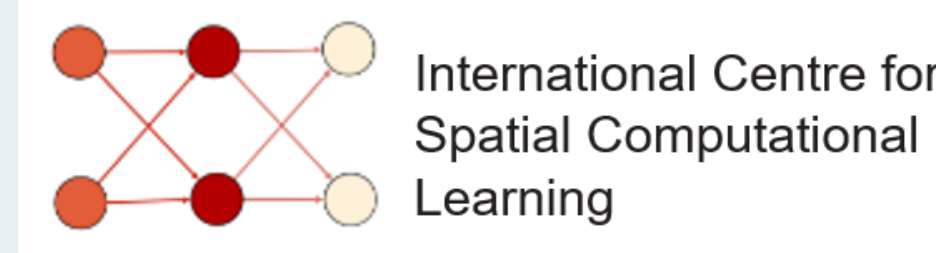


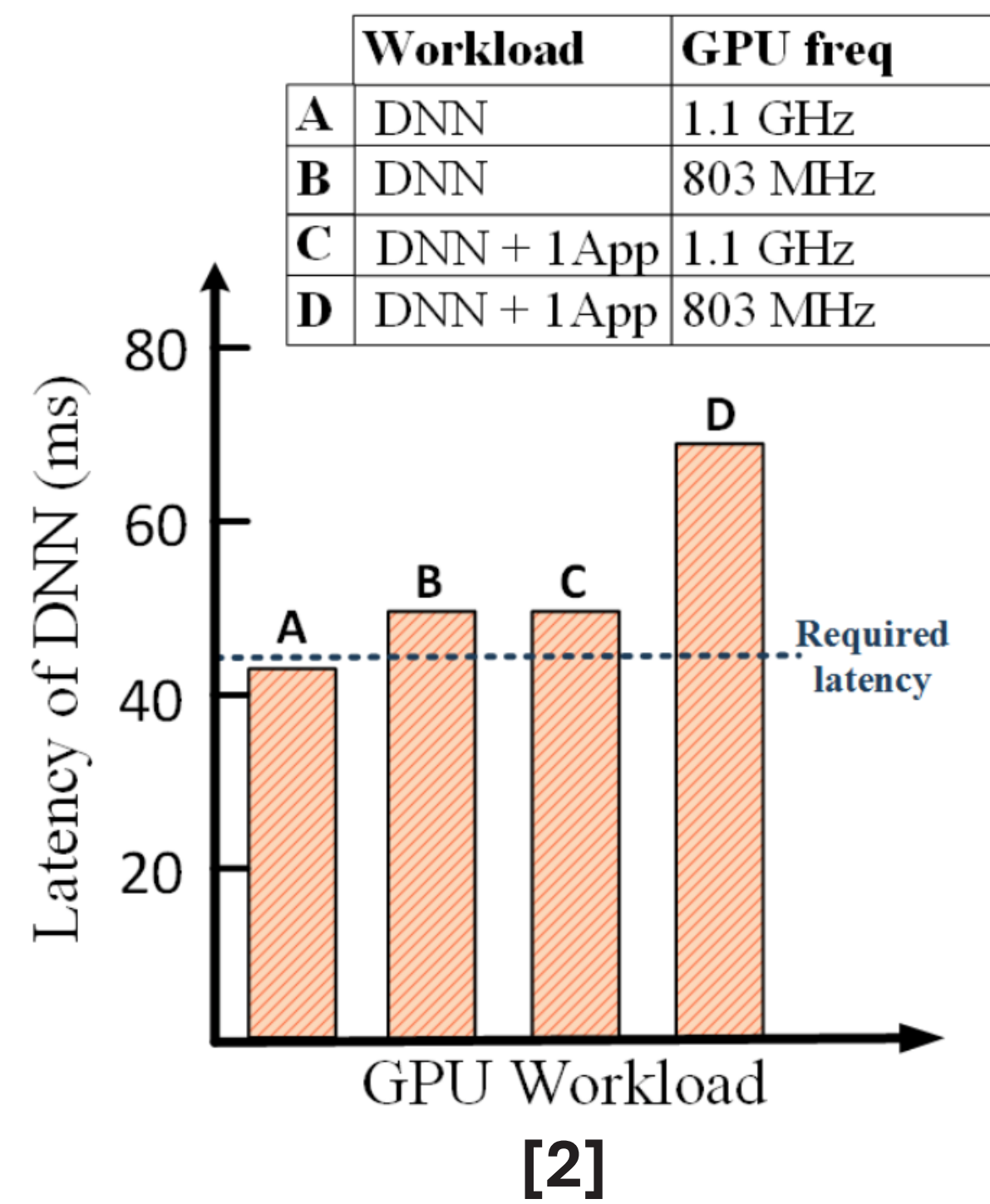
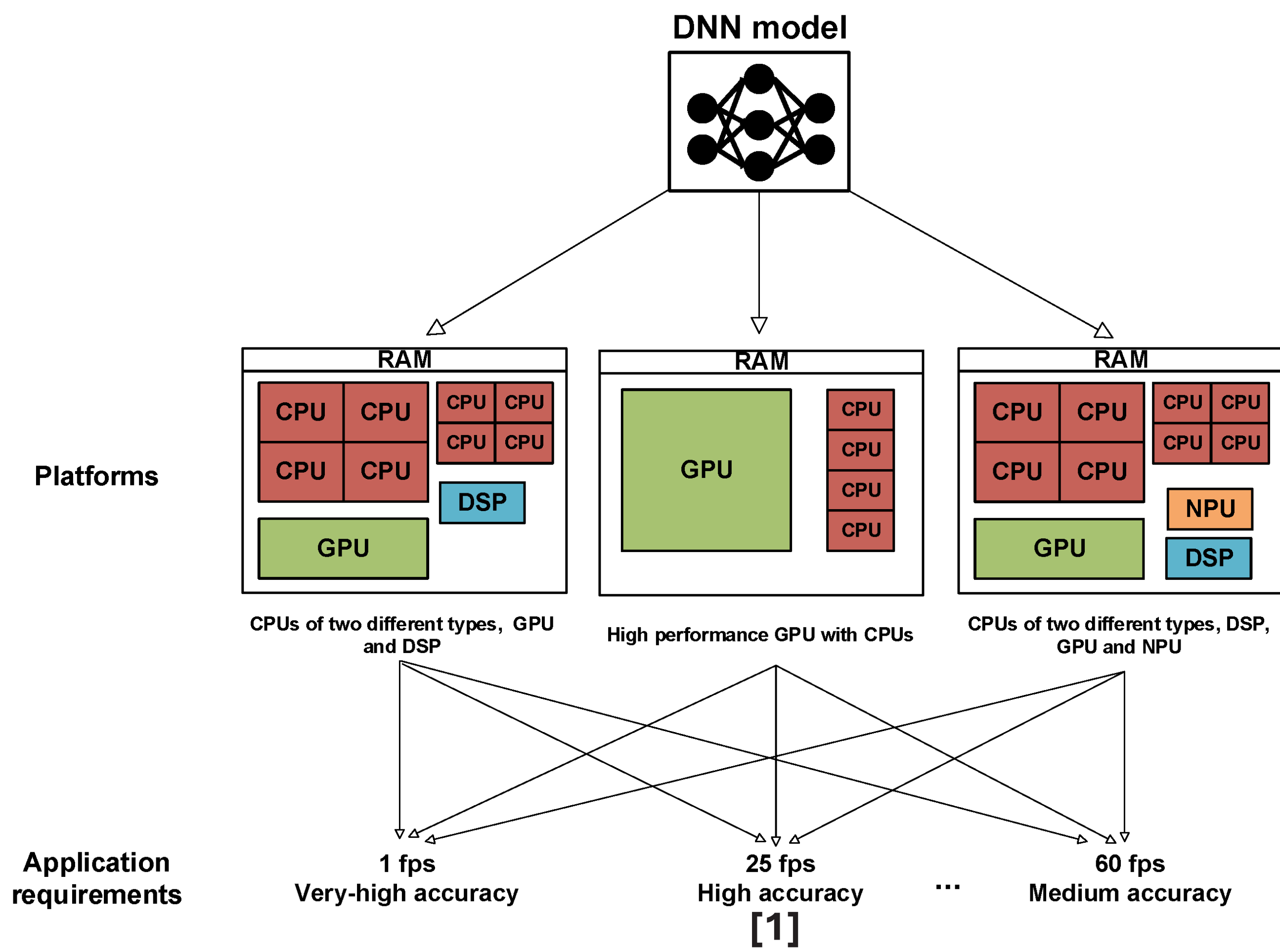
Lei Xun, Jonathon Hare, Geoff V. Merrett

School of Electronics and Computer Science, University of Southampton, UK

[l.xun@soton.ac.uk](mailto:l.xun@soton.ac.uk) / Lei Xun (in)



