SOAP3 & SOAP3-dp

GPU-based Compressed Indexing & Ultra-fast Parallel Alignment of Short Reads

- A collaboration between University of Hong Kong (HKU) & BGI

T.W. Lam, C.M. Liu, R. Luo, Thomas Wong, Edward Wu, S.M. Yiu, HKU
Yingrui Li, Bingqiang Wang, Chang Yu, BGI
X. Chu, K. Zhao, Baptist U
Ruiqianq Li, Peking U

Short read alignment

- First step of NGS (next generation sequencing) data analysis: Mapping a large number of short reads to a reference genome with a few mismatches allowed.
  - E.g., reference : human genome (~3 Gigabases);

    NGS output: 1.2 billion reads, each of length 100;

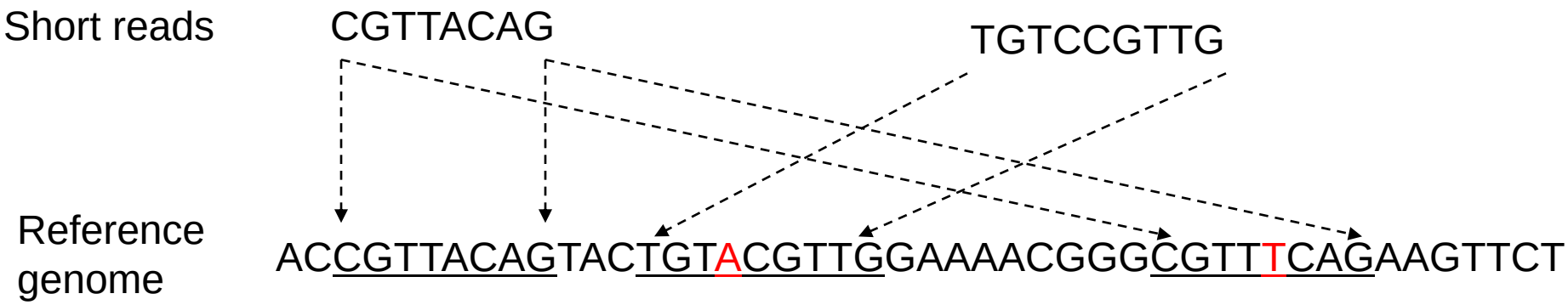
    2 to 4 mismatches.

Short read alignment

- First step of NGS (next generation sequencing) data analysis: Mapping a large number of short reads to a reference genome with a few mismatches allowed.
  - E.g., reference : human genome (~3 Gigabases);

    NGS output: reads, each of length 100;

    2 to 4 mismatches.



Data volume

- A high-throughput sequencer like Illumina HiSeq 2500 can generate 1.2G reads of length 100 in 27 hours (total size 120 Gigabases)
- Large genome centers like BGI have over 100 sequencers.
- The alignment software must be really fast.

Existing tools since 2008

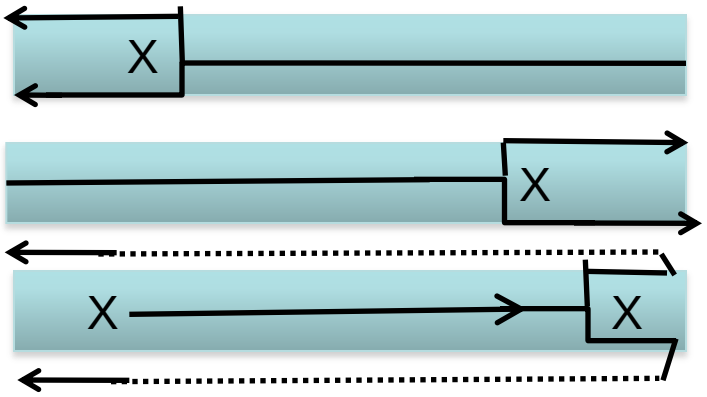
- Maq, SOAP2, ZOOM, Bowtie, BWA, …
- SOAP2 and BWA are known to be the fastest

