Distributed Transformer Inference on Low-Power MCUs

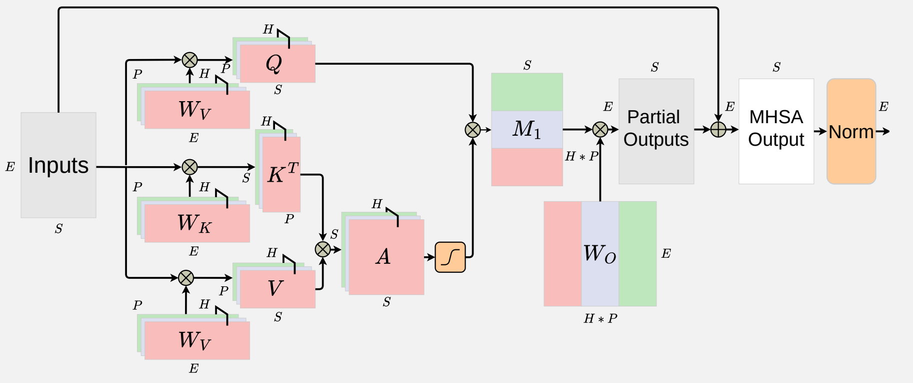
Severin Bochem[1], Victor J.B. Jung[2], Arpan Suravi Prasad[2], Francesco Conti[3], Luca Benini[2,3]
[1]D-ITET, ETH Zurich; [2]Integrated Systems Laboratory, ETH Zurich; [3]DEI, University of Bologna

1 Introduction

How to deploy **real-world Transformers** on low-power wearable devices? We propose a partitioning scheme on a distributed systems of MCUs with **minimal off-chip traffic** reaching **super-linear speedup** compared to single MCU system.

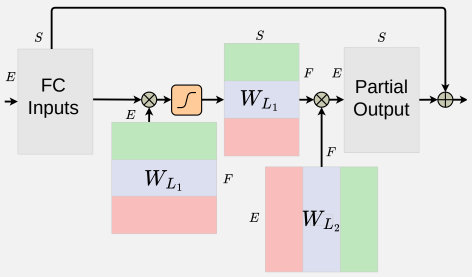
2 Method Overview

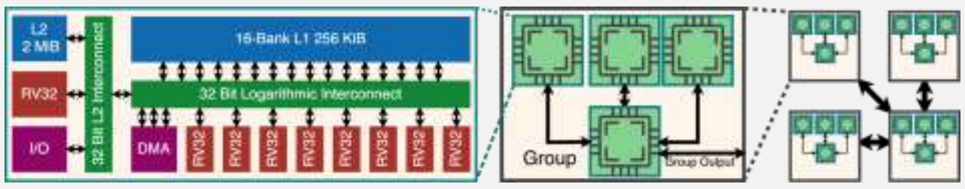
- Minimize **off-chip communication**, avoid **weight duplication**, and **synchronize only twice**.



S: Sequence Length | E: Embedding Dimension | P: Head Dimension | H: NumHeads

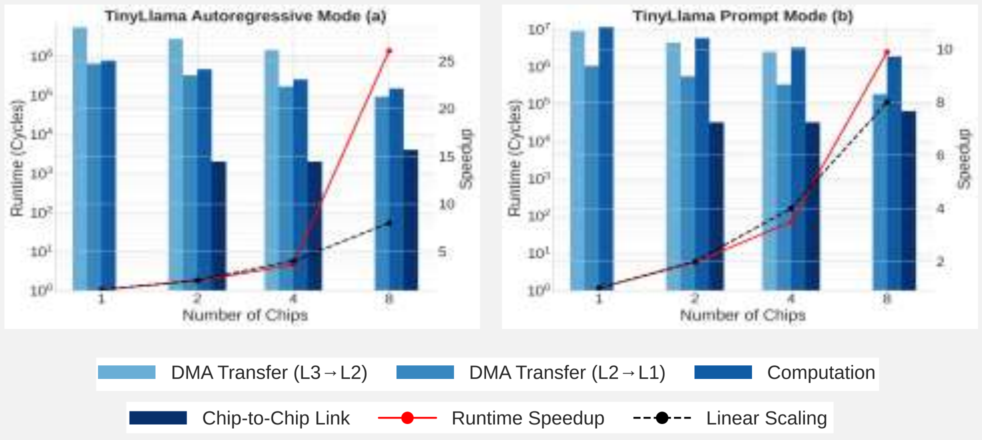
- Parallelize all **GEMMs**, **Softmax**, and **Activations**.
- **Accumulate** Partial Outputs and compute **Layer Norm** on one chip.
- Use MIPI interface to communicate between MCUs.



