

A close-up photograph of a man with dark hair, a beard, and a blue hair tie. He is smiling and looking down at a tablet device he is holding. The background is blurred, showing what appears to be a modern office or laboratory setting.

arm

Genomics at Arm

**“Life expectancy
exceeds 100 years”**

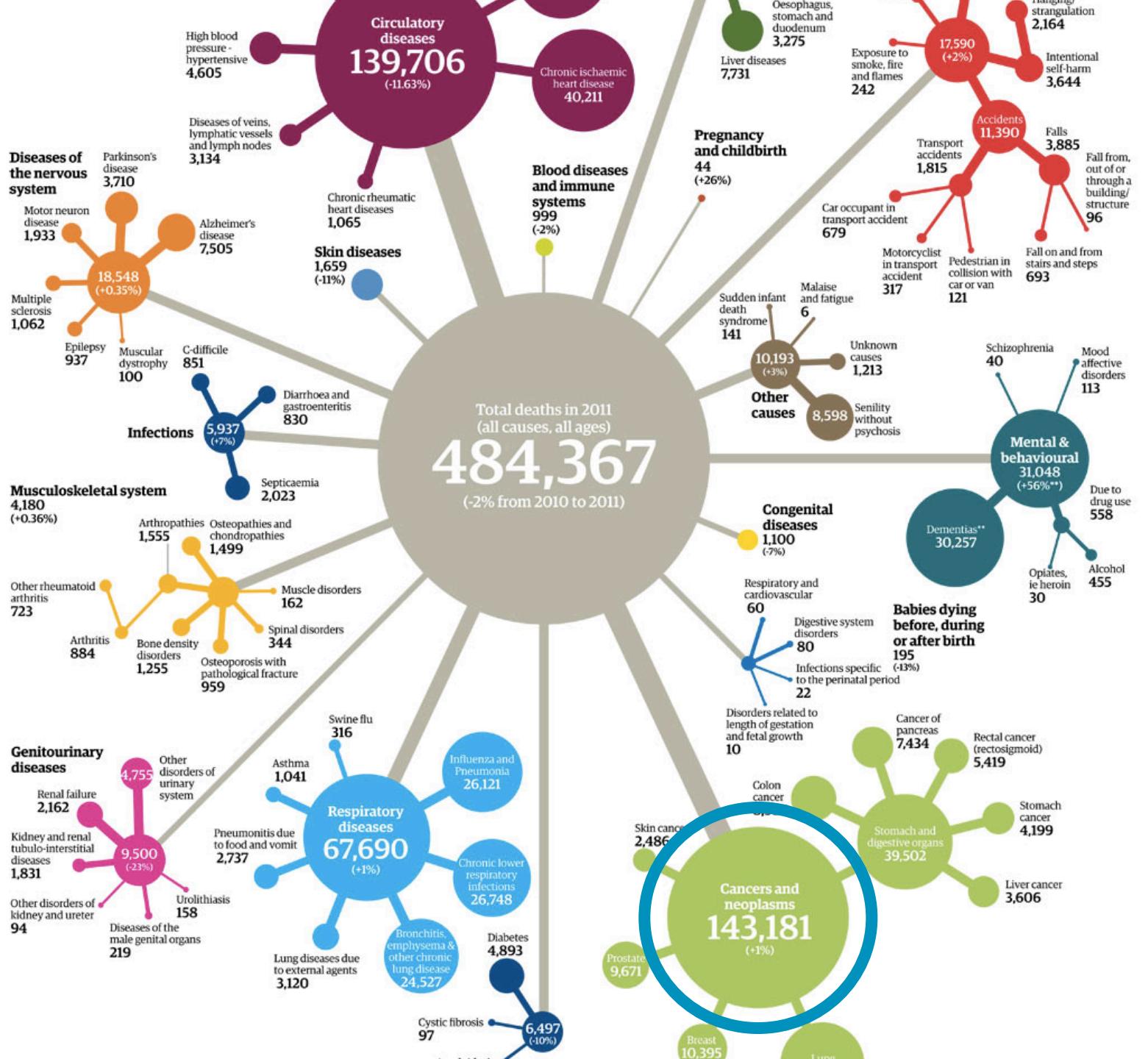
Masayoshi Son,
Chairman & CEO of SoftBank