SOAP → SOAP2 → SOAP3 → SOAP3-dp

- SOAP: first-generation short read alignment software
- SOAP2 (2008): 20 to 30 times faster than SOAP, less memory
  - first collaboration between HKU & BGI
  - Compressed indexing: bidirectional BWT (2BWT)
  - E.g., read 100 bp, 4 mismatches, best alignment :
    - 140 – 220 seconds per million reads (quad core)
- SOAP3 (2011): 10 to 30 times faster than SOAP2
  - GPU's parallel processing power; CPU memory: increase from a few to tens GB.
  - GPU-based indexing: GPU-2BWT
  - E.g., read 100 bp, 4 mismatches, best alignment :
    - 5 - 6 seconds (quad core + GPU); improved sensitivity
- SOAP3-dp (2012): 2 to 3 times faster than SOAP3; higher sensitivity
  - Consider alignment with INDELs (insert/delete) in addition to mismatches.
  - GPU: index-assisted dynamic programming

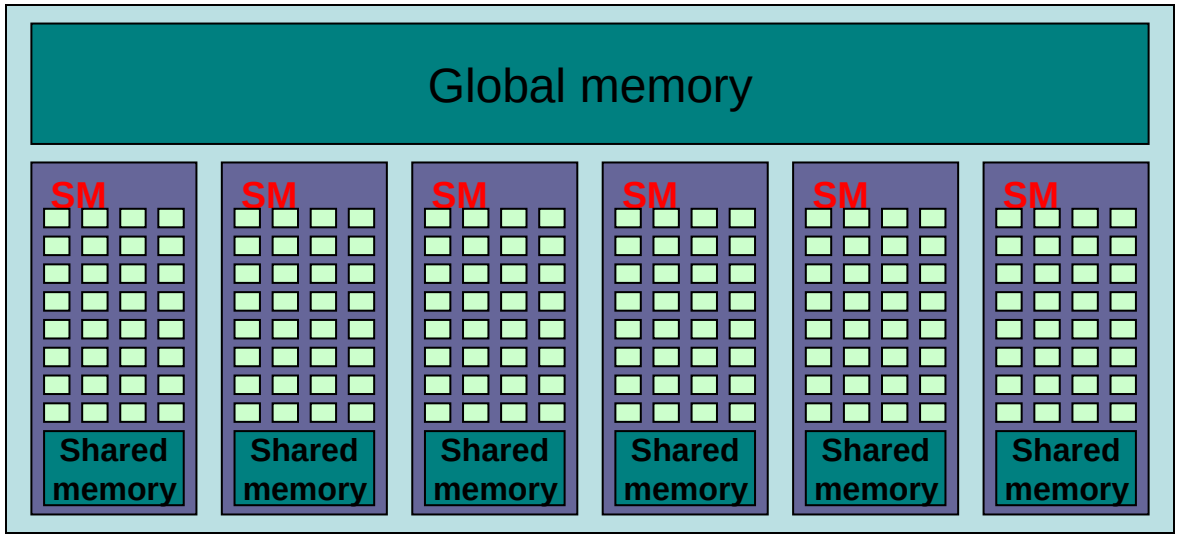
SOAP ➔ SOAP2 ➔ SOAP3

- Fast alignment software makes use of a compressed index of the reference genome in the main memory.
- BWA, Bowtie: BWT (Burrows-Wheeler transform) allows very efficient pattern matching in one direction.
- SOAP2, SOAP3: 2BWT (bidirectional BWT), allows very efficient pattern matching in both directions.