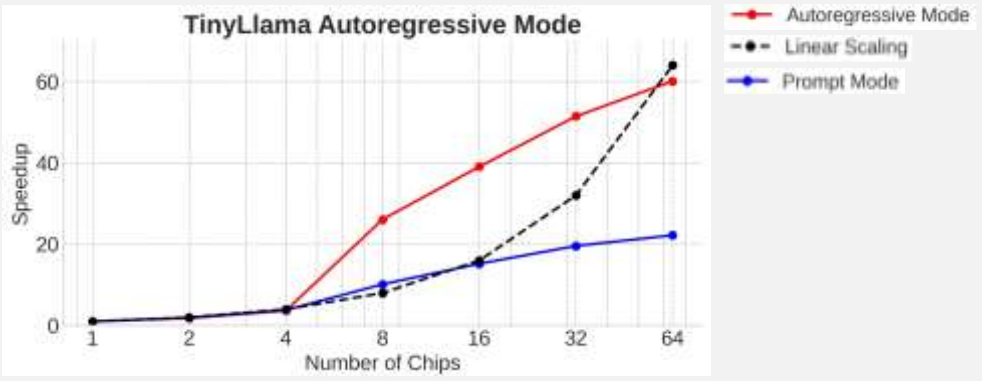


- Using Siracusa-like [1] MCUs, with 2MiB L2 SRAM memory.
- Hierarchical reduction in groups of chips to improve scalability.

3 Super-Linear Speedup on 8-MCUs

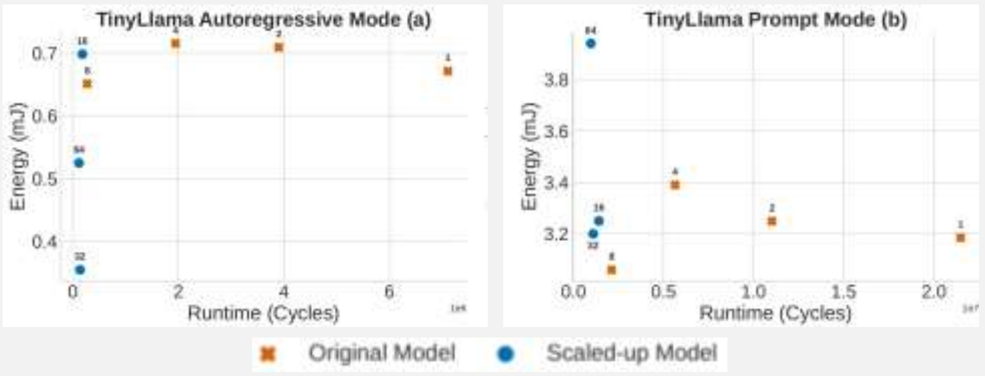


4 Scale-up Study



- Scaled up TinyLlama [2] from 8 to 64 heads, while keeping other parameters constant.
- Partitioning **falls behind linear scaling** at **64 MCUs** for Autoregressive mode and **32 MCUs** for Prompt mode.

5 Runtime/Energy Tradeoff



- We measured the runtime/energy tradeoff for the original and scaled-up TinyLlama in Autoregressive and Prompt mode with 1 to 64 MCUs.
- The **8 MCUs** configuration is **optimal for the original model.**
- The Autoregressive mode benefits maximally from partitioning as it is **more memory bound than the Prompt mode**.

6 Conclusion

We proposed and benchmarked a Transformer partitioning scheme for edge devices, where running the layers **solely from on-chip memories** leads to **super-linear speedup**. With the 8 MCUs system, we achieve a **26x speedup** on the Autoregressive TinyLlama model.

References

1. A. S. Prasad, M. Scherer, F. Conti et al., “Siracusa: A 16 nm heterogenous RISC-V SoC for extended reality with at-MRAM neural engine,” IEEE Journal of Solid-State Circuits, 2024.
2. M. Scherer et al., “Deeploy: Enabling Energy-Efficient Deployment of Small Language Models on Heterogeneous Microcontrollers,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2024.

github.com/pulp-platform/Deeploy