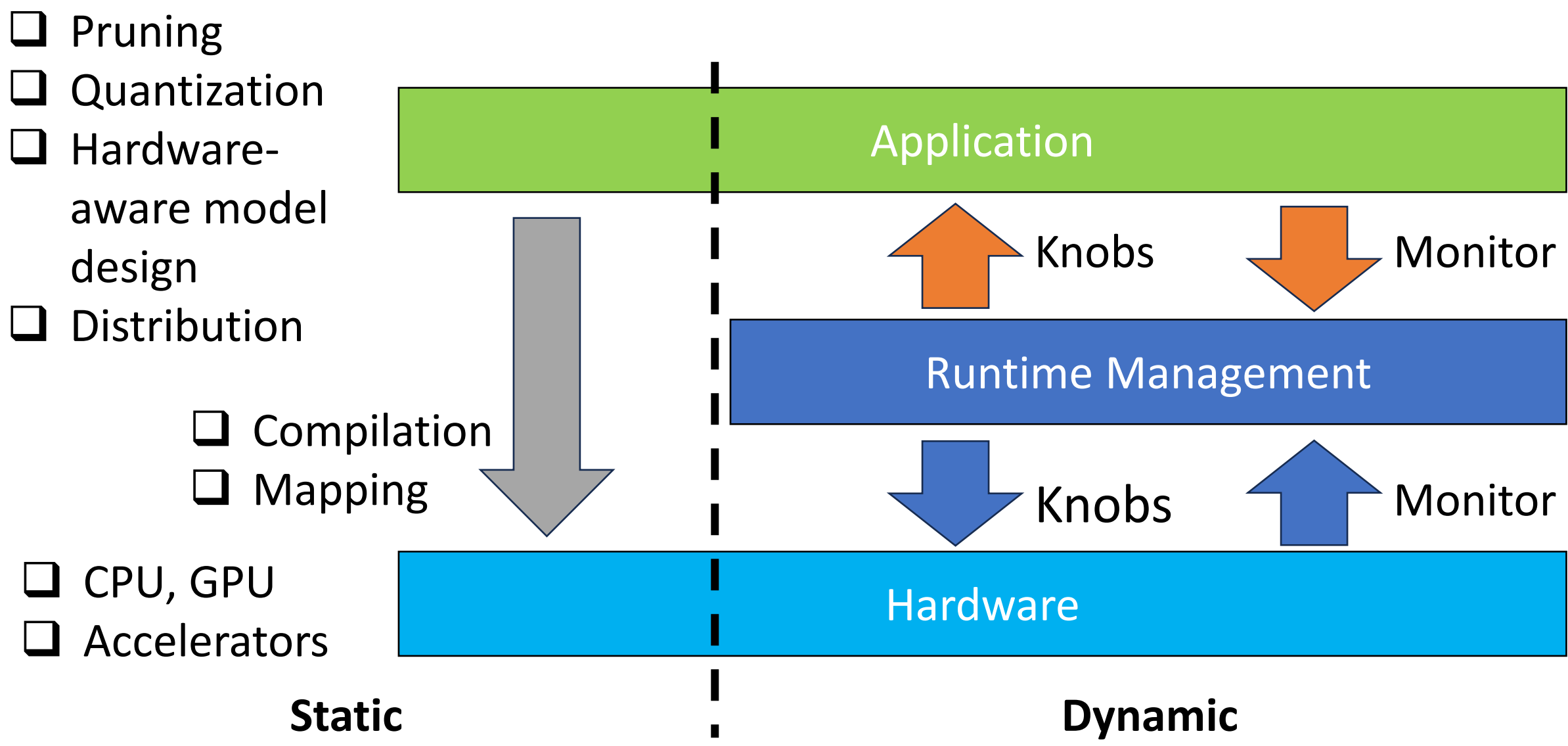
## Introduction



- DNN inference is increasingly being executed on mobile and embedded devices, thanks to its low latency and improved privacy. However, DNN models are both computationally and memory access intensive.
- Efficient deployment of DNN models faces two primary challenges at runtime:
  - [Runtime Hardware Availability]** Modern SoCs encounter fluctuating hardware resource availability due to dynamic workloads, varying core combinations, and changes in voltage and clock frequencies.
  - [Runtime Application Variability]** A single DNN model must serve various applications, each with different performance needs (e.g., ChatBots require low latency, while translation focuses on accuracy).
- Static optimization methods fail to address runtime challenges, as they cannot adapt to dynamic changes in hardware availability or application/user-specific performance needs.

