS.

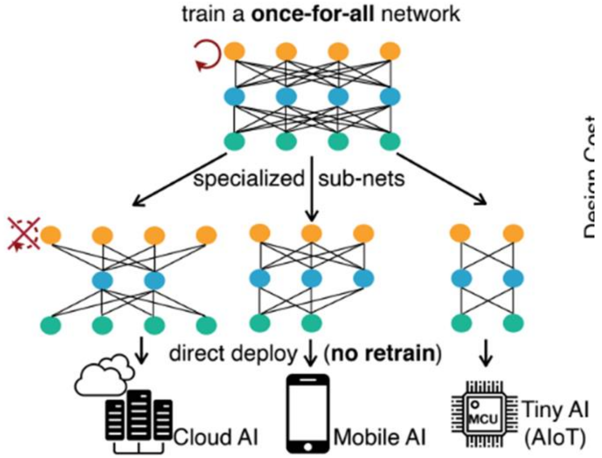
(b) Efficient CPU model found by ProxylessNAS.

(c) Efficient mobile model found by ProxylessNAS.

C. Han, L. Zhu, and S. Han, "ProxylessNAS: Direct neural architecture search on target task and hardware" in ICLR, 2019.

Once-for-all

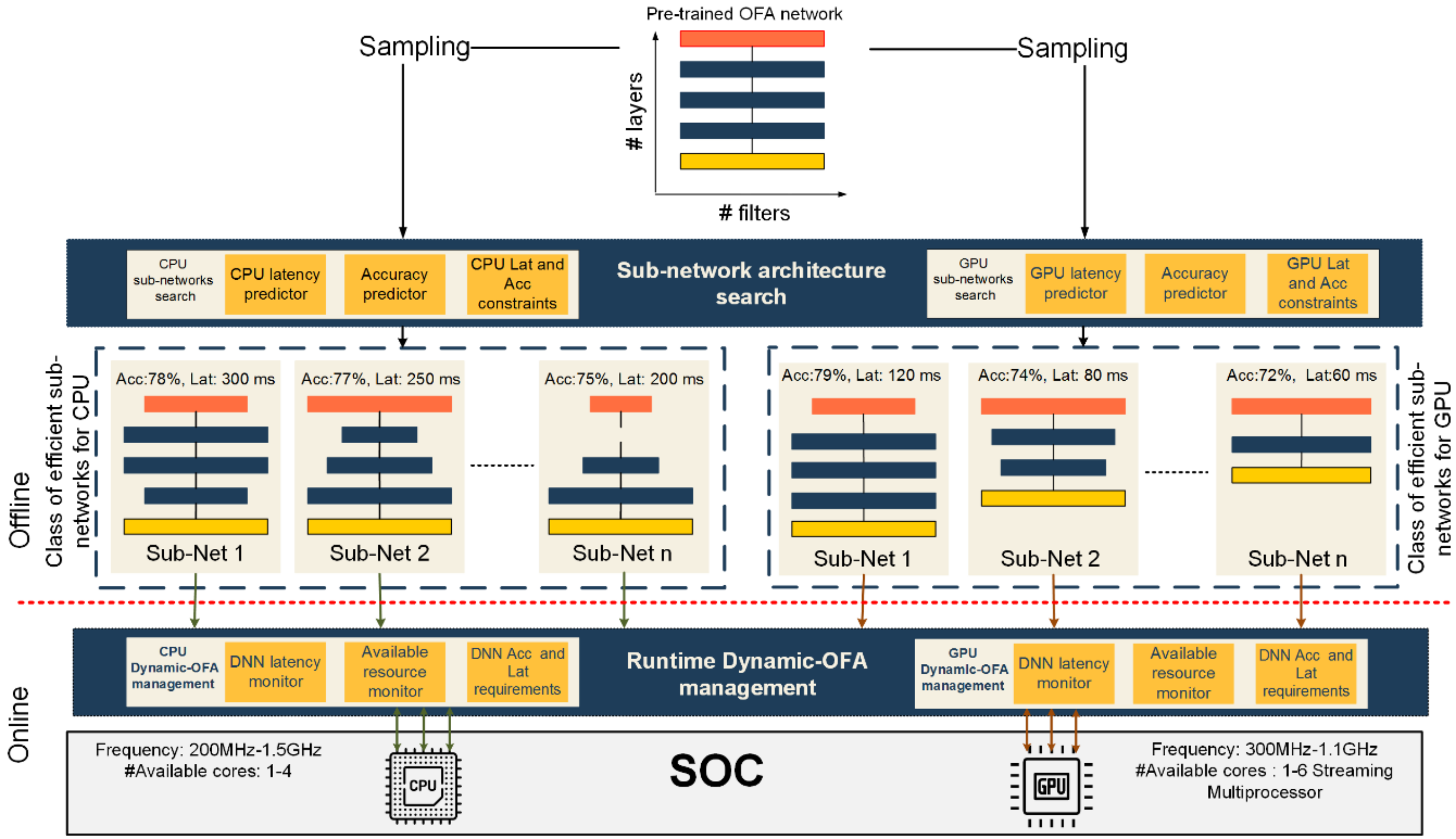
- Train model once for 10^19 sub-networks with different accuracy-latency trade-offs
- Model architecture changes at a fine level (i/p resolution, kernel size, layer, channel)
- Runtime search not feasible (and existing search designed for finding one model)

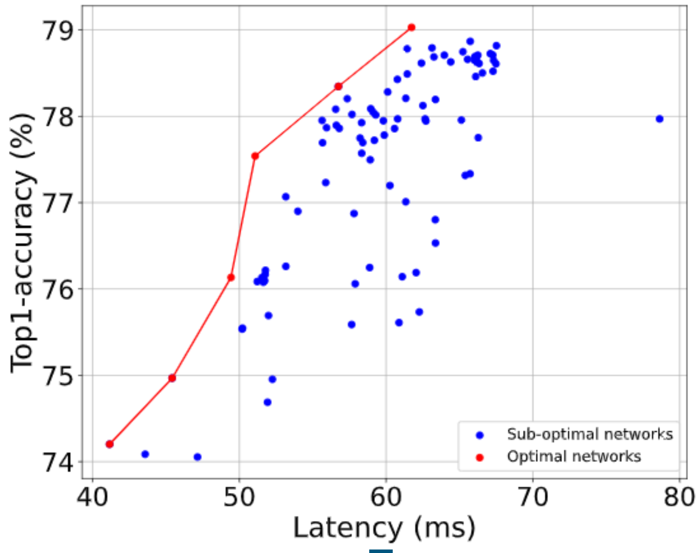


H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-all: Train one network and specialize it for efficient deployment" in ICLR, 2020.

