GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis

Damla Senol Cali

Carnegie Mellon University

(dsenol@andrew.cmu.edu)

Gurpreet S. Kalsi[2], Zulal Bingol[3], Can Firtina[4], Lavanya Subramanian[5], Jeremie Kim[1,4], Rachata Ausavarungnirun[6,1], Mohammed Alser[4], Juan Gomez-Luna[4], Amirali Boroumand[1], Anant Nori[2], Allison Scibisz[1], Sreenivas Subramoney[2], Can Alkan[3], Saugata Ghose[7,1], and Onur Mutlu[4,1,3]

[1] Carnegie Mellon [2] intel [3] Bilkent University [4] ETH zürich

[5] facebook [6] King Mongkut's University of Technology North Bangkok [7] University of Illinois Urbana-Champaign [1,4] SAFARI

Genome Sequencing

- **Genome sequencing:** Enables us to determine the order of the DNA sequence in an organism's genome
  - Plays a pivotal role in:
    - Personalized medicine
    - Outbreak tracing
    - Understanding of evolution

Genome

DNA

- Modern genome sequencing machines extract smaller randomized fragments of the original DNA sequence, known as **reads**
  - *Short reads:* a few hundred base pairs, error rate of ~0.1%
  - *Long reads:* thousands to millions of base pairs, error rate of 10–15%

Genome Sequence Analysis

- **Read mapping:** *First key step* in genome sequence analysis (GSA)
  - Aligns reads to one or more possible locations within the reference genome, and
  - Finds the matches and differences between the read and the reference genome segment at that location

- Multiple steps of read mapping require ***approximate string matching***
  - Approximate string matching (ASM) enables read mapping to account for sequencing errors and genetic variations in the reads

- Bottlenecked by the computational power and memory bandwidth limitations of existing systems