Causes of Deaths

**“1 in 2 people
in the UK
will get cancer”**

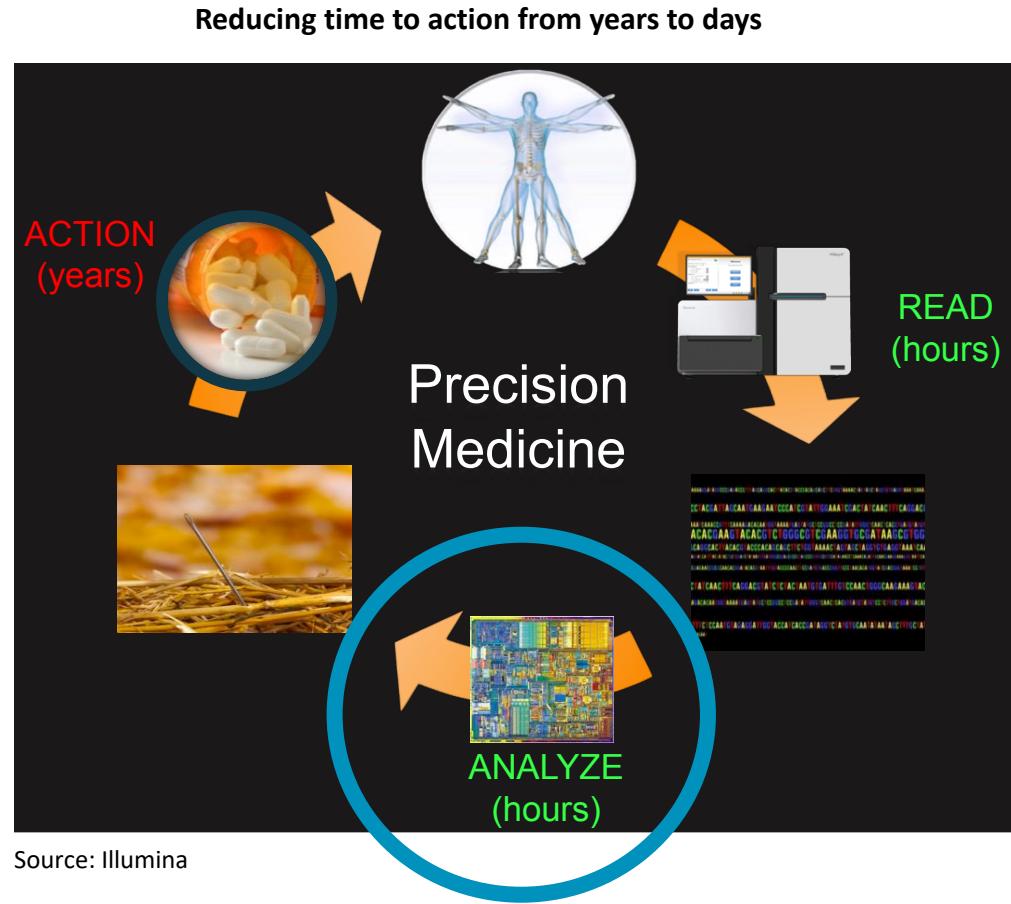
Cancer Research UK

arm

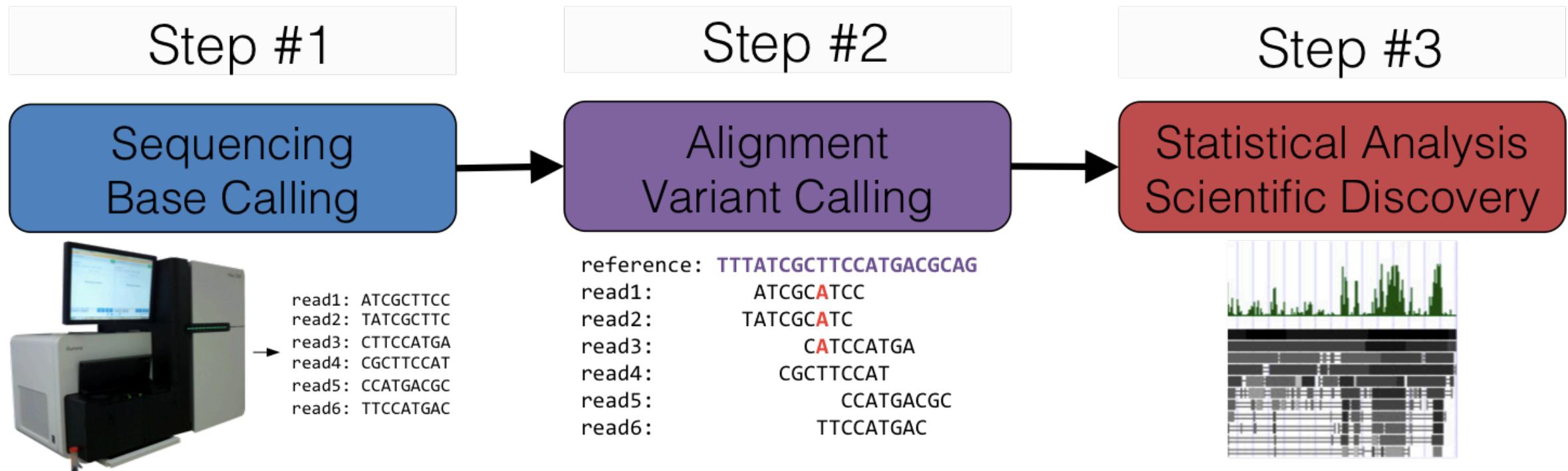


Genomics Accelerates Cancer Early Diagnosis and Treatments

- Cancers caused by genome mutations
- Genomics accelerates cancer early diagnosis and treatments
 - Genome sequencing for early cancer diagnosis, i.e., ctDNA
 - Genome editing for cancer treatments, i.e., CRISPR/cas9