DYNAMIC OFA

- Dynamic DNNs + Once-for-all = small number of best architectures

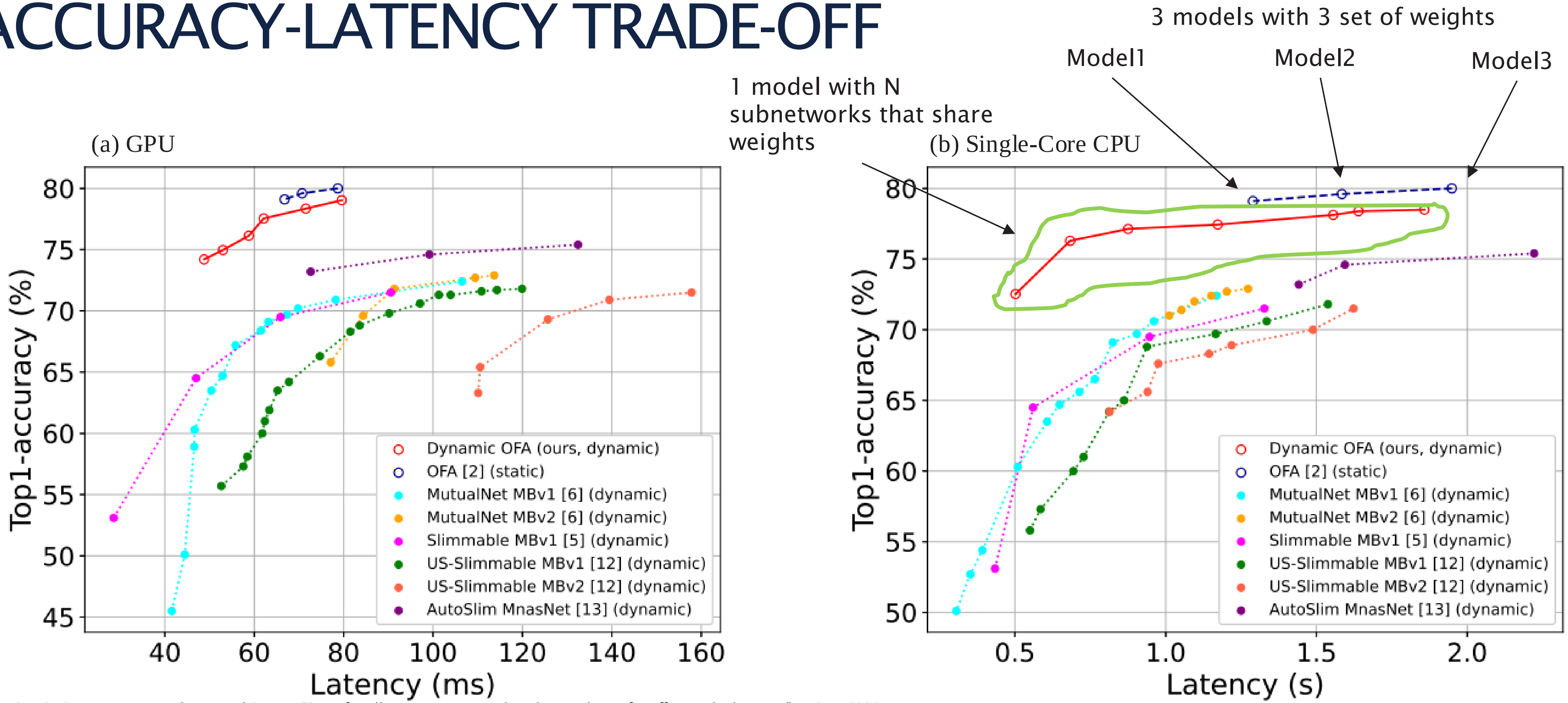




GPUs prefer shallow and wide DNN architectures, while CPUs prefer deep and narrow DNN architectures. So separated sampling is conducted.

W. Lou et al., "Dynamic-OFA: Runtime DNN Architecture Switching for Performance Scaling on Heterogeneous Embedded Platforms," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 2021, pp. 3104-3112

ACCURACY-LATENCY TRADE-OFF



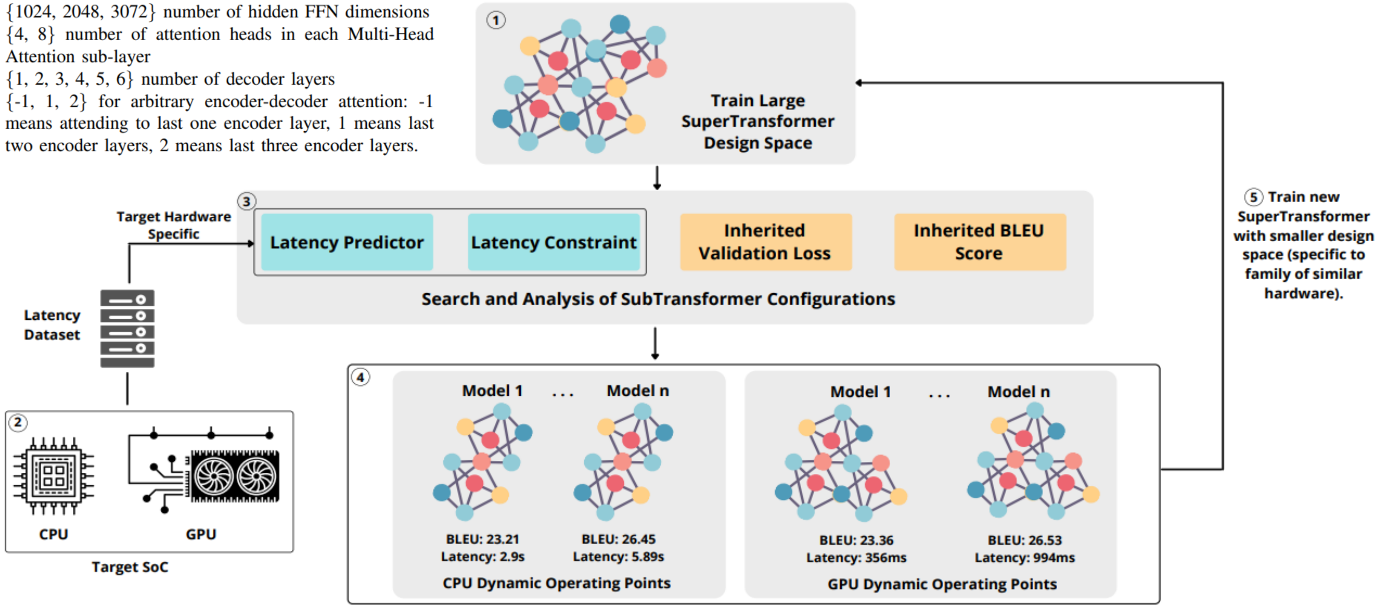
[2] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, “Once-for-all: Train one network and specialize it for efficient deployment” in ICLR, 2020.
[6] T. Yang, S. Zhu, C. Chen, S. Yan, M. Zhang, and A. Willis, “MutualNet: Adaptive convnet via mutual learning from network width and resolution” in ECCV, 2020.
[5] J. Yu, L. Yang, N. Xu, J. Yang, and T. Huang, “Slimmable neural networks” in ICLR, 2019.
[12] J. Yu and T. Huang, “Universally slimmable net-works and improved training techniques” in ICCV, 2019.
[13 J. Yu and T. Huang, “Autoslim: Towards one-shot architecture search for channel numbers” in arXiv 1903.11728, 2019.