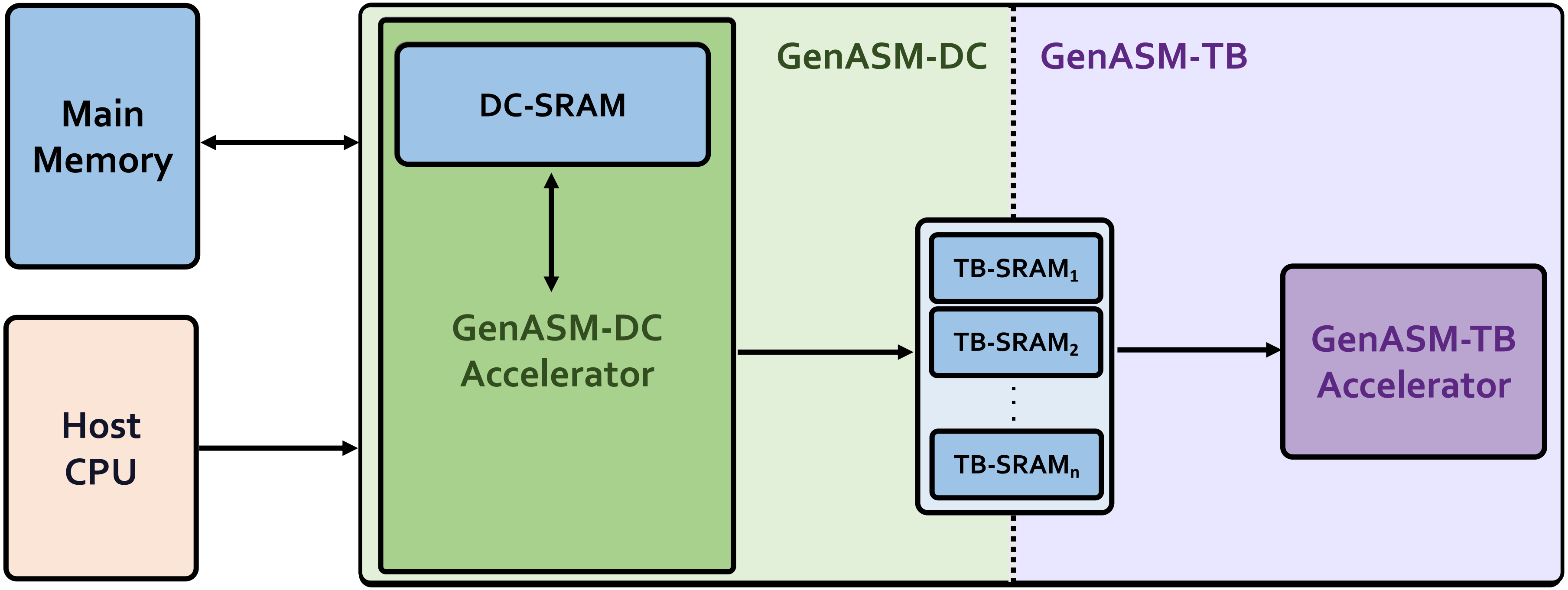
GenASM: ASM Framework for GSA

Our Goal:

Accelerate approximate string matching
by designing a fast and flexible framework,
which can accelerate *multiple steps* of genome sequence analysis

- **GenASM:** *First* ASM acceleration framework for GSA
  - Based upon the *Bitap* algorithm
    - Uses fast and simple bitwise operations to perform ASM
  - Modified and extended ASM algorithm
    - Highly-parallel Bitap with long read support
    - Bitvector-based novel algorithm to perform *traceback*
  - Co-design of our modified scalable and memory-efficient algorithms with low-power and area-efficient hardware accelerators