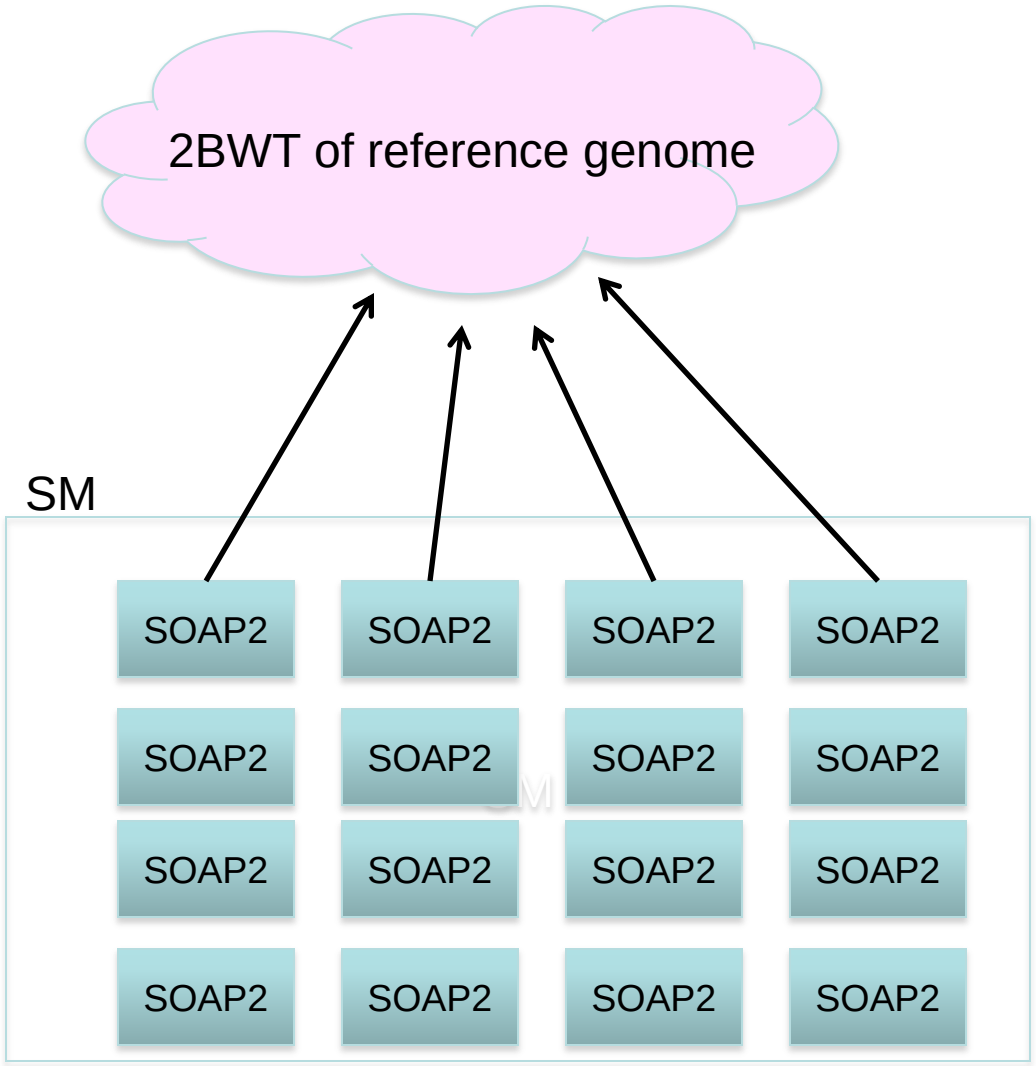
SOAP3 & GPU

- SOAP3 is a GPU-enhanced version of SOAP2.
- GPU: multiple streaming multiprocessers (SM)
  - Each SM contains dozens of processers (e.g. 32)
  - single-instruction multiple-thread (SIMT)



SOAP2 + GPU: A naïve approach

Each processor works on a different read.



The naïve approach sucks

- Initial attempt: SOAP2 on GPU is indeed slower than SOAP2 using CPU.
- GPU is tailor-made for running computational intensive process in parallel;
  - yet alignment with indexing is data intensive. I.e., the index is the bottleneck.
- Too much branching: GPU is ideal for running identical processes in parallel (SIMT); but
  - different reads have different alignments (numbers and positions of mismatches).
  - Very often, within an SM, many reads are waiting for a few "troublesome" reads to finish.