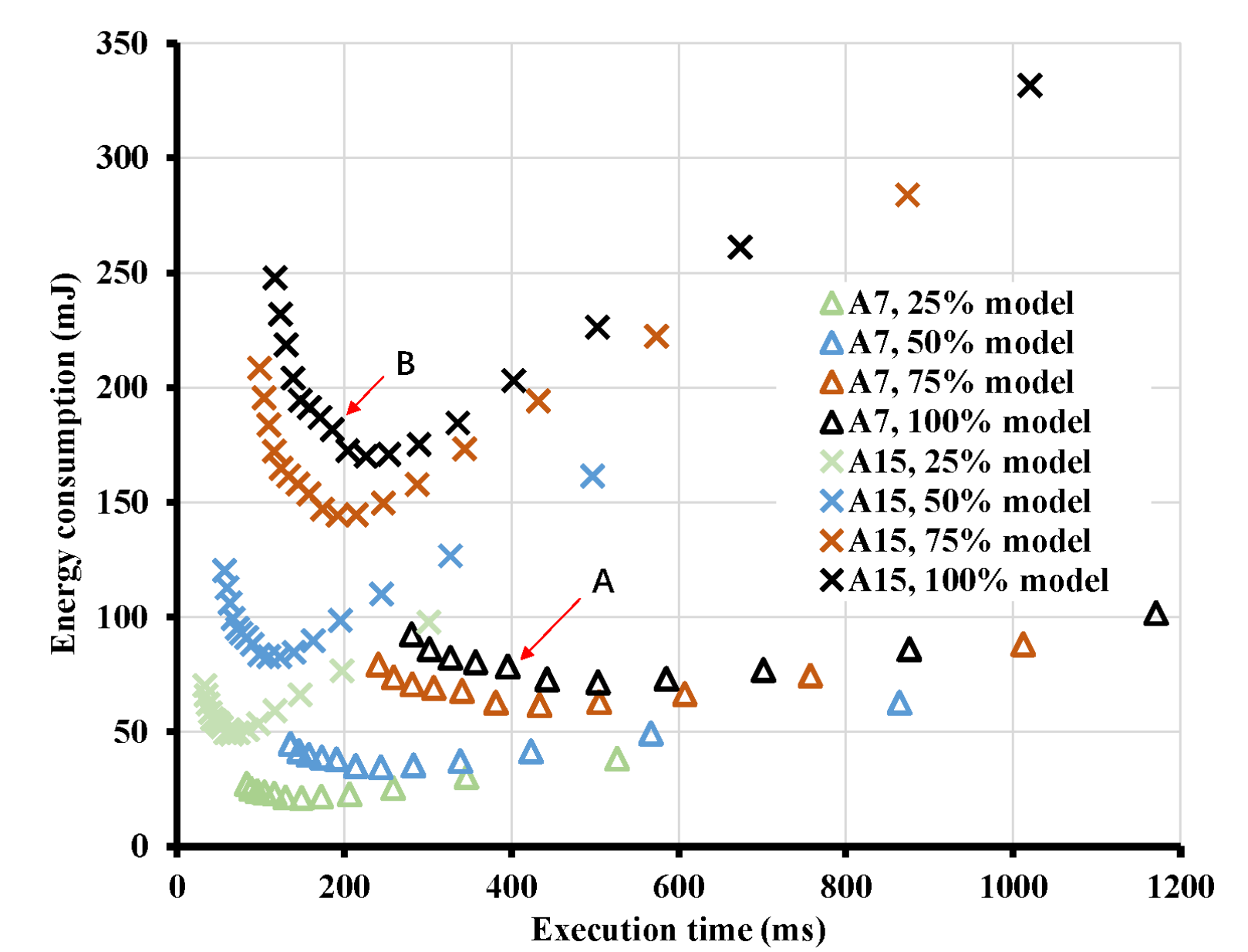
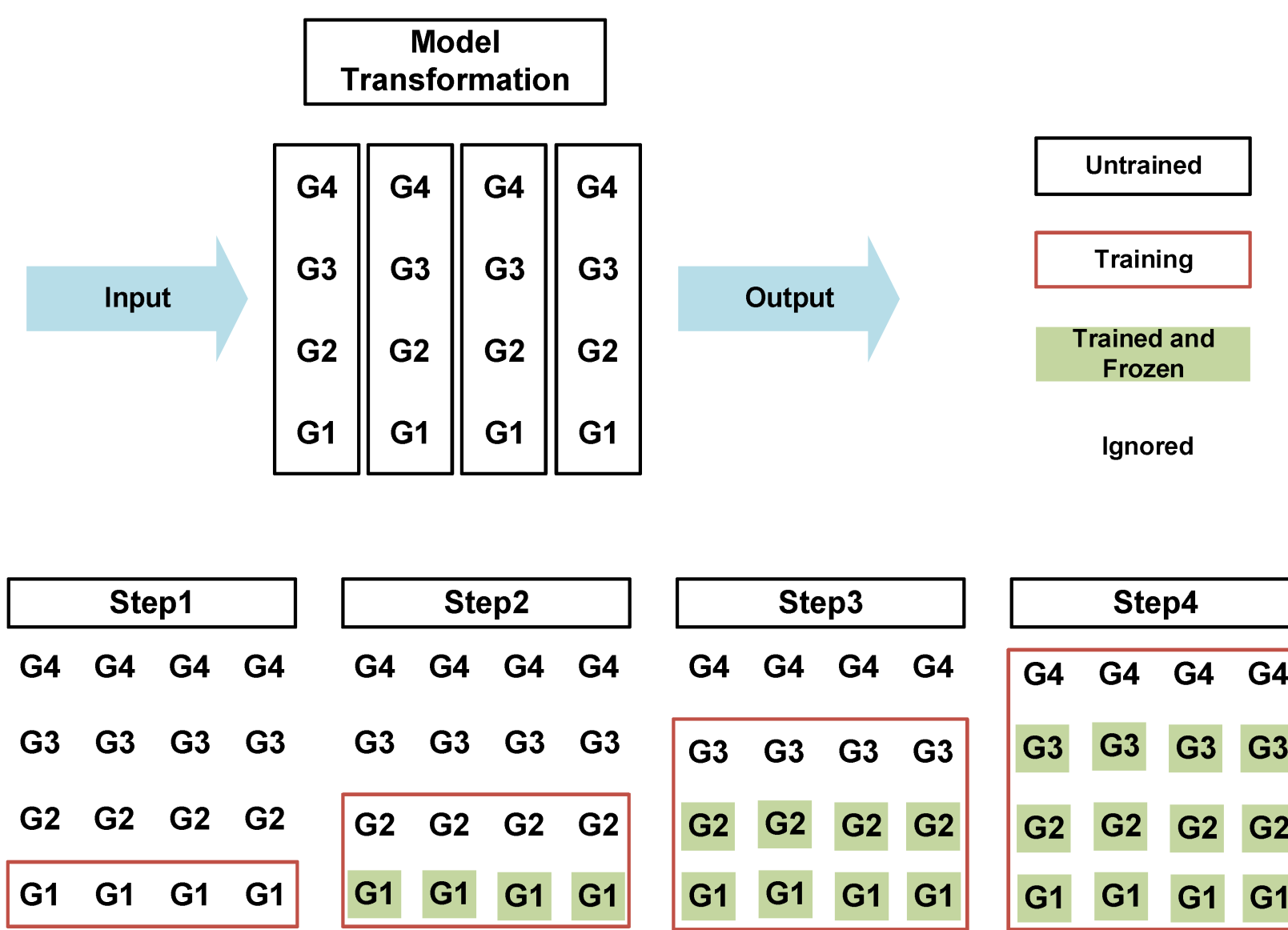
## Vertical System-level Optimization at Runtime [1,5]