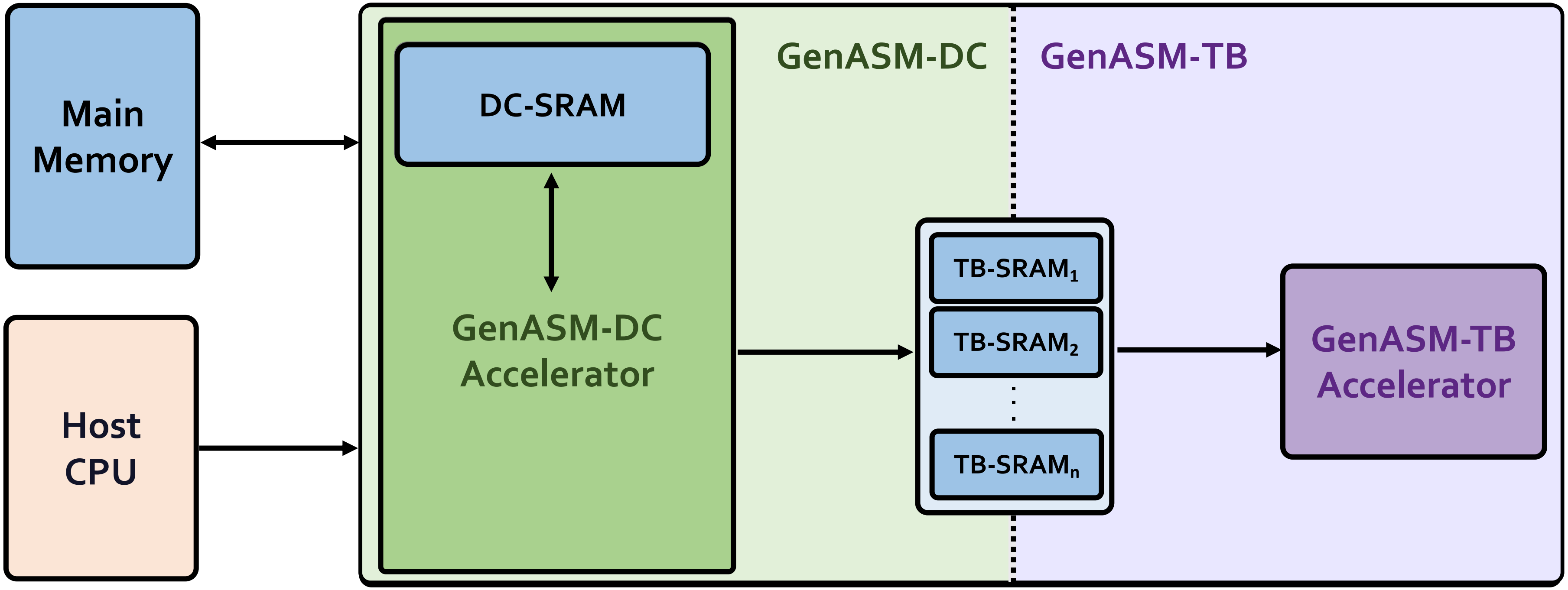
GenASM: Hardware Design



GenASM-DC:
generates bitvectors
and performs edit
Distance Calculation

GenASM-TB:
performs TraceBack
and assembles the
optimal alignment

GenASM: Hardware Design



Our *specialized compute units* and *on-chip SRAMs* help us to:

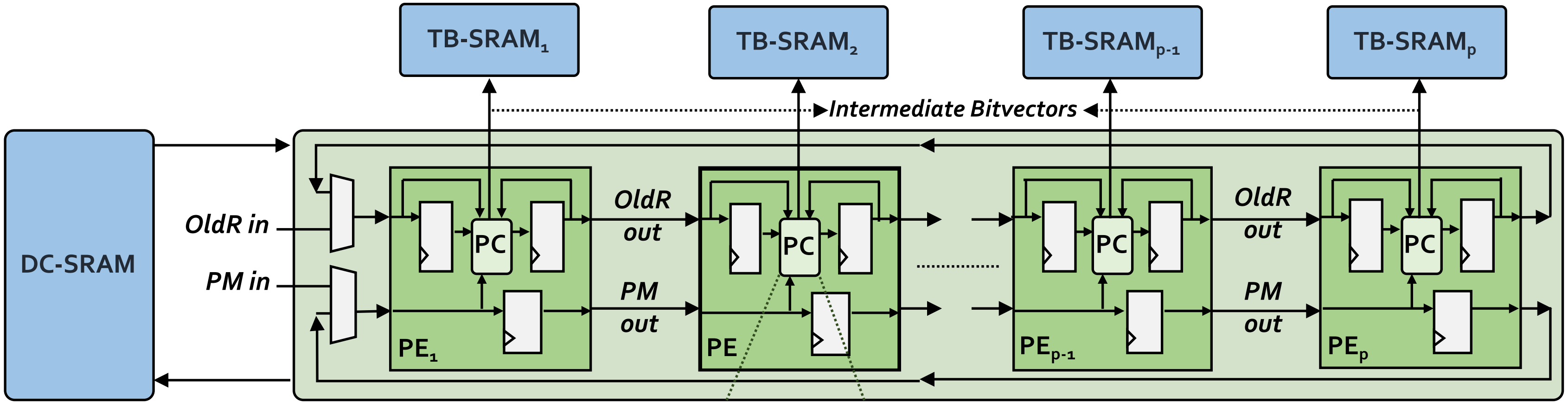
→ Match the rate of computation with memory capacity and bandwidth

→ Achieve high performance and power efficiency

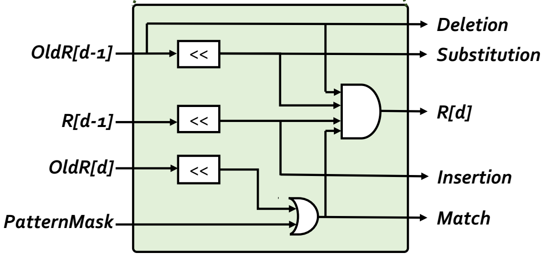
→ Scale linearly in performance with the number of parallel compute units that we add to the system