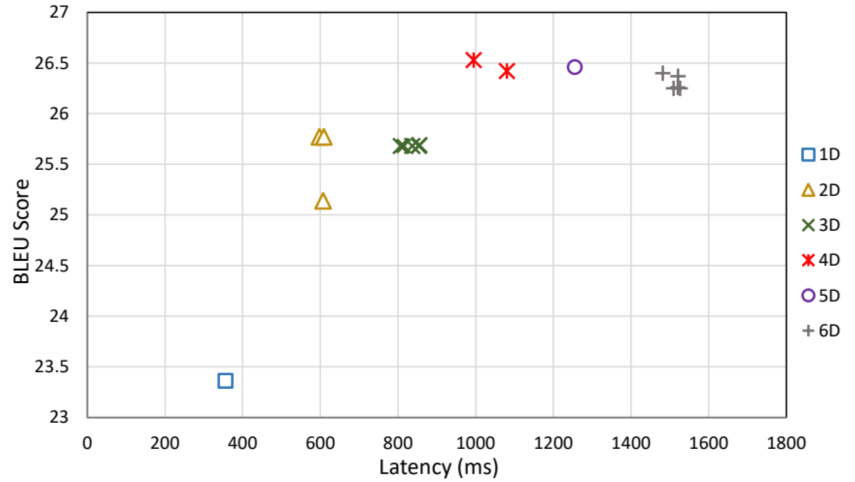
W. Lou et al., "Dynamic-OFA: Runtime DNN Architecture Switching for Performance Scaling on Heterogeneous Embedded Platforms," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 2021, pp. 3104-3112

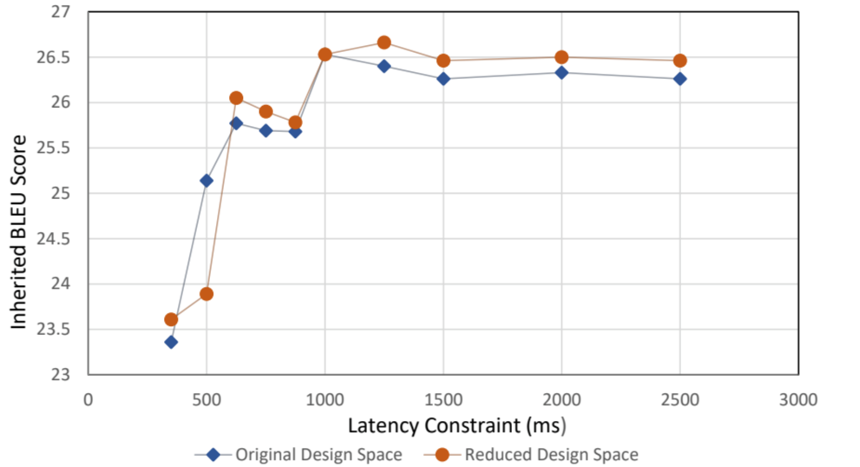
DYNAMIC TRANSFORMERS

- Also extended the idea to Dynamic-HAT, using Hardware-Aware Transformers (HAT) as a backbone.

- {512, 640} input embedding dimensions for the encoder/decoder
- {1024, 2048, 3072} number of hidden FFN dimensions
- {4, 8} number of attention heads in each Multi-Head Attention sub-layer
- {1, 2, 3, 4, 5, 6} number of decoder layers
- {-1, 1, 2} for arbitrary encoder-decoder attention: -1 means attending to last one encoder layer, 1 means last two encoder layers, 2 means last three encoder layers.





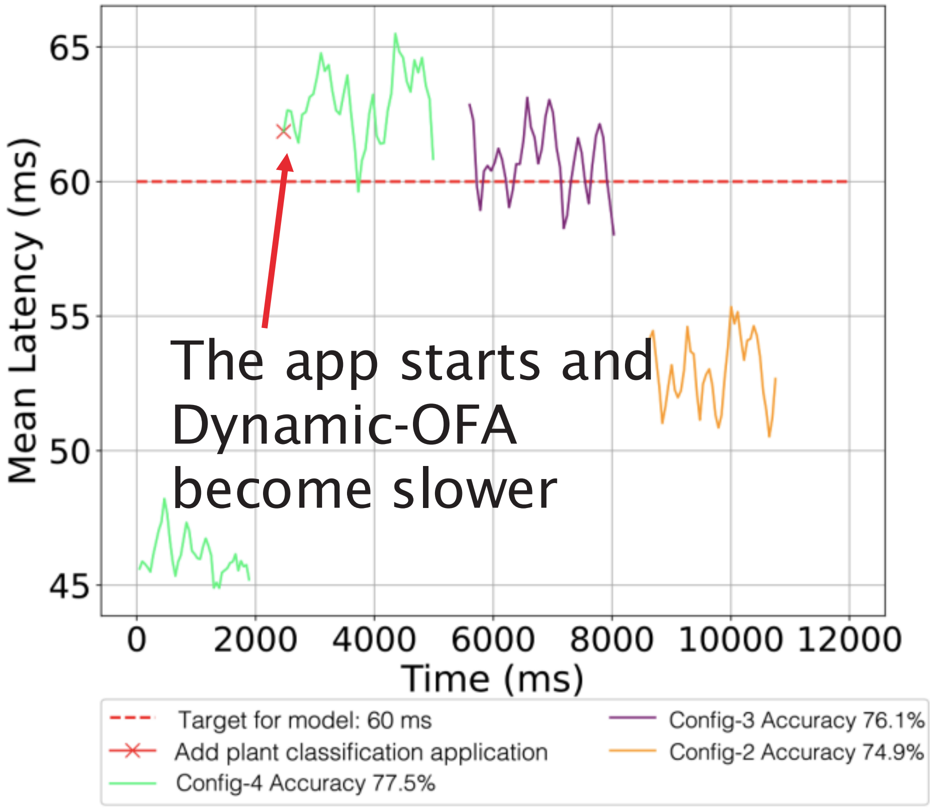


H. Wang, Z. Wu, Z. Liu, H. Cai, L. Zhu, C. Gan, S. Han, "HAT: Hardware-Aware Transformers for Efficient Natural Language Processing" in ACL, 2020.
H. Parry, L. Xun, A. Sabet, J. Bi, J. Hare and G. V. Merrett, "Dynamic Transformer for Efficient Machine Translation on Embedded Devices," 2021 ACM/IEEE 3rd Workshop on Machine Learning for CAD (MLCAD), Raleigh, NC, USA, 2021, pp. 1-6

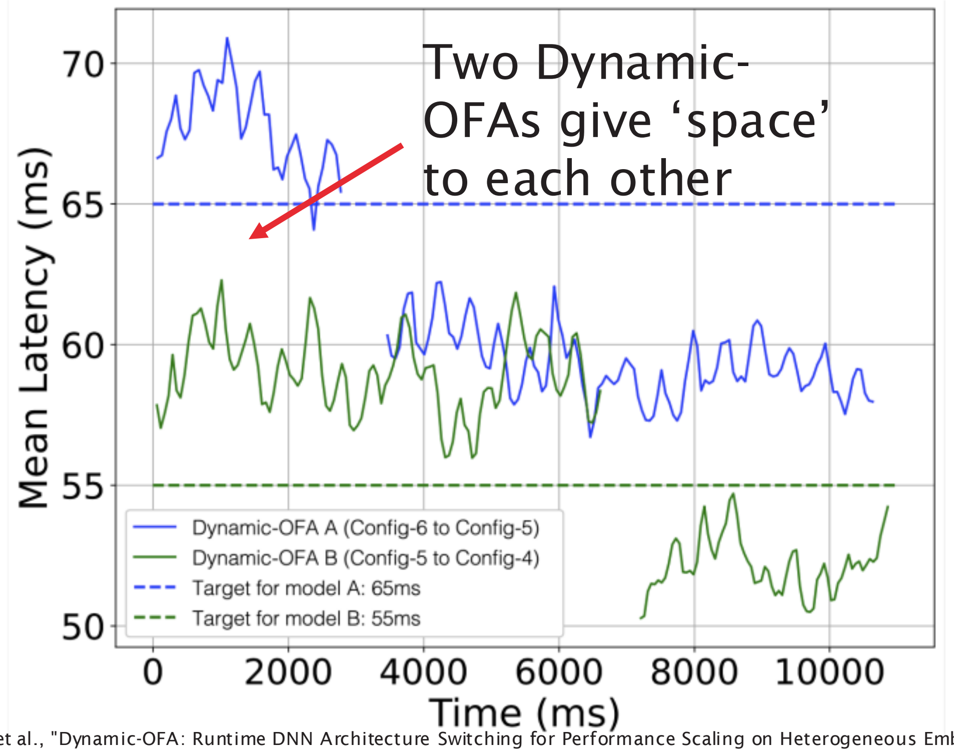
RUNTIME ADAPTATION

- Using approaches from previous work (PRiME), we could look at how to adapt and respond to changes.

