tinyML® Summit

Enabling Ultra-low Power Machine Learning at the Edge

Perspectives & Challenges for TinyML Hardware: a System-Level View

Dr. Francesco Conti, 30 Mar 2022

www.tinyML.org

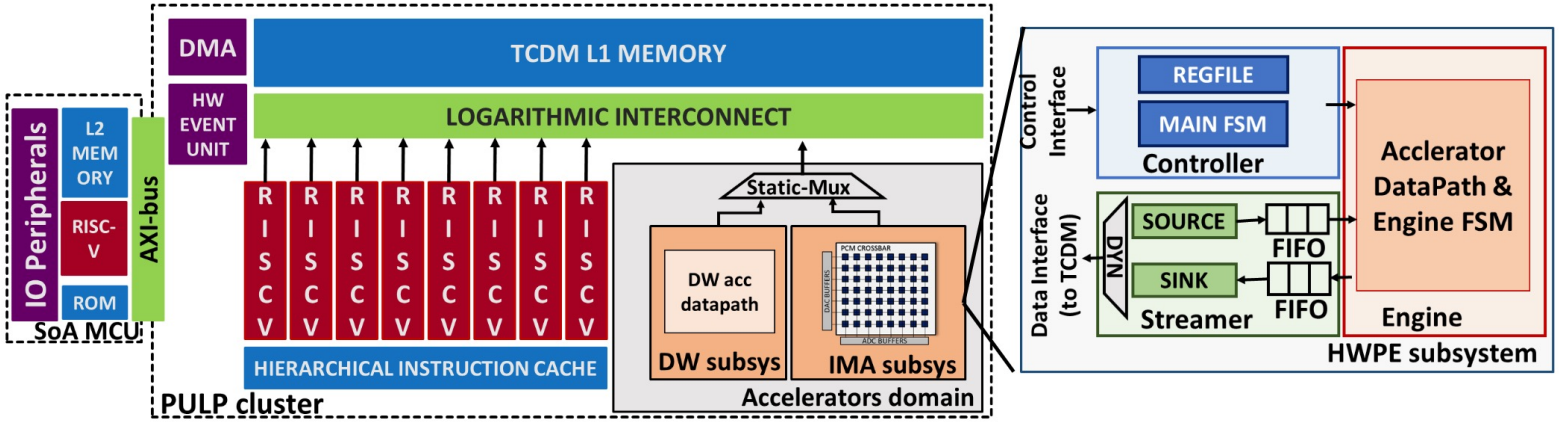
TINY ML SUMMIT

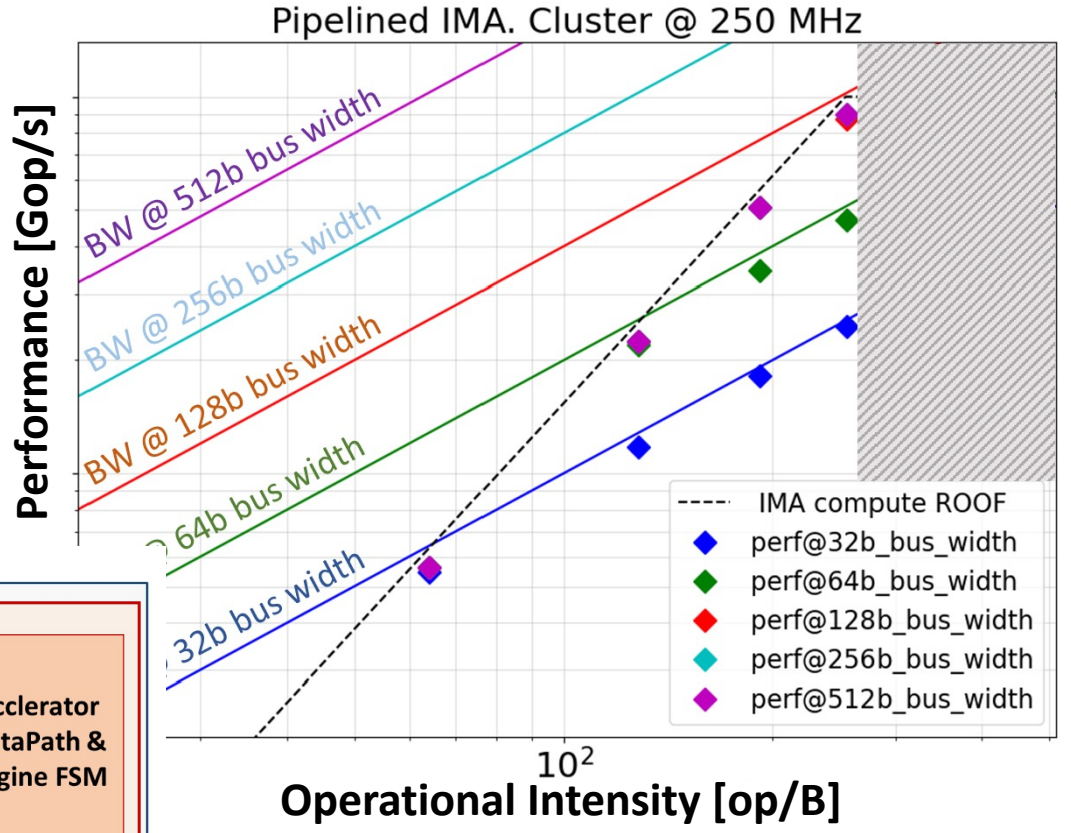
tinyHW Panel

Abstract: TinyML inference on ultra-low-power hardware is finally here to stay. But what are the perspectives for the future of this technology? Will Analog In-Memory Computing take up the