GenASM-DC: Hardware Design

- **Linear cyclic systolic array** based accelerator
  - Designed to maximize parallelism and minimize memory bandwidth and memory footprint

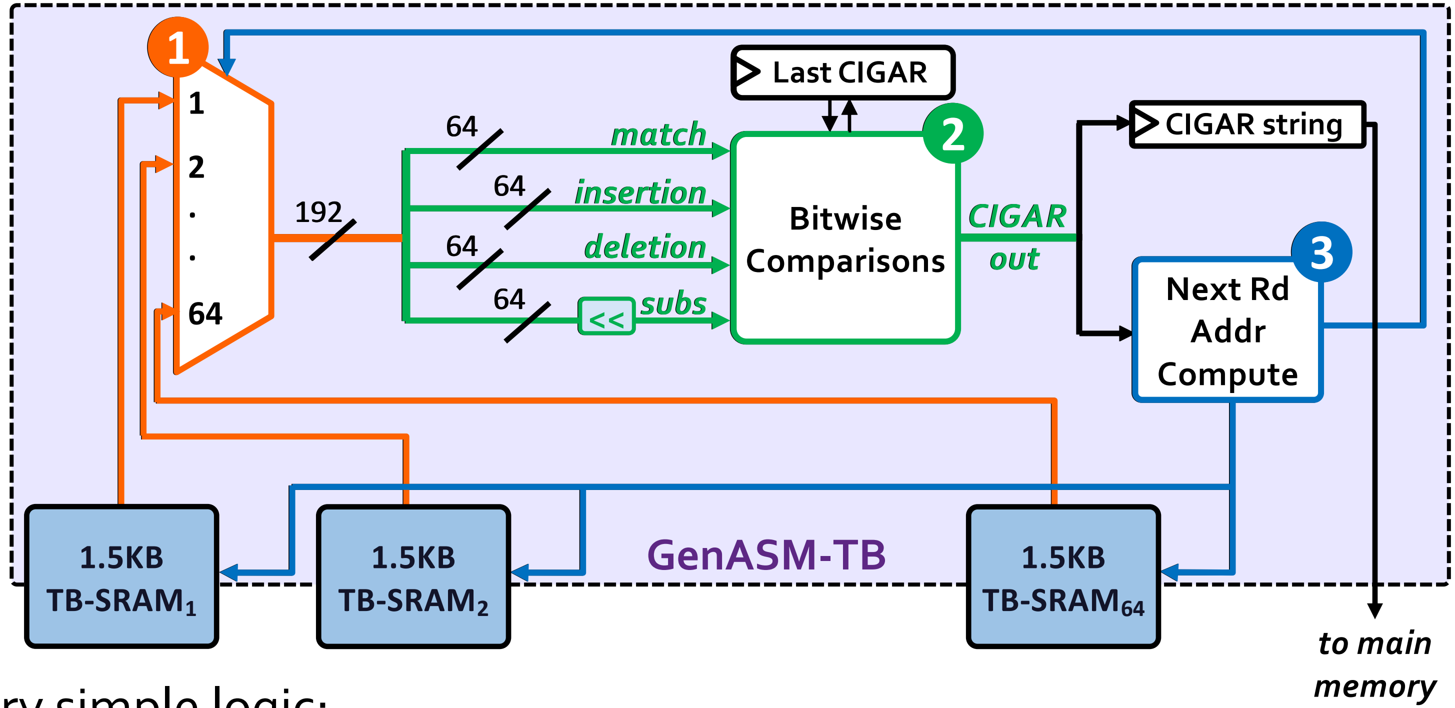


Processing Block (PB)



Processing Core (PC)