- Static system-level optimization has no adaptation to runtime challenges.
- Traditional runtime resource management primarily focuses on hardware adjustments (e.g., DVFS, task mapping), treating DNN models as general applications and missing domain-specific opportunities.
- In our research, we have co-designed Dynamic DNNs that facilitate runtime adjustments for both algorithms and hardware.



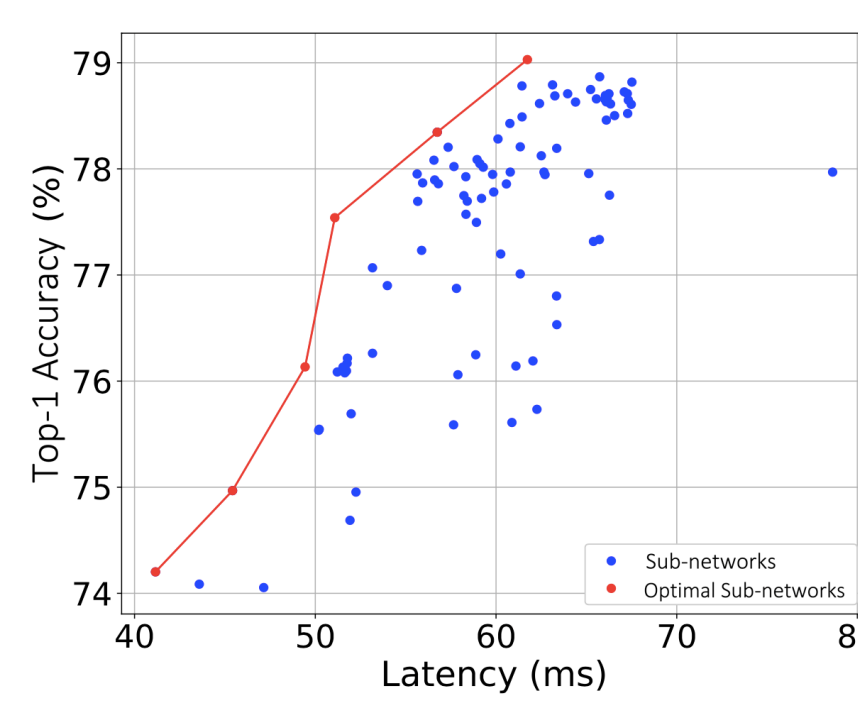
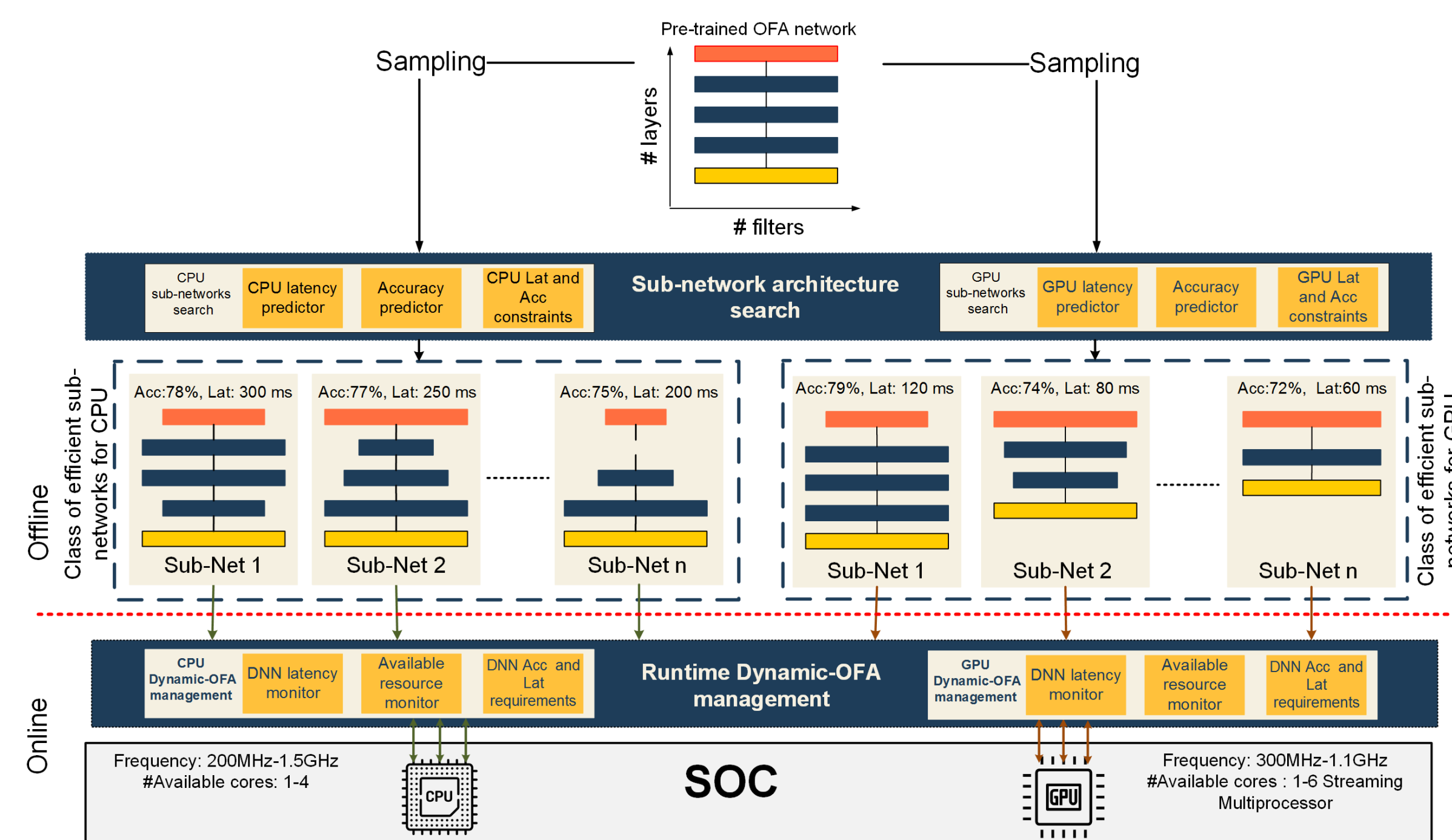
## Incremental Training and Group Convolution Pruning [1,3,4]

- Convolution layers are divided into groups, which are trained incrementally.
- A dynamic neural network with four sub-network configurations is created. Each sub-network offered unique accuracy, latency, and power/energy trade-offs.