scepter of crucial technology for TinyML from digital accelerators? Will TinyML hardware evolve to support learning as well as inference? In this session, we discuss the main challenges and opportunities ahead for TinyML hardware from the perspective of using it in real embedded systems.

Integrating AIMC @ system level

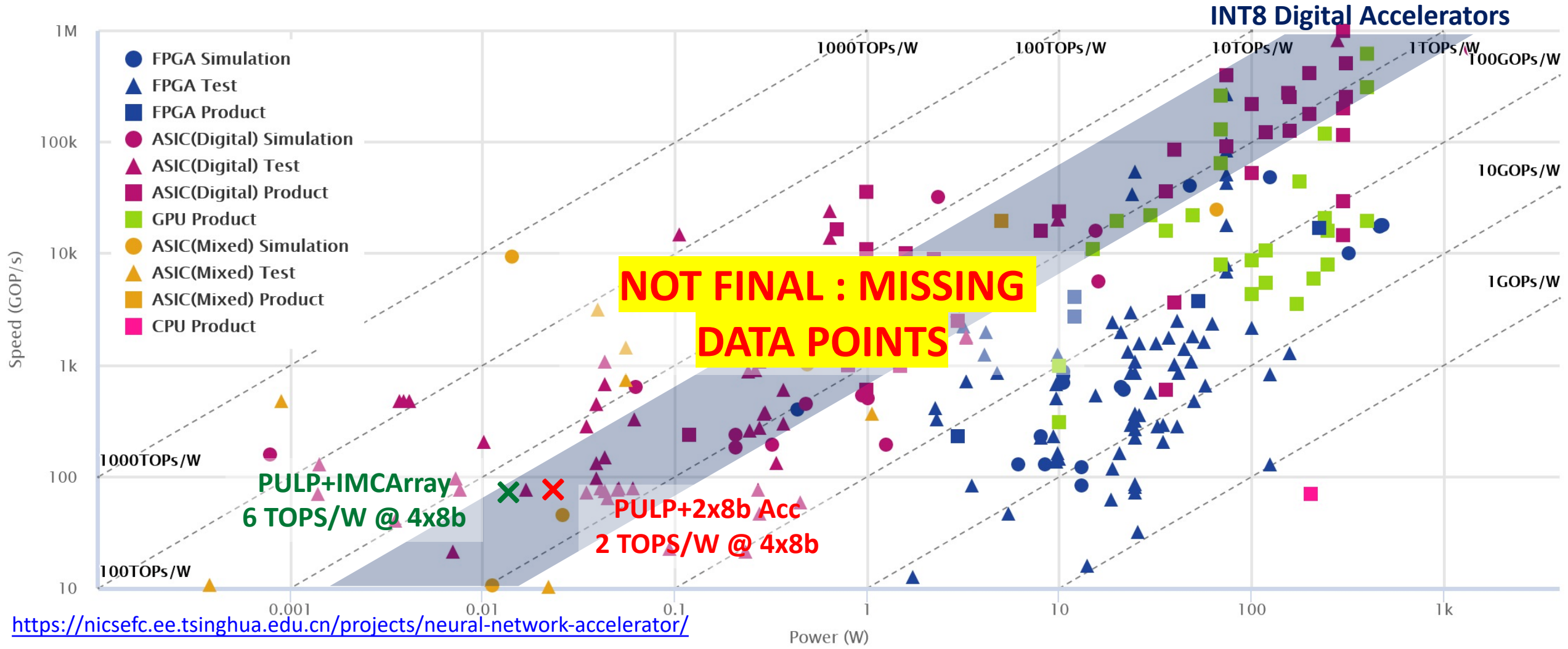
AIMC: Real performance depends on **system integration** (e.g., bandwidth) but also on **workload** (even more than digital?)