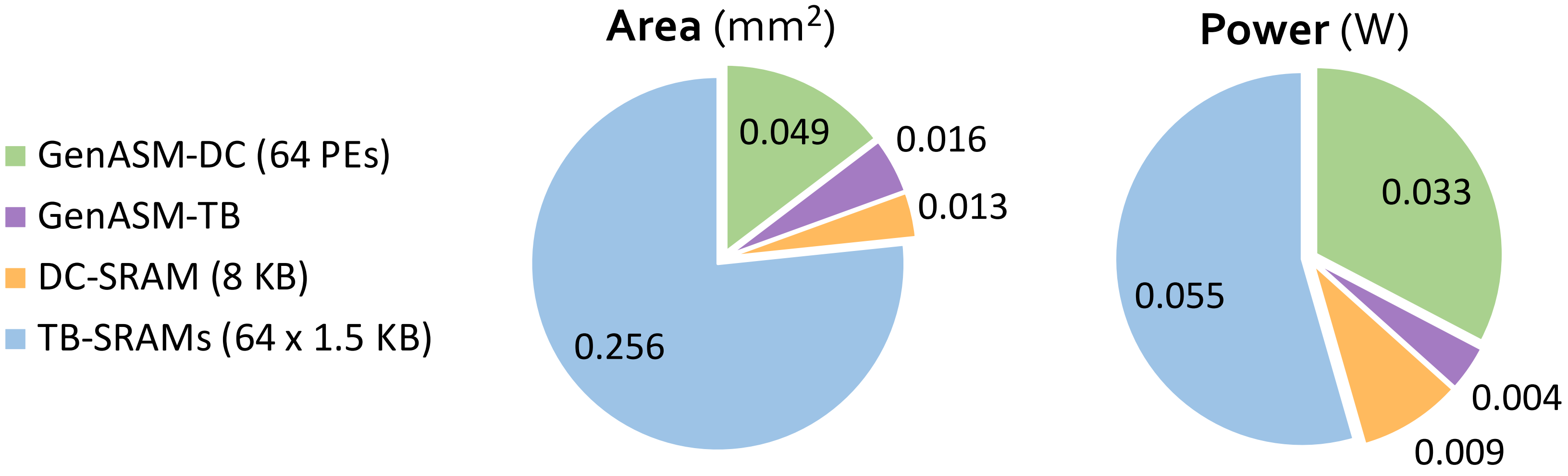
GenASM-TB: Hardware Design



- ❑ Very simple logic:
  - 1 Reads the bitvectors from one of the TB-SRAMs using the computed address
  - 2 Performs the required bitwise comparisons to find the traceback output for the current position
  - 3 Computes the next TB-SRAM address to read the new set of bitvectors

Key Results – Area and Power

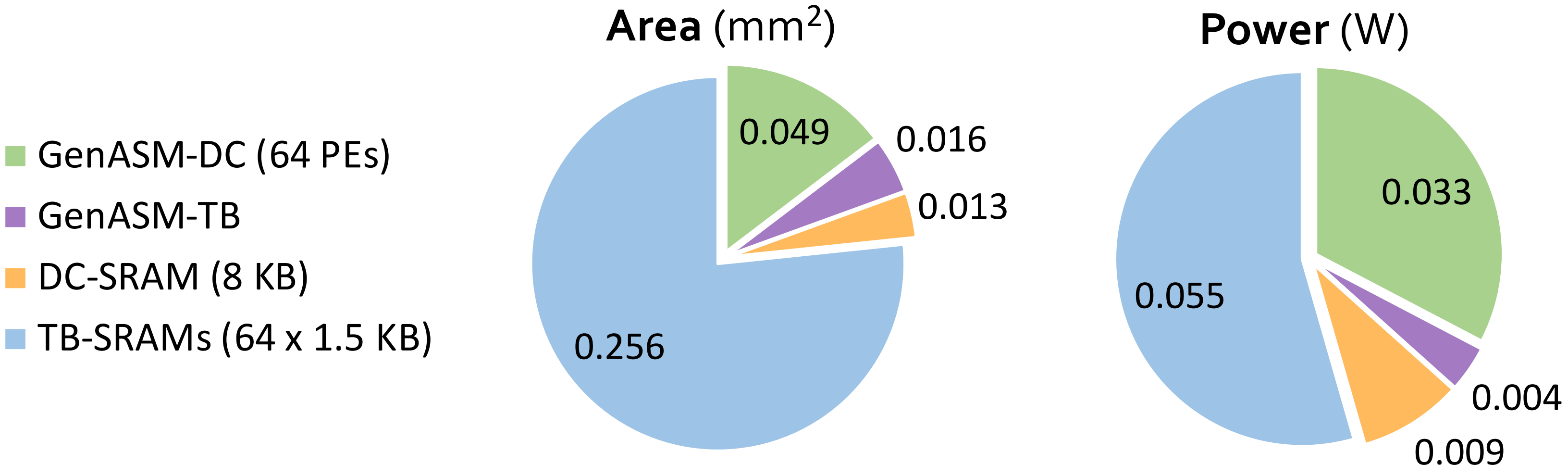
- Based on our **synthesis** of **GenASM-DC** and **GenASM-TB** accelerator datapaths using the Synopsys Design Compiler with a **28nm** LP process:
  - Both GenASM-DC and GenASM-TB operate **@ 1GHz**



