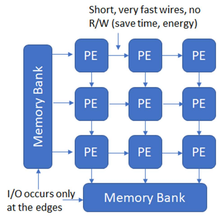
Systolic Array Matrix Multiplication Accelerator (SAMMA) ASIC Tape-out and Comparison

PIPETTE TEAM: VANCE BORUS, SACHAL SHAIKH, BYEONGGUK LEE, CHENYI WANG, NEAL CAUSEY, AND SEAN BUBERNAK
VANILLA TEAM: BRANDON CHANG, VERONICA GONG, YARA AL-SHORMAN, AND TEJASWINI KOPPALKAR

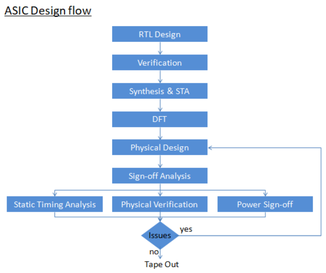
Systolic Arrays

- Modern AI accelerators rely on systolic arrays, which consist of a network of processing elements (PEs), for efficient matrix multiplication.
- The interconnect topology between PEs affects communication efficiency and hardware area.
- We compare power, performance, and area (PPA) between a traditional 2D mesh network (Vanilla) and a Twisted-Torus interconnect network (Pipette).



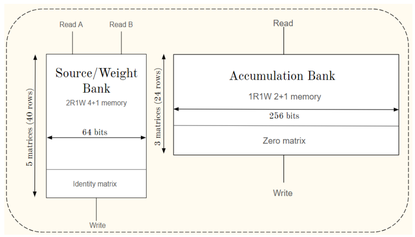
ASIC Design Flow

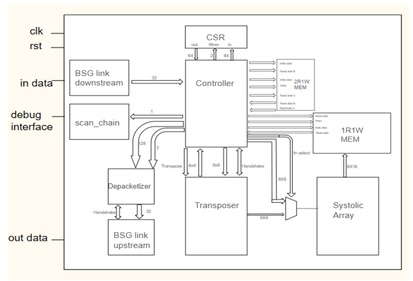
- Hammer, a custom EDA flow written in Python, is used to design the two chips.
- HAMMER integrates several industry-standard EDA tools such as Genus, Innovus, Calibre, Klayout, Conformal, and Verdi for a full ASIC design flow.
- The flow enables iterative design, testing, and optimization across all stages of ASIC implementation.



Shared Component Layout

- To enable direct architectural comparison, both chips share several modules, a common instruction set architecture (ISA) and memory organization structure.
- This sharing also includes I/O and chip controller interfaces, showing how the compute unit is the only major change to compare against.
- Memory is partitioned into two banks: one for INT8 and one for INT32 for their specialized purposes.





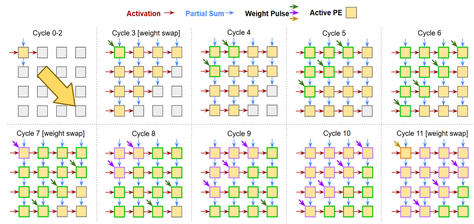
PPA Comparison

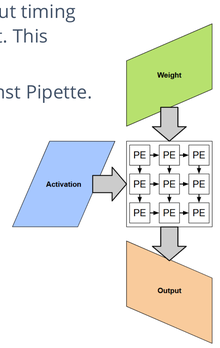
PPA stands for Power, Performance, and Area, which are important metrics in ASIC design.

- The Pipette compute unit achieves approximately **10% lower systolic array area**.
- Pipette requires only 56 mW, which is about **15% less power than Vanilla.**
- Pipette shows modest latency reductions of **around 8 cycles** for a single computation.