Dynamic-OFA model shares GPU with App

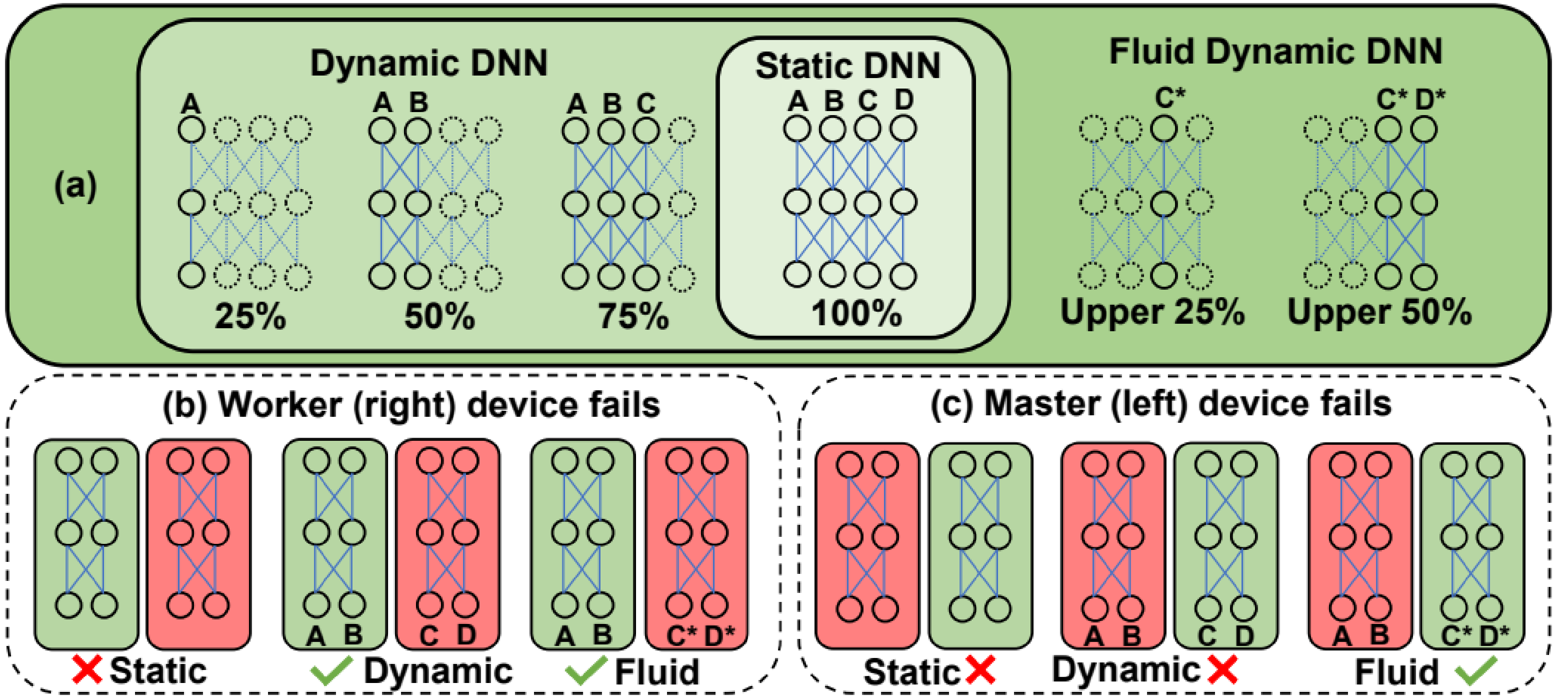


2 Dynamic-OFA models share the GPU



W. Lou et al., "Dynamic-OFA: Runtime DNN Architecture Switching for Performance Scaling on Heterogeneous Embedded Platforms," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 2021, pp. 3104-3112

IMPROVING RELIABILITY IN DISTRIBUTED DNNS

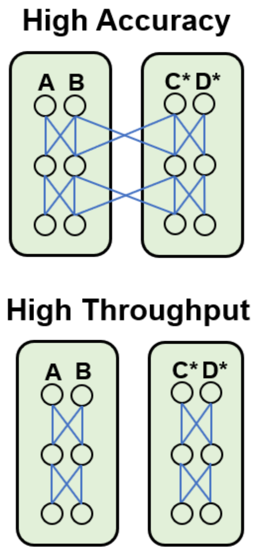


Model configurations:

Static DNN: ABCD

Dynamic DNN:
1. A
2. AB
3. ABC
4. ABCD

Fluid Dynamic DNN:
1. A
2. AB
3. ABC*
4. ABC*D*
5. C*
6. C*D*



- A Fluid Dynamic DNN model trained by incremental training, reducing dependencies between sub-networks and enhancing reliability and adaptability.

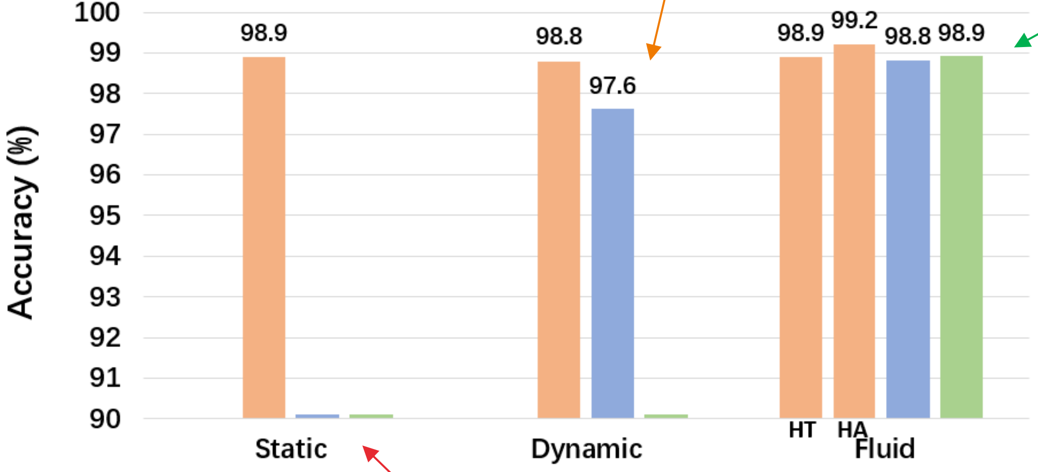
L. Xun, M. Hu, H. Zhao, A. K. Singh, J. Hare and G. V. Merrett, "Fluid Dynamic DNNs for Reliable and Adaptive Distributed Inference on Edge Devices," 2024 Design, Automation & Test in Europe Conference & Exhibition (DATE), Valencia, Spain, 2024, pp. 1-2