 - Genomics enables precision medicine and reduces drug discovery cycles



Genomic Data Analysis Pipeline



Primary: Image/signal processing for base calling, i.e., canny edge detection, microfluidics control

Secondary: BWA alignment (incl. Smith-Waterman) and variant calling GATK (incl. PairHMM)

Tertiary: Machine learning algorithms: k-means, SVM, Fast-LMM etc.

Arm Accelerates Genomics

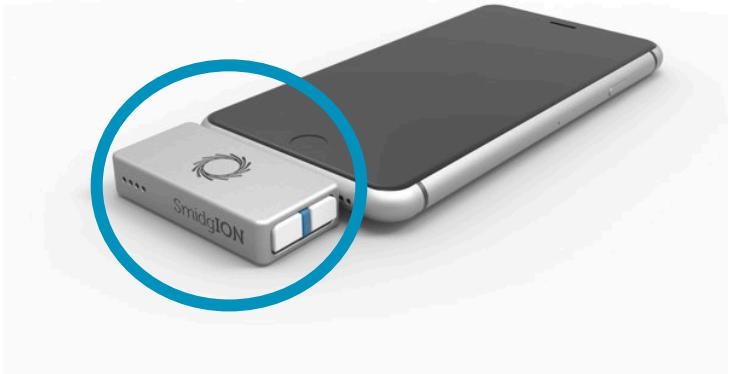
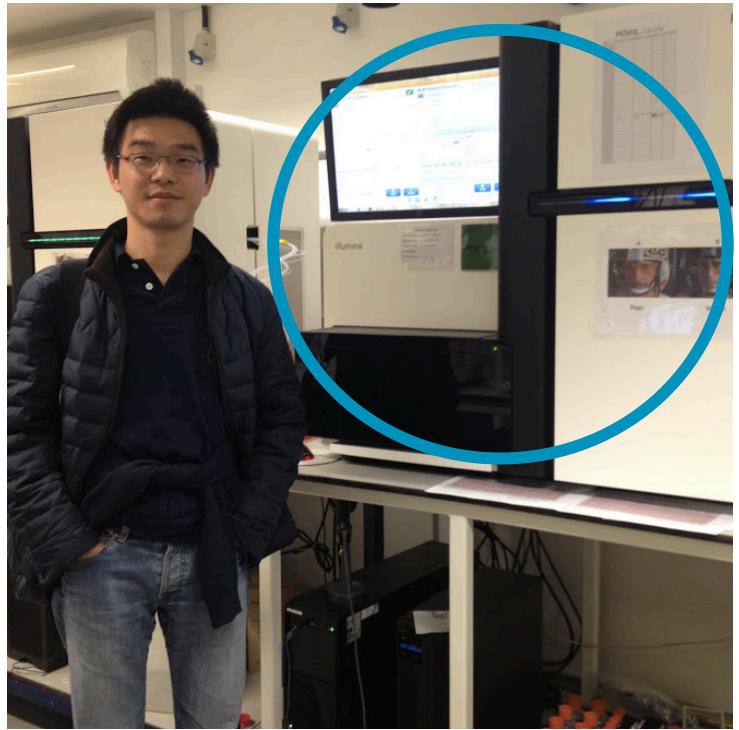
Arm servers for NGS