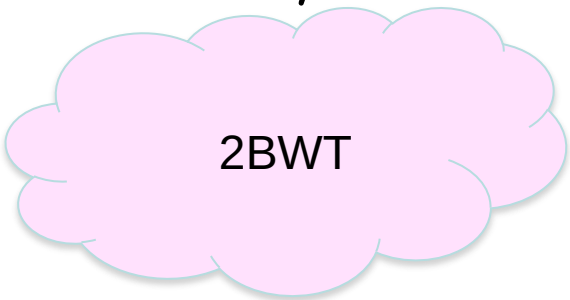
SOAP3 core ideas

- Reduce memory access
- Reduce branching effect

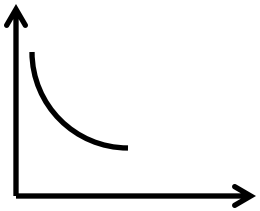
How ?

Reduce memory access

- Engineering the index: E.g,
  - 2-level sampling becomes 1-level sampling;
  - group data items according to retrieval patterns instead of logical functions
  - redundancy



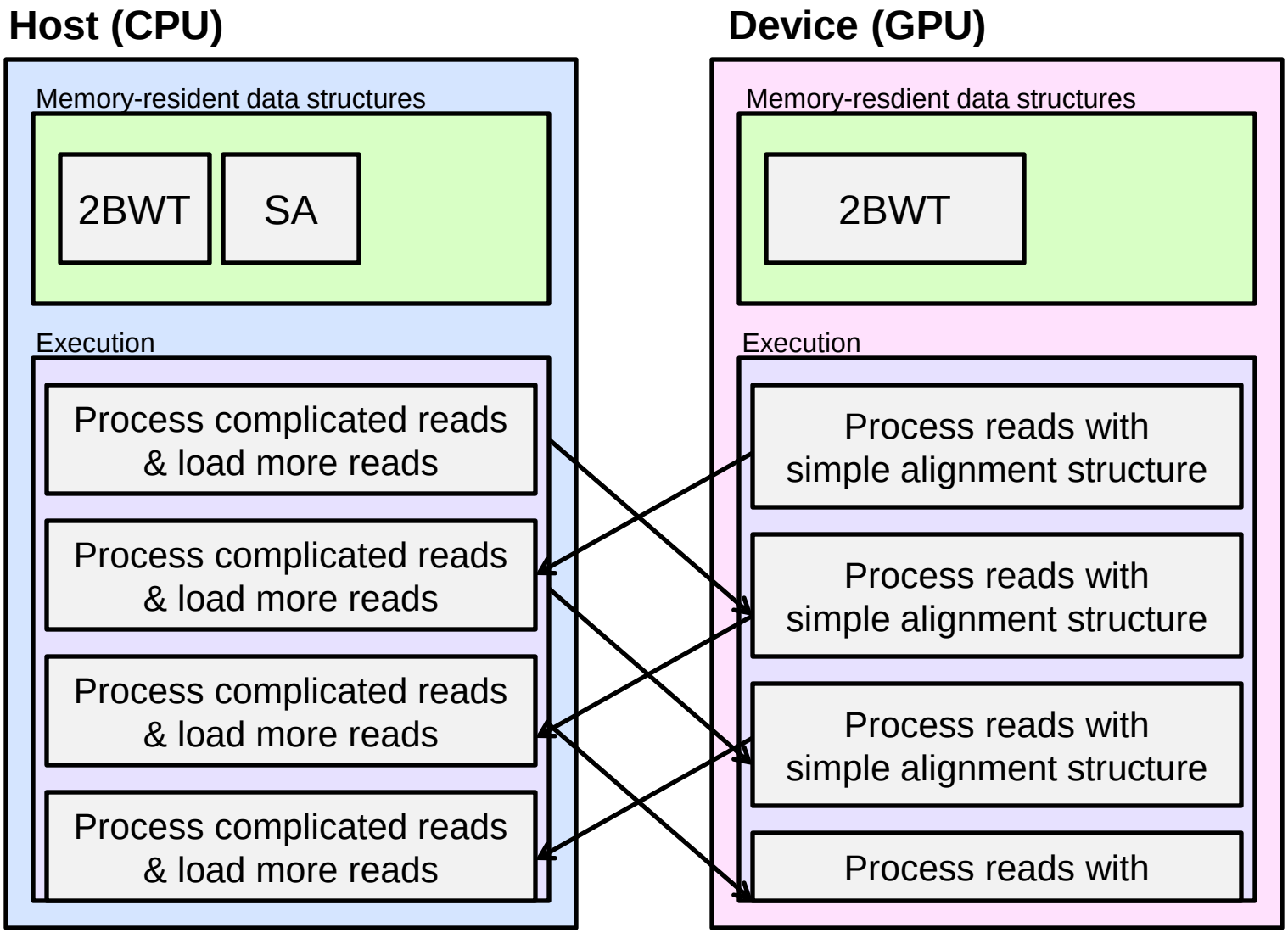
- In SOAP2, a search step takes four 32-bit & two 256-bit memory accesses;
- In SOAP3, it takes two 32-bit & two 128-bit memory accesses