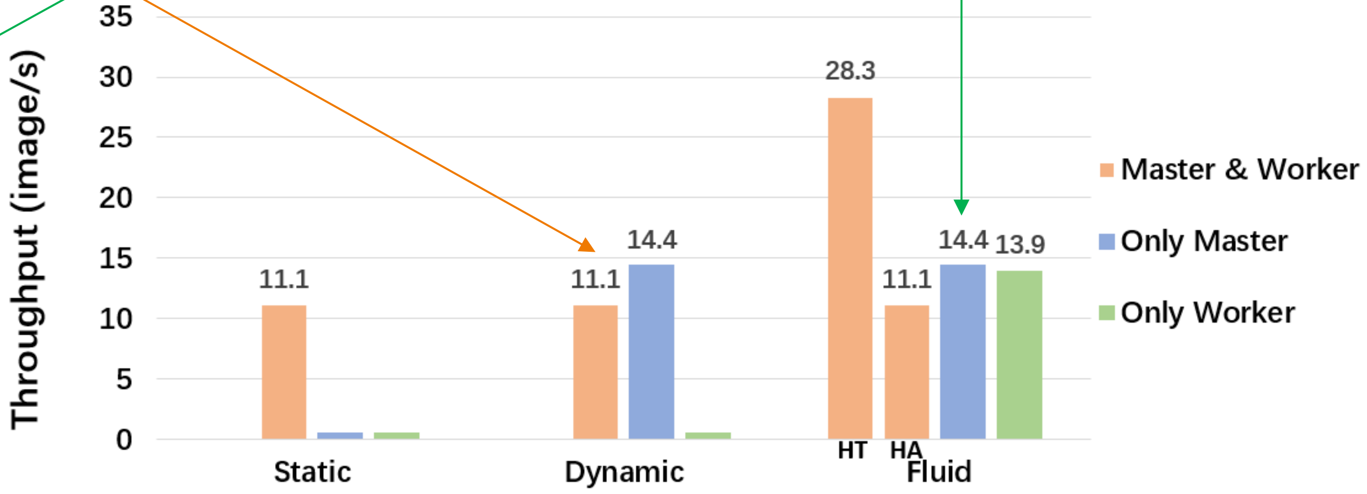
INITIAL RESULTS

- Small DNN, MNIST dataset, evaluated on the CPU of Nvidia Jetson Xavier NX platform.

Dynamic DNN can still work when worker device fails, i.e. run the 50% model on the Master device at reduced accuracy but increased throughput

- Fluid Dynamic DNN can still work when any one devices fails, i.e. run the 50% model
- High-Throughput (HT) mode and High-Accuracy (HA) mode when no device fails





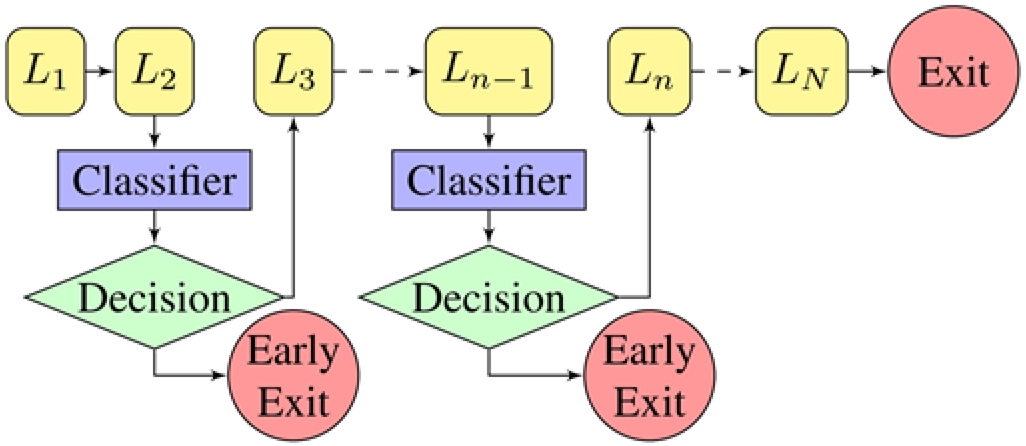
Static DNN fails when any device fails

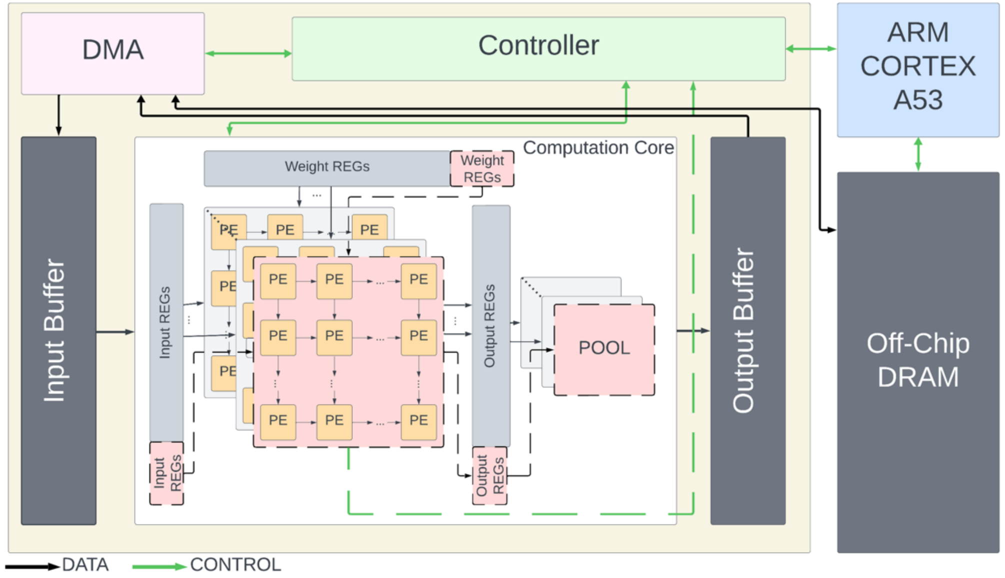
L. Xun, M. Hu, H. Zhao, A. K. Singh, J. Hare and G. V. Merrett, "Fluid Dynamic DNNs for Reliable and Adaptive Distributed Inference on Edge Devices," 2024 Design, Automation & Test in Europe Conference & Exhibition (DATE), Valencia, Spain, 2024, pp. 1-2

ACCELERATING DYNAMIC NETWORKS

Are the advantages of dynamic networks realised on accelerated hardware?

- Input-dependent early-exit networks
- The **backbone** network, which is the 'static' original network.
- The **intermediate classifiers**, which are typically placed between layers and decide the parts of the DNN to be executed.





A. Dimitriou, L. Xun, J. Hare and G. V. Merrett, "Realisation of Early-Exit Dynamic Neural Networks on Reconfigurable Hardware," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

A. Dimitriou, B. Biggs, J. Hare and G. V. Merrett, "FPGA Acceleration of Dynamic Neural Networks: Challenges and Advancements,"2024 IEEE International Conference on Omni-layer Intelligent Systems (COINS), London, United Kingdom, 2024