Vanilla Compute Unit

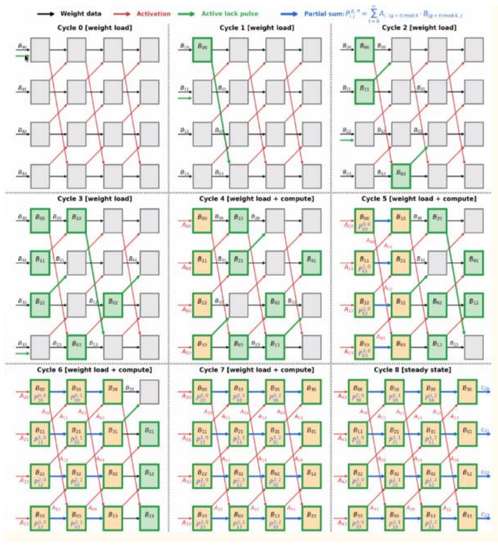
- The weight-stationary 2D mesh networks necessitates a skewed input timing for timing calculations, which in turn leads to de-skewing the output. This requires the use of shift register delays for data re-alignment.
- This design serves as the baseline architecture for comparison against Pipette.





Pipette Compute Unit

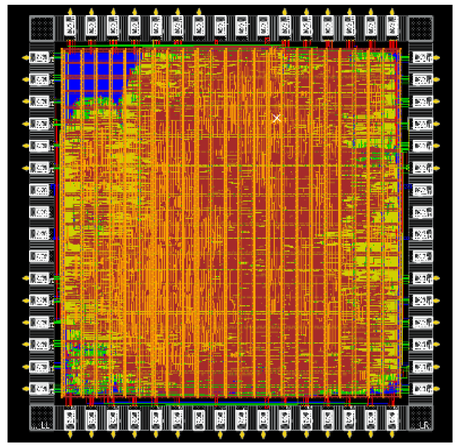
- Pipette's replacement of the 2D mesh with a Twisted-Torus interconnect eliminates Vanilla's boundary skew shift registers.
- Calculations travel along a diagonal torus-like link, removing the need for input skewing and output de-skewing.
- Preliminary values are loaded along a bypass path during the current calculation to enable starting the next calculation as soon as possible.
- This shift register removal eliminates the staggered loading step, directly impacting Pipette's area, power, and latency gains over Vanilla.

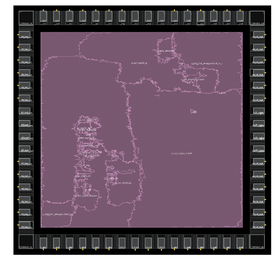


Cadence Innovus Chip Layout

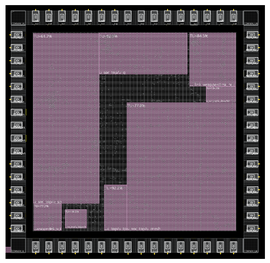
To the right is the raw logical layout of the chip, showing the memory, compute unit (Pipette's in this case), and control modules:

- 64 I/O pads in a ring around the chip's edge, with 16 used for PWD/GND.
  - The prominent horizontal and vertical lines on the chip are for power and ground distribution.
- Each color represents a different metal layer from METAL1 up to METAL6, which serve as the interconnections between logic gates on the bottom layer.





Amoeba View
- Shows each lower-level module.
- Reveals the boundaries each module's gates take up.



Floorplan View
- Created Guides for each module.
- Helps tool place our modules in the defined locations with a set density.

References

- This lab was advised by Ang Li and Jiayi Wang, and their research lab PNCel, which is the source of the Pipette architecture.
- Vanilla sourced their design from industry papers such as Google's TPU7x Ironwood, for comparison with Pipette.

J. Li and Z. Jiang, “Performance analysis of cambricon MLU100,” SpringerLink.

“TPU7x (Ironwood),” Google Cloud Documentation, 2026. https://docs.cloud.google.com/tpu/docs/tpu7x.

Xu, Rui, et al. "A survey of design and optimization for systolic array-based DNN accelerators." ACM Computing Surveys 56.1 (2023): 1-37.

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISERS: ANG LI, JIAYI WANG
SPONSORS: TSMC and APPLE