- Arm servers for secondary and tertiary genomic data analysis
- Arm SoC for primary data analysis inside NGS machines

Arm for mobile 3GS

- MinION for Ebola and Zika surveillance*
- PoC data processing required for primary data analysis
- On-board secondary data analysis desirable
- Machine learning in edge devices

* Reference: Quick et al, Real-time, portable genome sequencing for Ebola surveillance



Top: Illumina HiSeq
Bottom: Nanopore SmidgION

Current Work in Genomics

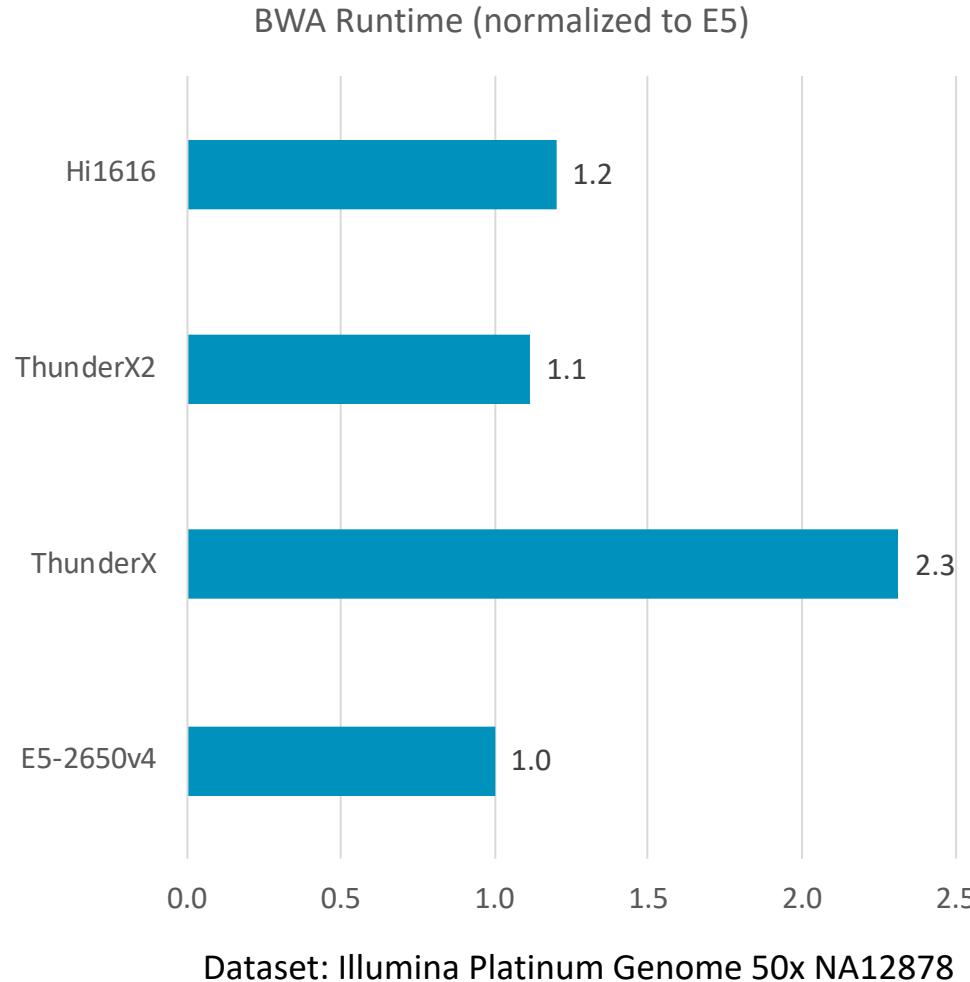
Make genomics workloads run efficiently on Arm processors

- Arm HPC tools
 - Compiler
 - Profiler
- Arm genomics library
 - SIMD, i.e., NEON and SVE
 - Accelerators, i.e., zip offload engine

Make future Arm processors better at running genomics workloads

- Create precision medicine benchmarks suite
- Propose new instructions for Arm SVE
- Processing in memory