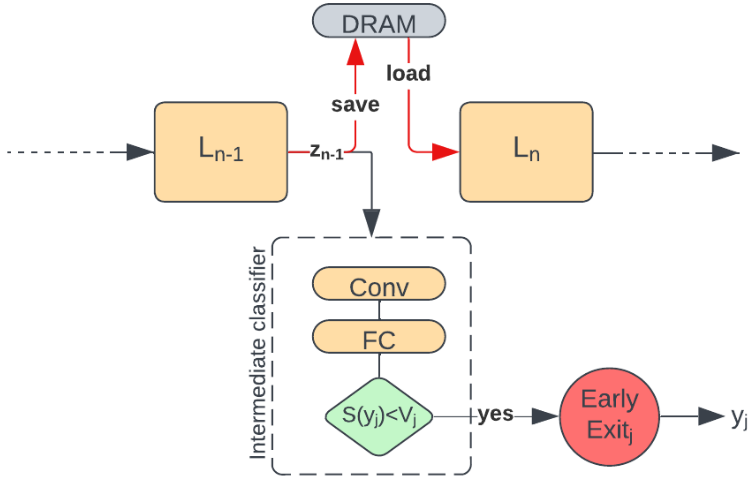
DECISION SUB-NETWORK DESIGN

Sequential Execution

✓ Reuses existing IP; lower area (and hence power) needs

✗ Increased latency when full depth is required

✗ Requires the intermediate output to be stored in memory

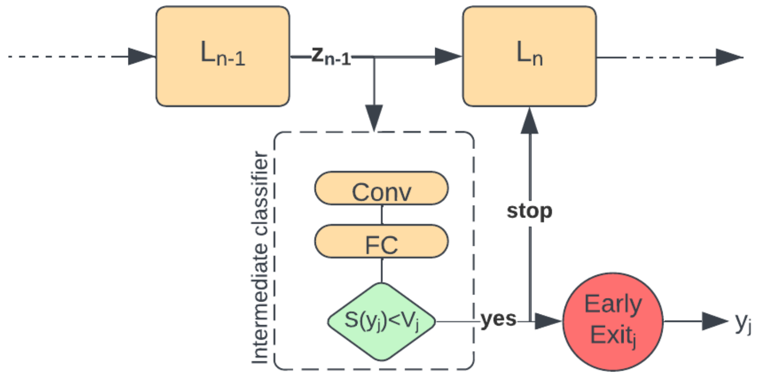


Parallel Execution

✓ No latency drop of the backbone execution

✓ Lower memory requirements

✗ Higher area (and hence power) requirements

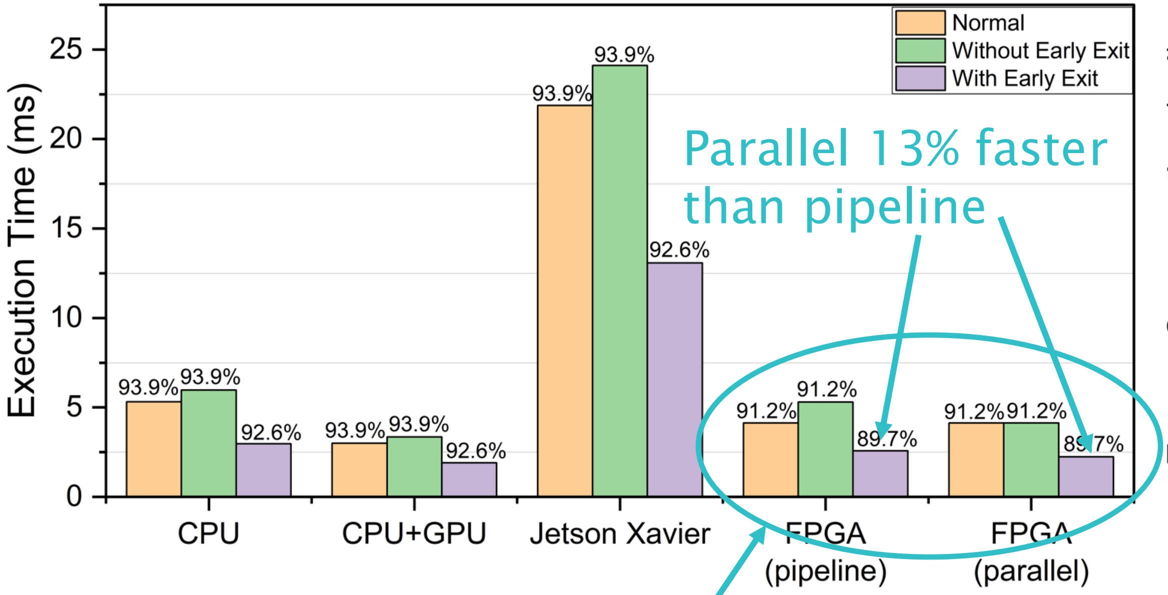


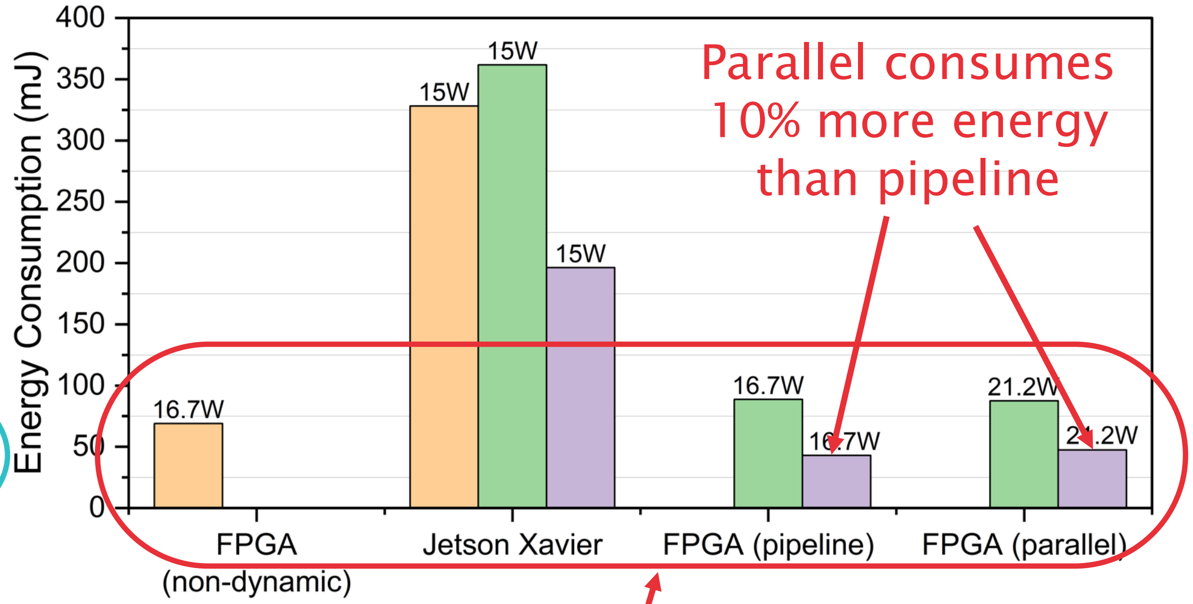
A. Dimitriou, L. Xun, J. Hare and G. V. Merrett, "Realisation of Early-Exit Dynamic Neural Networks on Reconfigurable Hardware," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

UNIVERSITY OF Southampton

ACCELERATING DYNAMIC NETWORKS - RESULTS

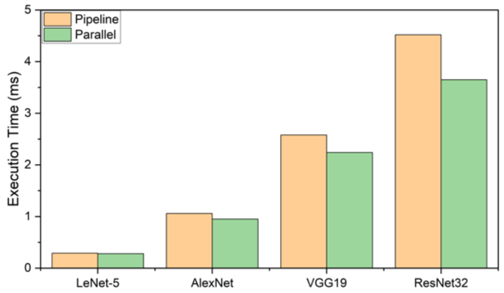
- VGG19 with BranchyNet on Cifar-10; Zynq UltraScale+ FPGA





Early-exiting speeds up inference by at least 1.4x, with less than 1.5% loss of accuracy

FPGA energy consumption reduced by 1.8x, despite the increase in power consumption.



