

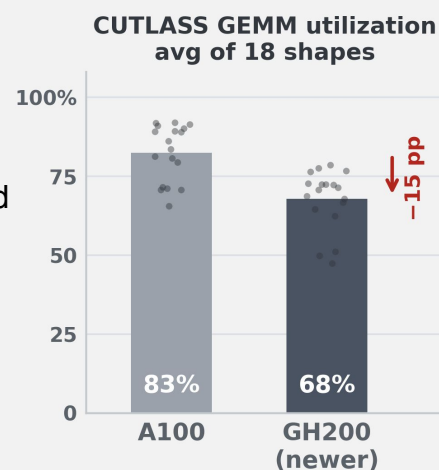
Design in Tiles: Automating GEMM Deployment on Tile-Based Many-PE Accelerators

Aofeng Shen¹, Chi Zhang¹, Yakup Budanaz², Alexandru Calotoiu², Bowen Wang¹, Torsten Hoefer², Luca Benini¹

¹Integrated Systems Laboratory, ²Department of Computer Science — ETH Zürich, Zürich, Switzerland

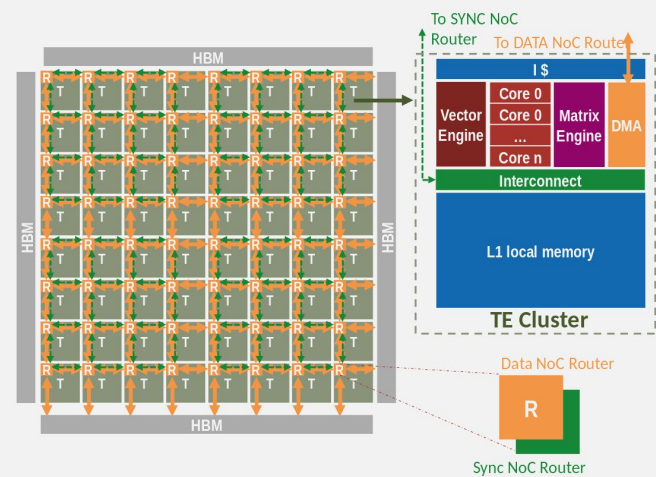
1 Motivation

- GPUs lose utilization at scale:** HW-managed caches give obscure dataflow control (GH200 < A100)
- Can tile-based accelerators do better?** With their SW-managed L1 + NoC collectives + distributed HBM promise explicit dataflow control [1]
- Mapping is HW-coupled:** manual kernel tuning impractical, and **no counterpart of CUTLASS / DeepGEMM** exists yet → DiT fills the gap



2 SoftHier: executable accelerator model

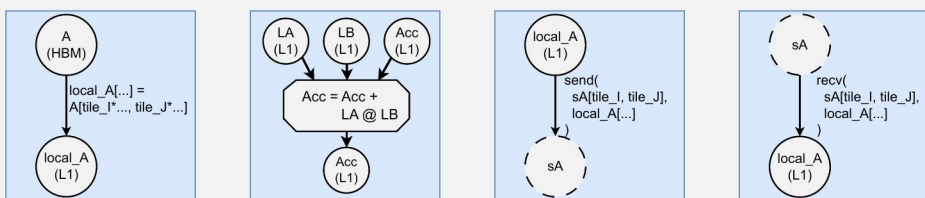
- RTL-calibrated** (<10% timing error), built on GVSoc; HBM via DRAMSys5
- Mask-based NoC collectives:** broadcast / reduce over arbitrary tile groups without software fan-out



3 DiT: automated end-to-end deployment

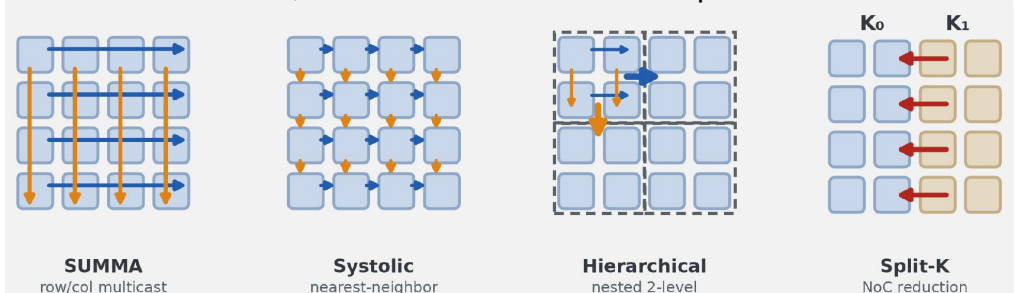


- Deployment via DaCe [2]:** SDFG IR captures per-PE workload, data movement & inter-tile communication
- Automated exploration:** enumerate → prune → simulate 300–1000+ schedules per shape; picks the best
- SDFG building blocks** — copy · MMAD · stream push / pop completely describe all collective communication patterns:



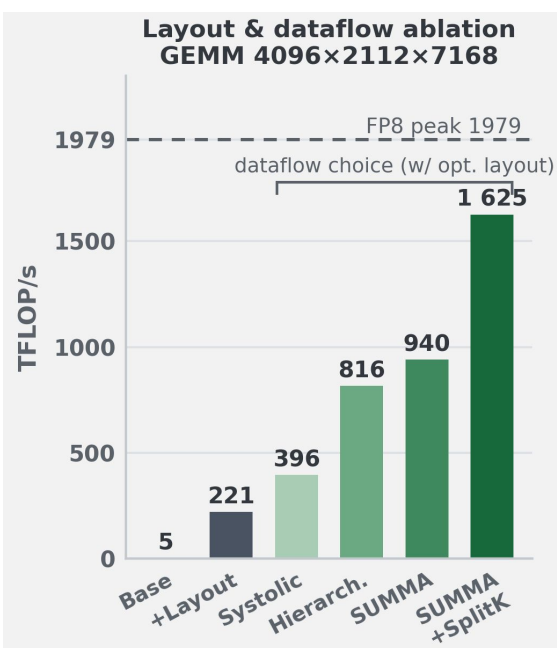
4 Design in tiles: schedule abstraction

- Tiling & mapping** — output-stationary 2D, or 3D split-K with NoC reduction; logical cluster remapping (e.g. 32×32 → 1×1024)
- Data layout** — split + placement schemes distribute matrices across HBM channels, avoiding contention
- Dataflow** — BSP supersteps (compute · communicate · barrier), double-buffered; four mask-based collective patterns:

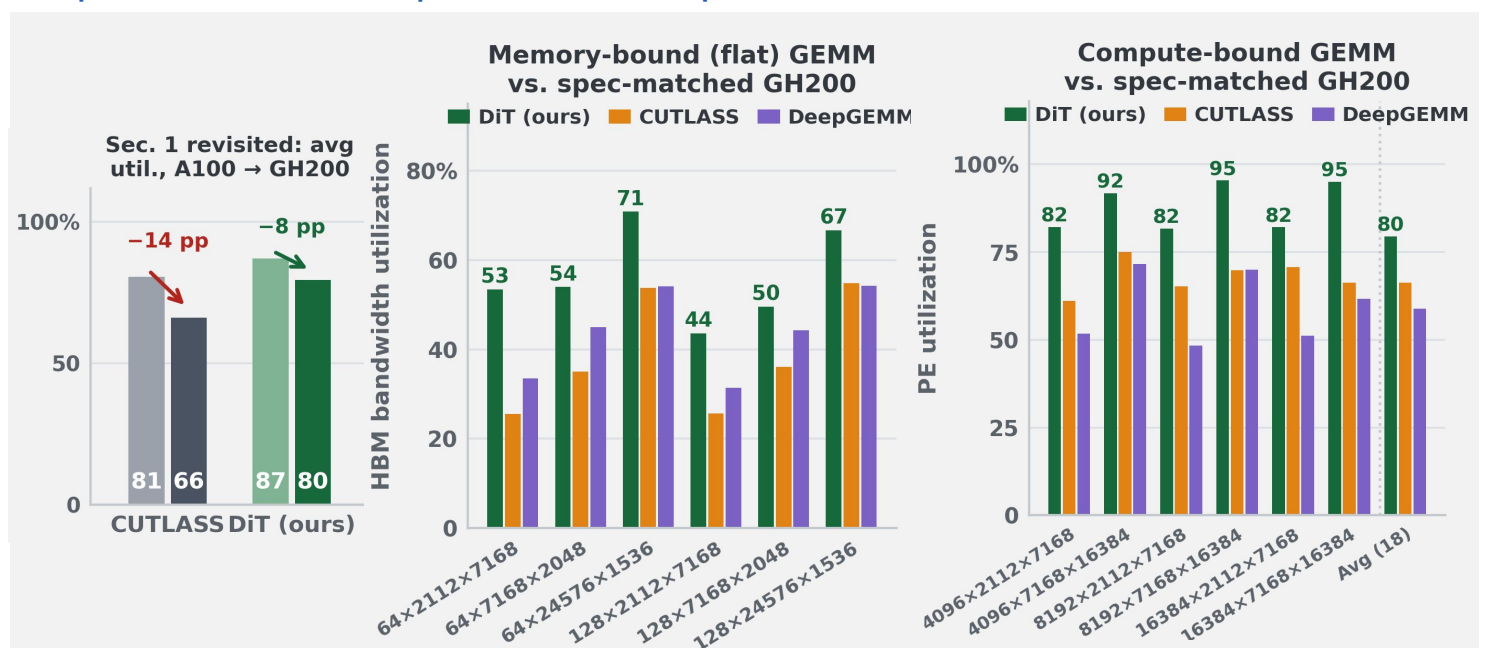


SUMMA multicast wins compute-bound shapes; systolic / nested variants suit store-intensive ones; Split-K + remap rescue irregular GEMM

5 Ablation vs. our baseline



6 Spec-matched comparison with expert-tuned GPU libraries



1.2 – 2.0× speedup over expert-tuned CUTLASS & DeepGEMM on a spec-matched GH200

References

- C. Zhang et al., "FlatAttention: Dataflow and fabric collectives co-optimization for efficient multi-head attention on tile-based many-PE accelerators," IEEE ISVLSI, 2025.
- T. Ben-Nun et al., "Stateful dataflow multigraphs," Proc. Int'l Conf. for High Performance Computing, Networking, Storage and Analysis (SC), 2019. — Full paper: arXiv:2512.13638 · Contact: Aofeng Shen, aoshen@iis.ee.ethz.ch