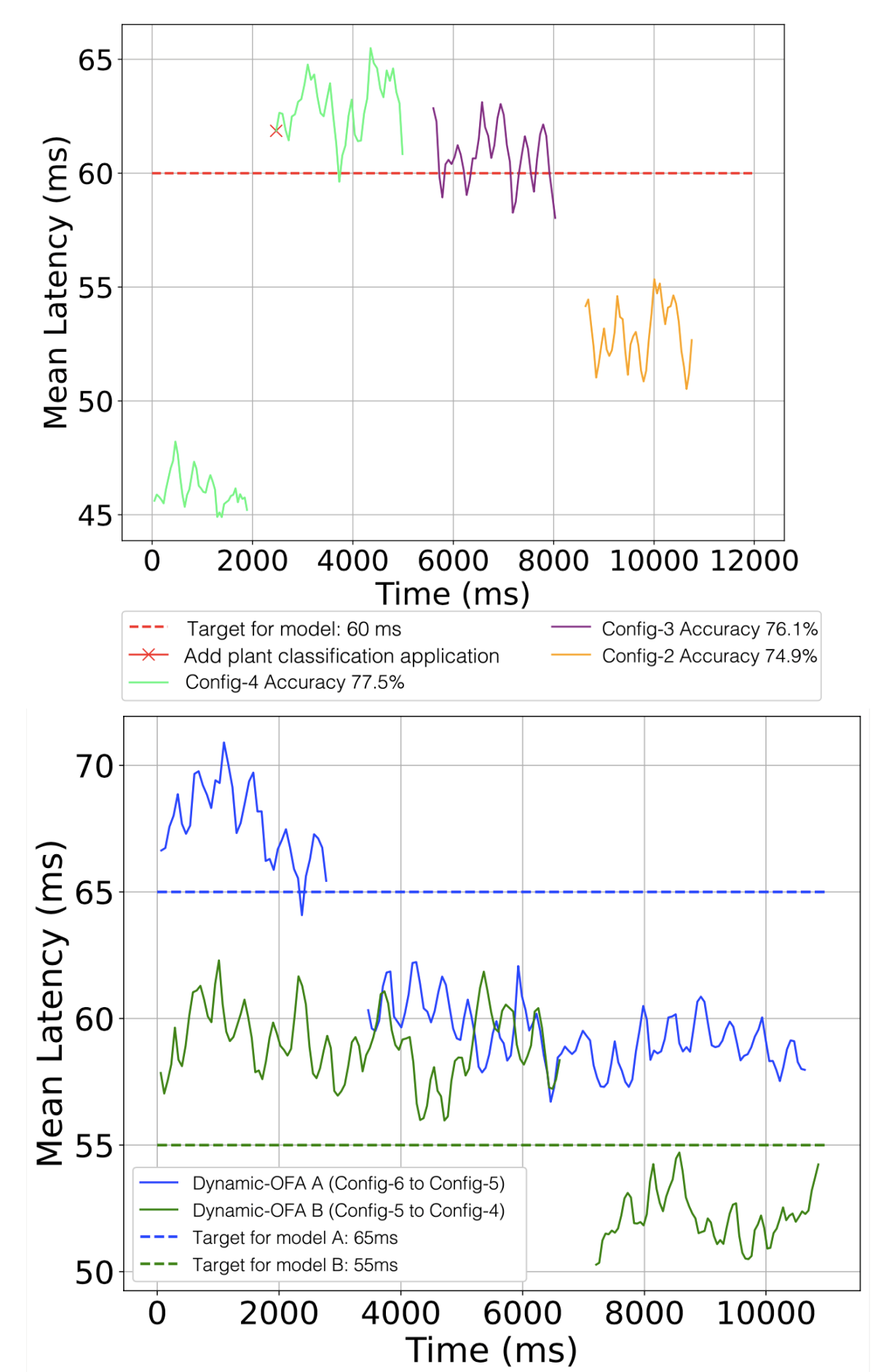