Total (1 vault):	0.334 mm²	0.101 W
Total (32 vaults):	10.69 mm²	3.23 W
% of a Xeon CPU core:	**1%**	**1%**

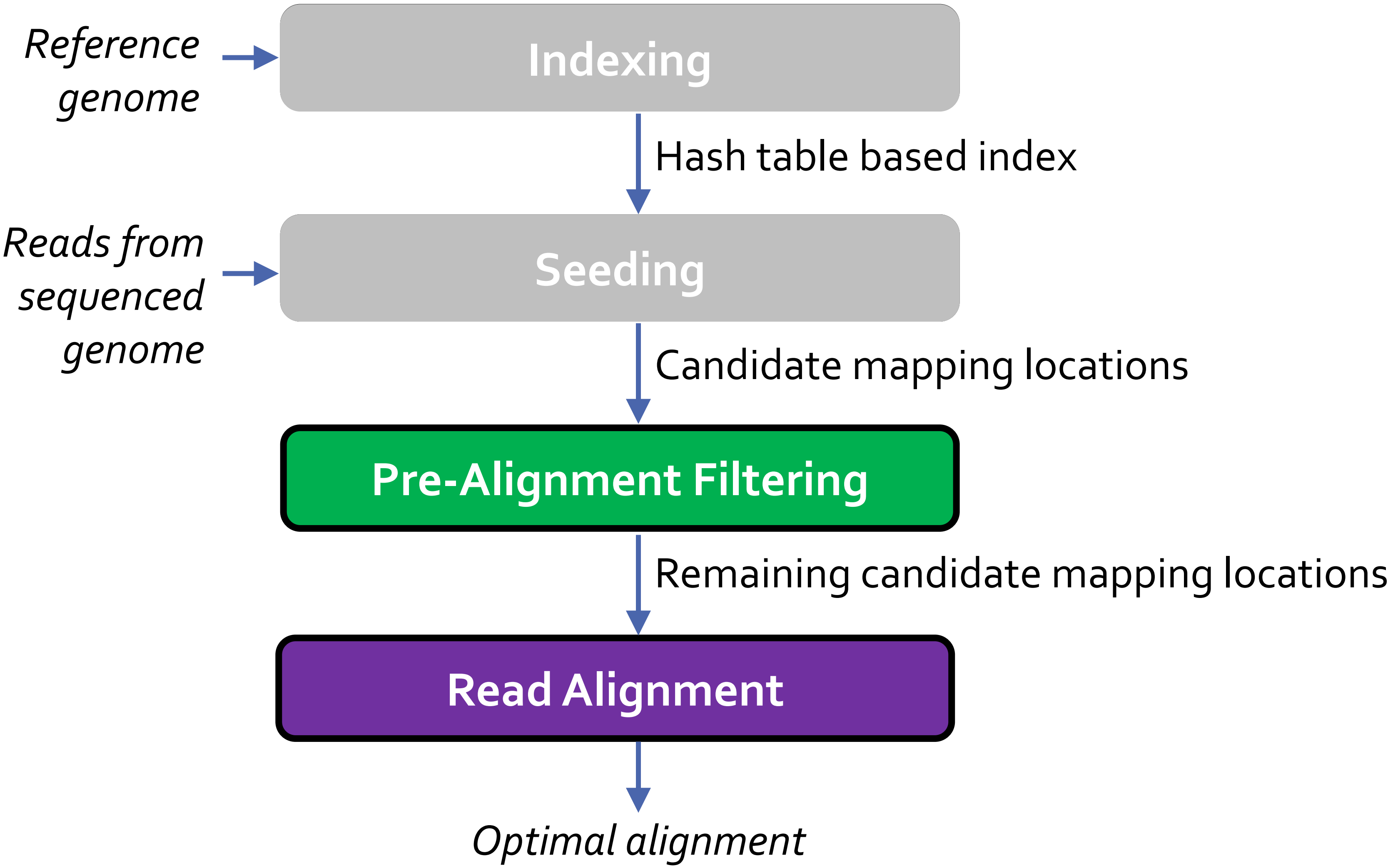
Key Results – Area and Power

- Based on our **synthesis** of **GenASM-DC** and **GenASM-TB** accelerator datapaths using the Synopsys Design Compiler with a **28nm** LP process:
  - Both GenASM-DC and GenASM-TB operate **@ 1GHz**



GenASM has low area and power overheads

Use Cases of GenASM



Use Cases of GenASM (cont'd.)

(1) Read Alignment Step of Read Mapping

- ○ Find the optimal alignment of how reads map to candidate reference regions

(2) Pre-Alignment Filtering for Short Reads

- ○ Quickly identify and filter out the unlikely candidate reference regions for each read

(3) Edit Distance Calculation

- ○ Measure the similarity or distance between two sequences

❑ We also discuss other possible use cases of GenASM in our paper:

- ○ Read-to-read overlap finding, hash-table based indexing, whole genome alignment, generic text search

Key Results

(1) Read Alignment

- ❑ **116×** speedup, **37×** less power than **Minimap2** (state-of-the-art **SW**)
- ❑ **111×** speedup, **33×** less power than **BWA-MEM** (state-of-the-art **SW**)