[G. Ottavi et al., End-to-end 100-TOPS/W Inference with Analog In-Memory Computing: Are We There Yet?, AICAS 2021]
[A. Garofalo et al., A Heterogeneous IMC Cluster for Flexible End-to-End Inference of Real-World DNNs https://arxiv.org/pdf/2201.01089.pdf]

AIMC or Digital?



https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/

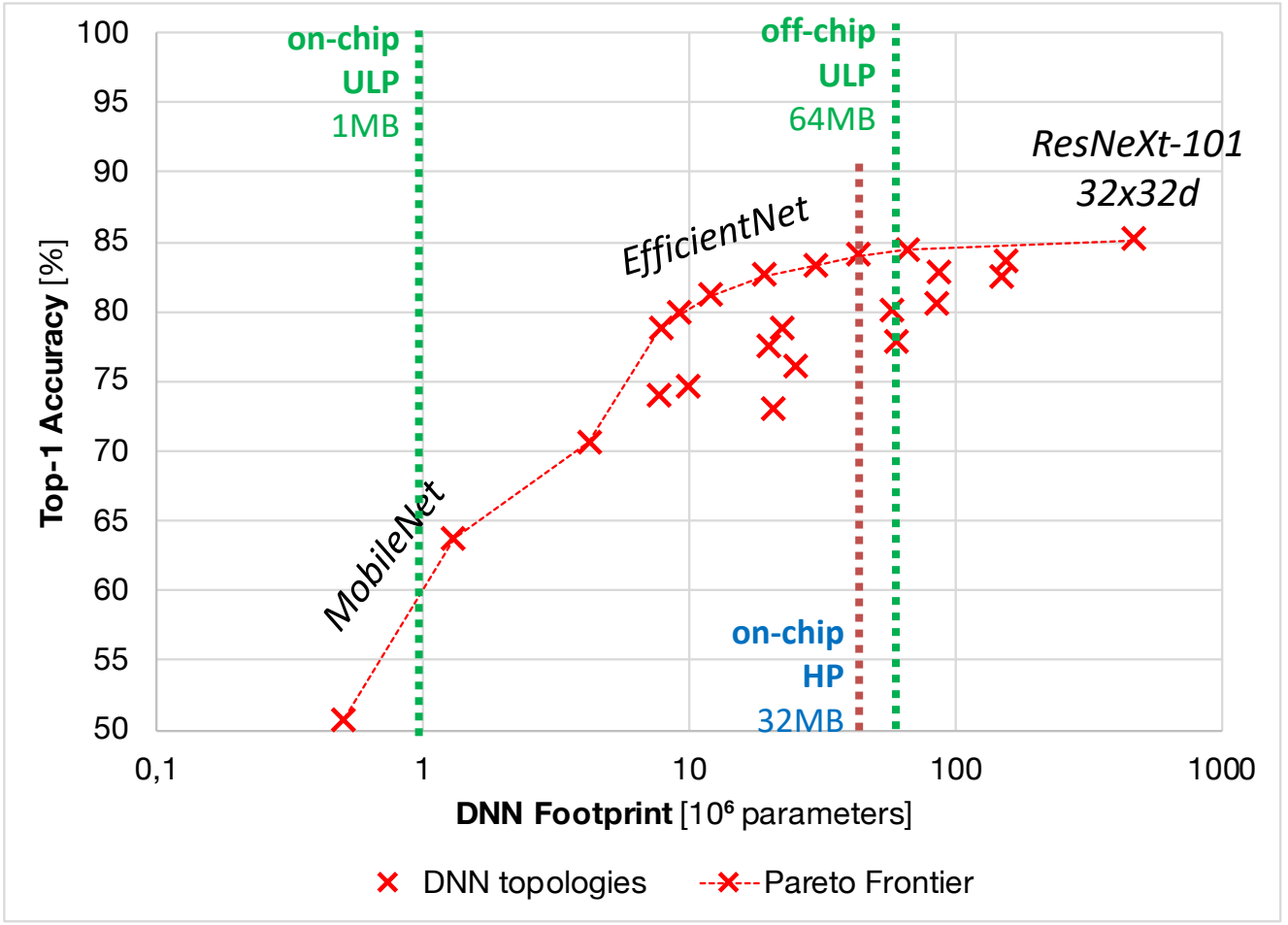
What technology will emerge as go-to for TinyML hardware? Is the efficiency gain of AIMC enough to move TinyML systems in this direction?

TINY ML SUMMIT

DNN Memory Footprint (2019/20)

ImageNet Top-1 Accuracy
vs **Memory Footprint**

- Most entries > 10 MB
- Pareto Frontier Acc vs Memory
  (from 50% @ 0.5Mparam
  to 85% @ 445 Mparam)
- Not necessarily representative of real-world problems (might be more complex)
- **<u>Almost always require off-chip DRAM!</u>**



Memory *is one of the primary constraints when looking at TinyML hardware from a system perspective*

How can we design TinyHW that supports upscaling the complexity of TinyML (e.g., larger nets)?

TINY ML SUMMIT

Learning on device?

The next level of "intelligence" is being able to capture and retain knowledge from the environment with on-device learning and adaptation.

What steps are necessary to get there?

tinyML Summit 2022 Sponsors

Copyright Notice

This presentation in this publication was presented as a tinyML® Summit 2022. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org