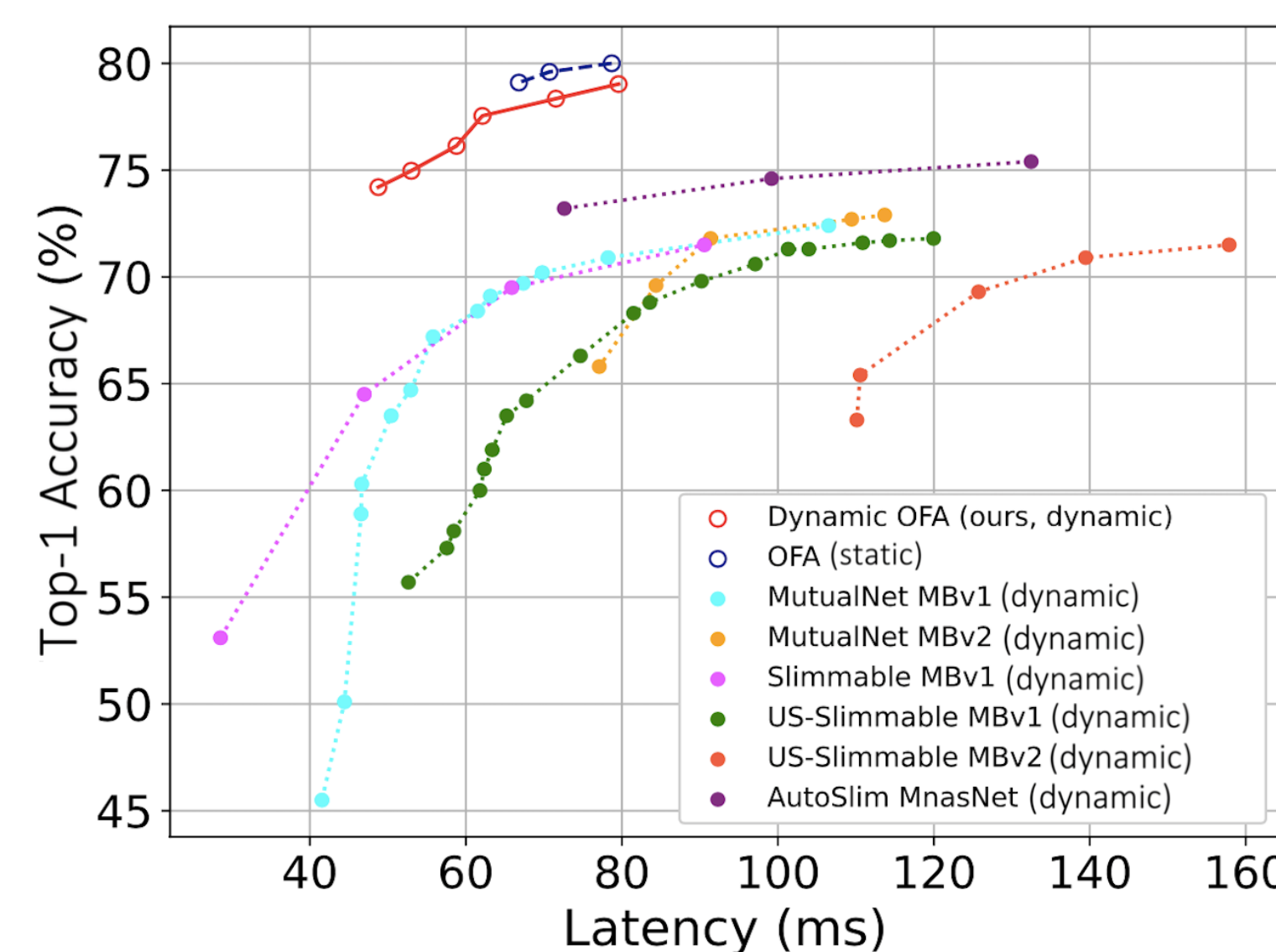