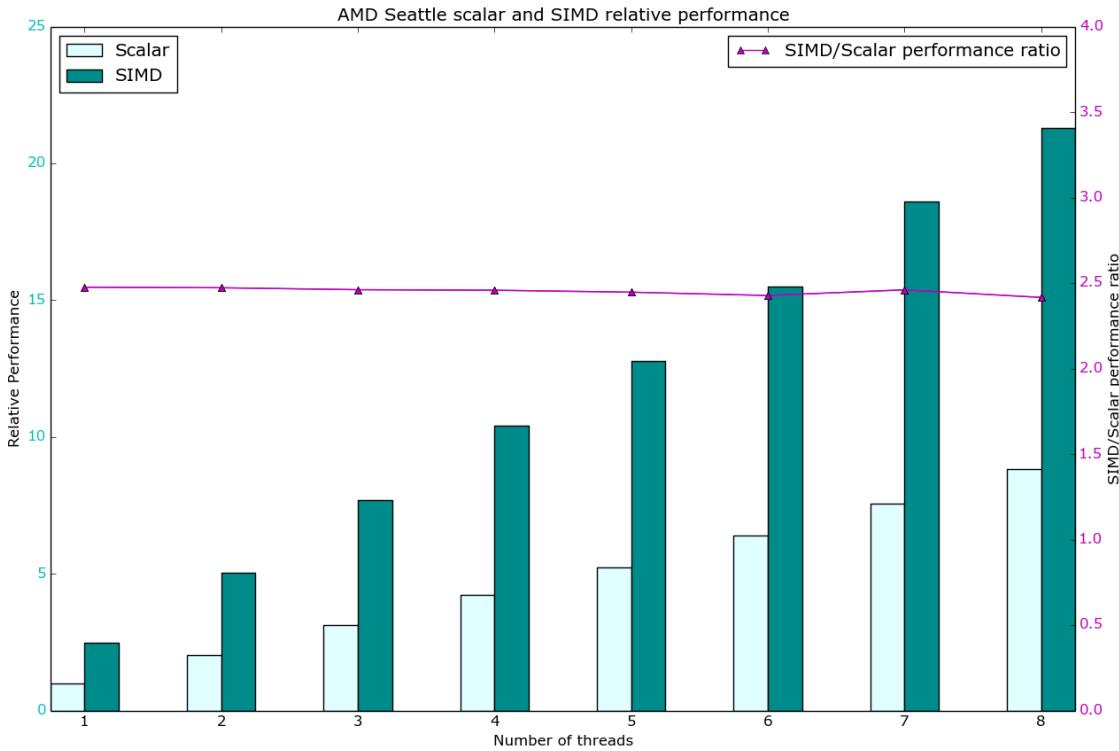
Burrows-Wheeler Aligner on Arm



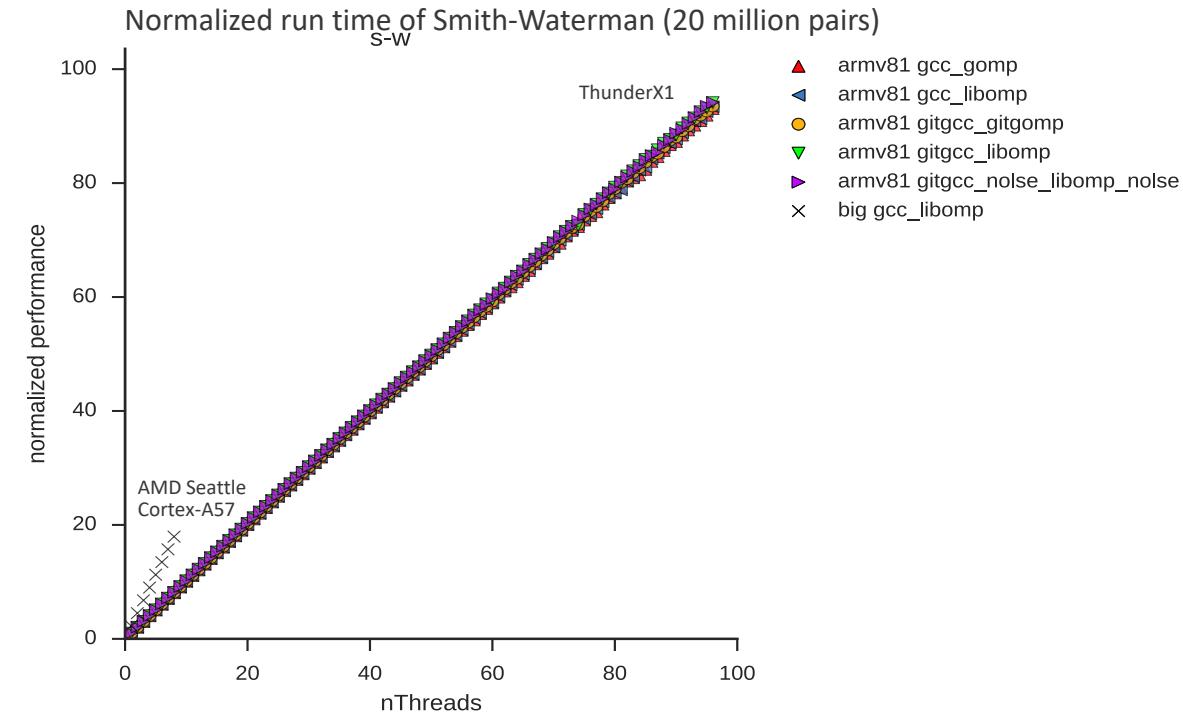
- Genome alignment packages run on Arm platforms
 - BWA and Isaac
- Performance competitive with alternative solutions
 - ThunderX2 and Hi1616 early prototypes
 - NEON not hand optimized as SSE2 in BWA