- Similar trends across LeNet-5 (MNIST), AlexNet (CIFAR10), ResNet32 (CIFAR100) – for the latter, parallel 20% faster for 11% more energy

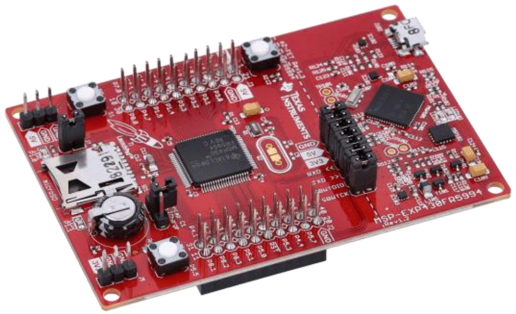
S. Teerapittayanon, B. McDanel, and H. T. Kung, “Branchynet: Fast inference via early exiting from deep neural networks,” ICPR, 2016.

A. Dimitriou, L. Xun, J. Hare and G. V. Merrett, "Realisation of Early-Exit Dynamic Neural Networks on Reconfigurable Hardware," in IEEE TCAD

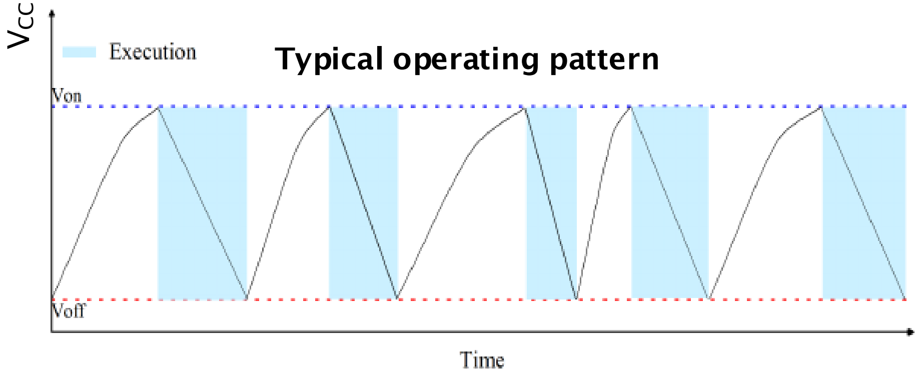
DYNAMIC INFERENCE ON MCUS

Can Dynamic Inference be effectively applied to MCUs?

- Constrained MCU-based systems powered from the environment
  - With minimal energy storage, the system operates intermittently



Available power varies over time and space

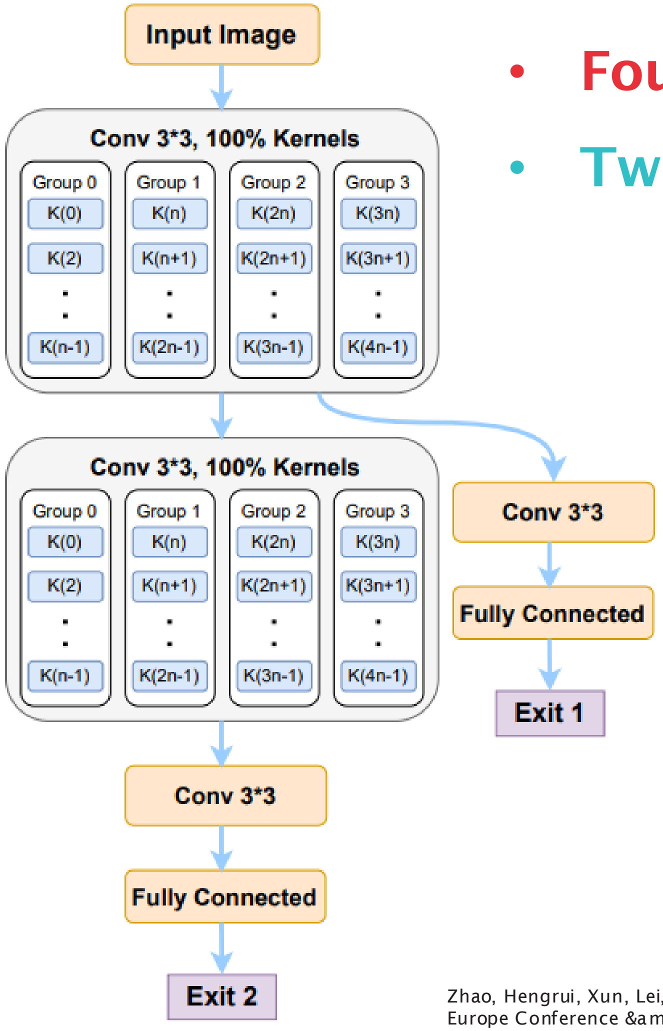


Dynamic DNNs- adapt to power constraints

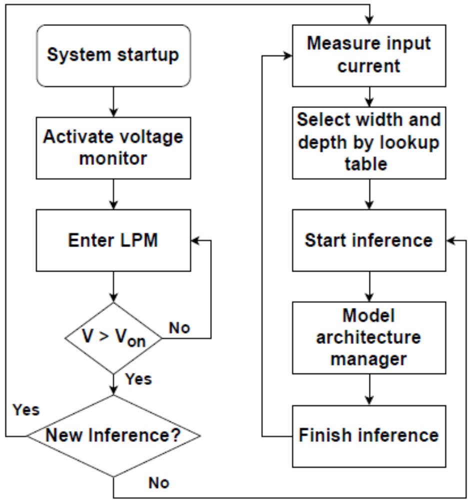
- Can dynamic DNNs offer a performance/latency trade-off for MCUs?
- Can we utilize this to enable systems to meet inference deadlines under variable/intermittent supply?

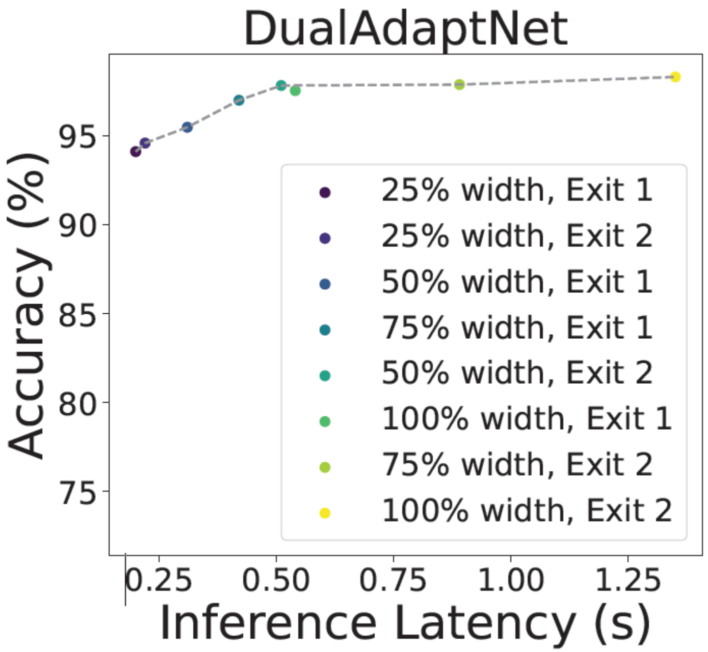
Zhao, Hengrui, Xun, Lei, Chauhan, Jagmohan and Merrett, Geoff (2024) Power- and deadline-aware dynamic inference on intermittent computing systems. In 2025 Design, Automation & Test in Europe Conference & Exhibition. IEEE. 7 pp . (In Press)