## Dynamic Super-networks [2,6]

- Sub-networks are sampled from a pre-trained Super-network, and Dynamic DNN is created using the sub-networks located on the Pareto-front of performance trade-off.
- The sampling process is conducted separately for CPUs and GPUs, as the most efficient sub-network architectures differed for these heterogeneous computing resources. Each sub-network provided unique accuracy, latency, and power/energy trade-offs.
- During runtime, the sampled sub-networks can be switched in real-time to meet the dynamic changes in hardware availability or application/user-specific performance needs.



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



## Conclusion

- Efficient DNN deployment demands anticipating runtime changes, not just initial optimization.
- Dynamic DNNs offer crucial flexibility but necessitate a full-stack approach for true efficiency.
- Deploying Dynamic DNNs showcases the benefits and reveals the need for adaptable hardware, compilers, and dynamic mapping, etc.

## References

- Lei Xun, Long Tran-Thanh, Bashir M Al-Hashimi, and Geoff V Merrett. Optimising Resource Management for Embedded Machine Learning. In Design, Automation and Test in Europe Conference (DATE), 2020.
- Wei Lou\*, Lei Xun\*, Amin Sabet, Jia Bi, Jonathon Hare, and Geoff V Merrett. Dynamic-OFA: Runtime DNN Architecture Switching for Performance Scaling on Heterogeneous Embedded Platforms. In Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021.
- Lei Xun, Long Tran-Thanh, Bashir M Al-Hashimi, and Geoff V Merrett. Incremental Training and Group Convolution Pruning for Runtime DNN Performance Scaling on Heterogeneous Embedded Platforms. In ACM/IEEE 1st Workshop on Machine Learning for CAD (MLCAD), 2019.
- Lei Xun, Bashir M Al-Hashimi, Jonathon Hare, and Geoff V Merrett. Runtime DNN Performance Scaling through Resource Management on Heterogeneous Embedded Platforms. In tinyML EMEA Technical Forum, 2021.
- Lei Xun, Bashir M Al-Hashimi, Jonathon Hare, and Geoff V Merrett. Dynamic DNNs Meet Runtime Resource Management on Mobile and Embedded Platforms. In 4th UK Mobile, Wearable and Ubiquitous Systems Research Symposium (MobiUK), 2022.
- Hishan Parry, Lei Xun, Amin Sabet, Jia Bi, Jonathon Hare, and Geoff V Merrett. Dynamic Transformer for Efficient Machine Translation on Embedded Devices. In ACM/IEEE 3rd Workshop on Machine Learning for CAD (MLCAD), 2021.

## Acknowledgement

These works were supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/S030069/1.