- ❑ **3.9×** better throughput, **2.7×** less power than **Darwin** (state-of-the-art **HW**)
- ❑ **1.9×** better throughput, **82%** less logic power than **GenAx** (state-of-the-art **HW**)

(2) Pre-Alignment Filtering

- ❑ **3.7×** speedup, **1.7×** less power than **Shouji** (state-of-the-art **HW**)

(3) Edit Distance Calculation

- ❑ **22–12501×** speedup, **548–582×** less power than **Edlib** (state-of-the-art **SW**)
- ❑ **9.3–400×** speedup, **67×** less power than **ASAP** (state-of-the-art **HW**)

GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis

Damla Senol Cali

Carnegie Mellon University

(dsenol@andrew.cmu.edu)

Gurpreet S. Kalsi[2], Zulal Bingol[3], Can Firtina[4], Lavanya Subramanian[5],
Jeremie Kim[1,4], Rachata Ausavarungnirun[6,1], Mohammed Alser[4],
Juan Gomez-Luna[4], Amirali Boroumand[1], Anant Nori[2], Allison Scibisz[1],
Sreenivas Subramoney[2], Can Alkan[3], Saugata Ghose[7,1], and Onur Mutlu[4,1,3]

[1] Carnegie Mellon [2] intel [3] İhsan Doğramacı Bilkent University 1984 Bilkent University [4] ETH zürich

[5] facebook [6] KING MONGKUT'S UNIVERSITY OF TECHNOLOGY NORTH BANGKOK [7] UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN [1,4] SAFARI