How to control branching effect

- Simple idea: GPU only finishes those structurally simple reads, and stops those complicated and time-consuming ones from completion.
- Multiple rounds:
  - Round 1: only finish those really simples one;
  - Round 2: group those complicated reads together for another round;
  - Round 3: extremely complicated ones are left to the CPU.
- How to define the complexity?
  - number of SA ranges (groups of alignments answers).

SOAP3 Architecture



Experimental setup

- GPU: NVIDIA GTX 580 (3 GB RAM); US $400
- 2.8 GHz quad-core CPU, 24 GB RAM

Experimental results

Reference: human genome 37.1

Reads: YH1 Cell-line DNA, 70 M read-pairs (length 100 x 2)

Output: best alignment

	SOAP3	SOAP2	BWA	Bowtie
3 mismatches	17 minutes	306 minutes	176 minutes	486 minutes
% of reads aligned	79.4%	76.5%	79.2%	79.4%
4 mismatches	33 minutes	309 minutes	229 minutes	not supported
% of reads aligned	81.5%	77.1%	81.0%	

NB. Soap3 time include load index time 2 minutes.

SOAP3 alignment time

- Single-end alignment (4 mismatches)
  - Find all alignments: ~ 38 seconds per million reads
  - Find a best alignment: ~ 5 seconds per million reads
- Paired-end alignment (4 mismatches)
  - Find all alignments: ~ 75 seconds per million read-pairs
  - Find a best alignment: ~ 20 seconds per million read-pairs

SOAP3-dp

Faster & higher sensitivity

- GPU: index-assisted dynamic programming (semi-global alignment)
- Alignment with INDELs (insert/delete) & mismatches
- YH1 data set: 70 M read-pairs (length 100 x 2):
  - SOAP3 : 33 minutes; alignment sensitivity 81.5%
  - SOAP3-dp: 19 minutes; alignment sensitivity 95.0%
  - BWA (mismatches only): 309 minutes; 81.0%
  - BWA (DP): 355 minutes; 90.7%
  - Bowtie2 (DP): 215 minutes: 90.5%

SOAP3-dp version 1.4

SOAP3-dp version 1.3: ww.cs.hku.hk/2bwt-tools/soap3-dp

Version 1.4 will be available in late May.

Other SOAP3 link:

- BGI (http://soap.genomics.org.cn/soap3.html)
- NIH (http://biowulf.nih.gov/apps/soap3.html)

Reference:

Liu et al. SOAP3: Ultra-fast GPU-based parallel alignment tool for short reads, Bioinformatics 2012.