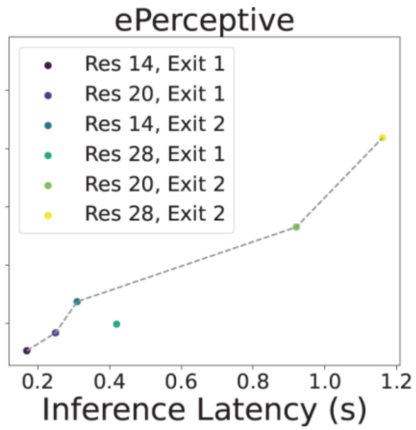
DYNAMIC INFERENCE ON MCUS: DualAdaptNet

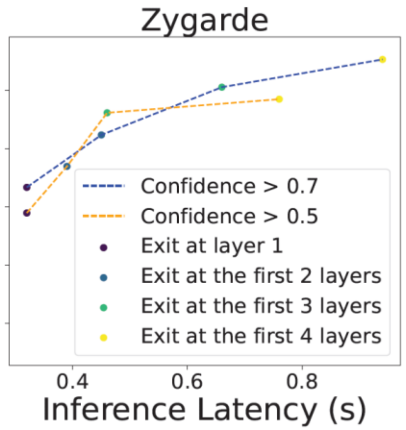


- **Four** widths: Conv kernels divided into 4 groups
- **Two** depths: Exit 1 & Exit 2





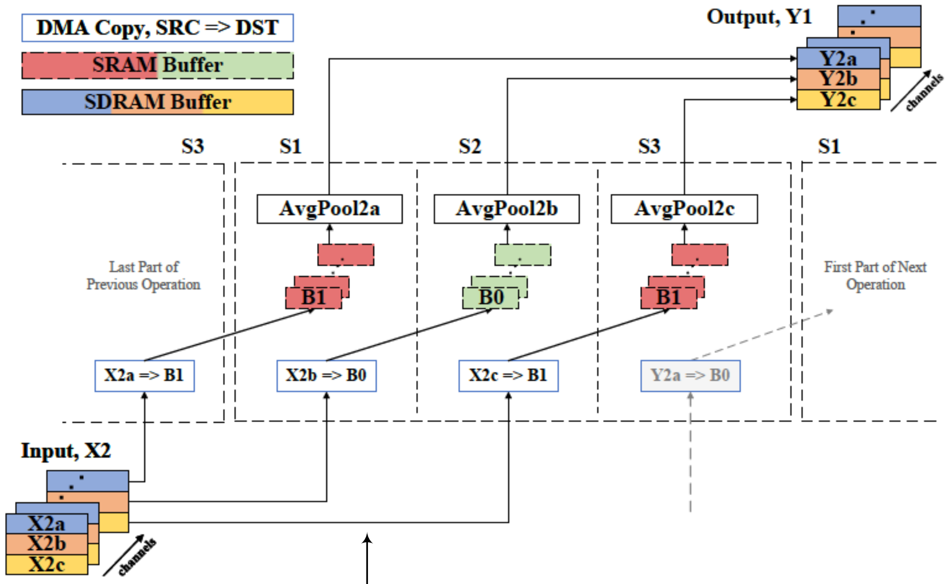


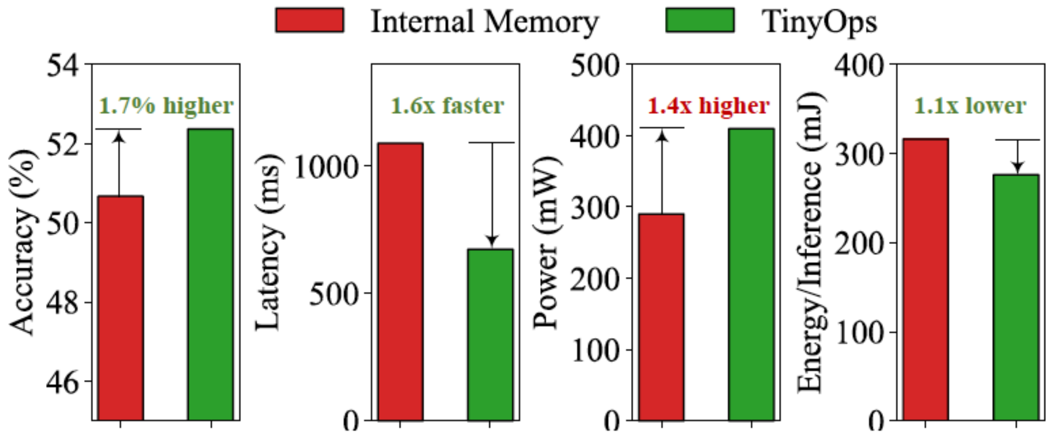


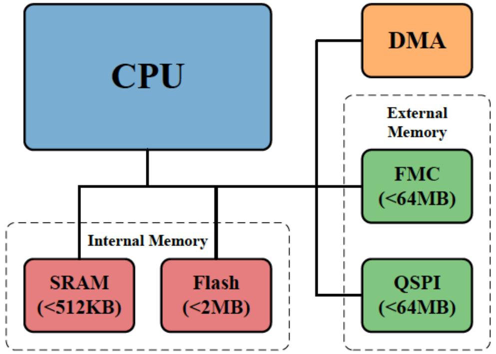
Zhao, Hengrui, Xun, Lei, Chauhan, Jagmohan and Merrett, Geoff (2024) Power- and deadline-aware dynamic inference on intermittent computing systems. In 2025 Design, Automation & Test in Europe Conference & Exhibition. IEEE. 7 pp . (In Press)

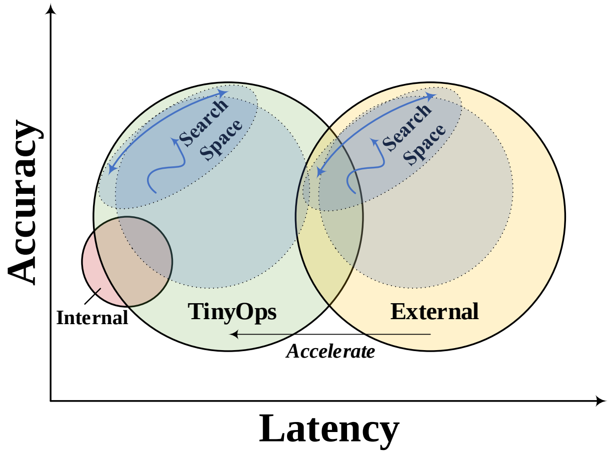
RECONSIDERING THE MCU DESIGN SPACE

- Majority of existing MCU approaches are constrained by the size of internal memory
- TinyOps enables MCU inference of large models in external memory with internal memory like latency
- ImageNet classification with 6% higher accuracy and 2.1x low inference latency









S. Sadiq, J. Hare, S. Craske, P. Maji and G. Merrett, "Enabling ImageNet-Scale Deep Learning on MCUs for Accurate and Efficient Inference," in IEEE Internet of Things Journal, vol. 11, no. 7, pp. 11471-11479, 1 April1, 2024
Sadiq, S., Hare, J., Merrett, G., Prasun, P., & Craske, S. J. (2024). U.S. Patent Application No. 17/813,396.

CONCLUSIONS

- Efficient DNN deployment demands anticipating runtime changes, not just initial optimization.
- Dynamic DNNs enable flexibility and offer benefits, but highlight need for adaptable hardware, compilers, mapping, etc.
- We need improved approaches to manage resources in systems while providing *acceptable* performance

> *"Companies will learn to make trade-offs between accuracy and computational efficiency, though that will have unintended, and antisocial, consequences too"*
>
> *John Naughton: Emeritus Professor of the Public Understanding of Technology at the Open University*

Photo by Patrick Schneider on Unsplash

YOUR QUESTIONS

Professor Geoff Merrett

e: gvm@ecs.soton.ac.uk
w: www.geoffmerrett.co.uk
@g_merrett

UNIVERSITY OF Southampton

UNIVERSITY OF
Southampton