Speed up Smith-Waterman on Arm

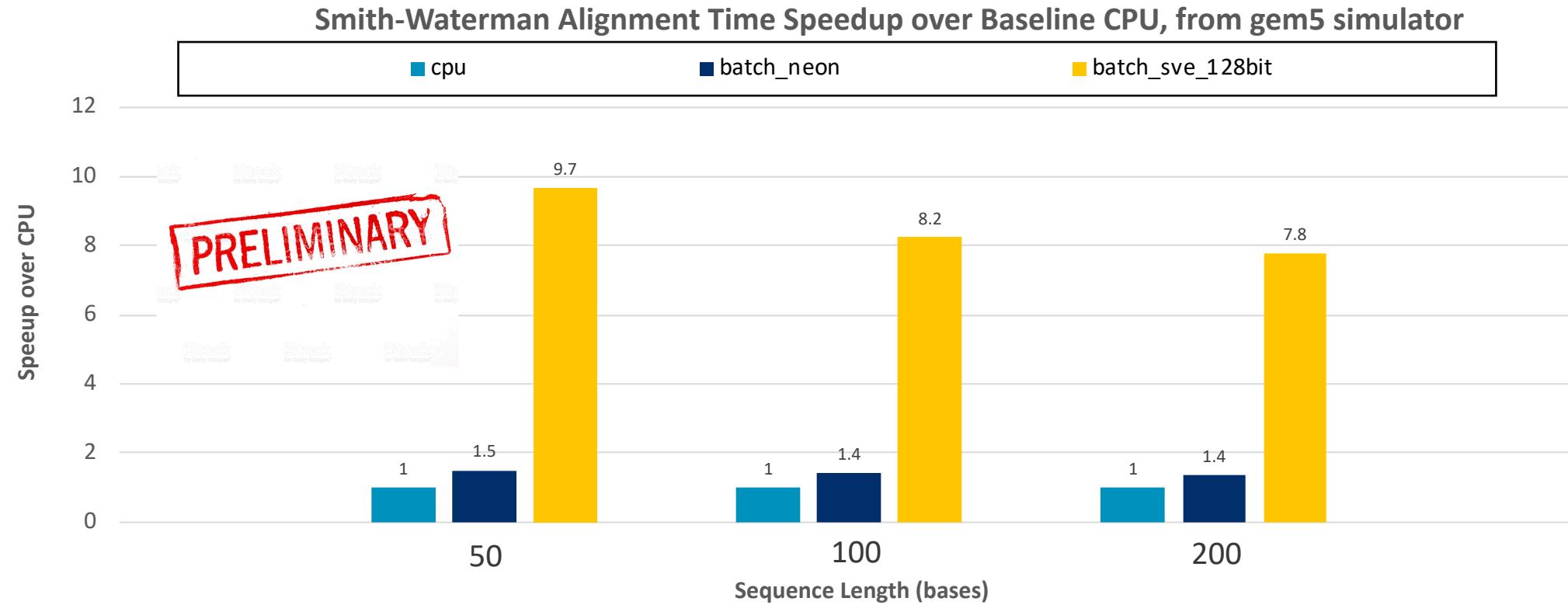
- NEON speeds up Smith-Waterman by ~2.5x



- Smith-Waterman scales up on Arm

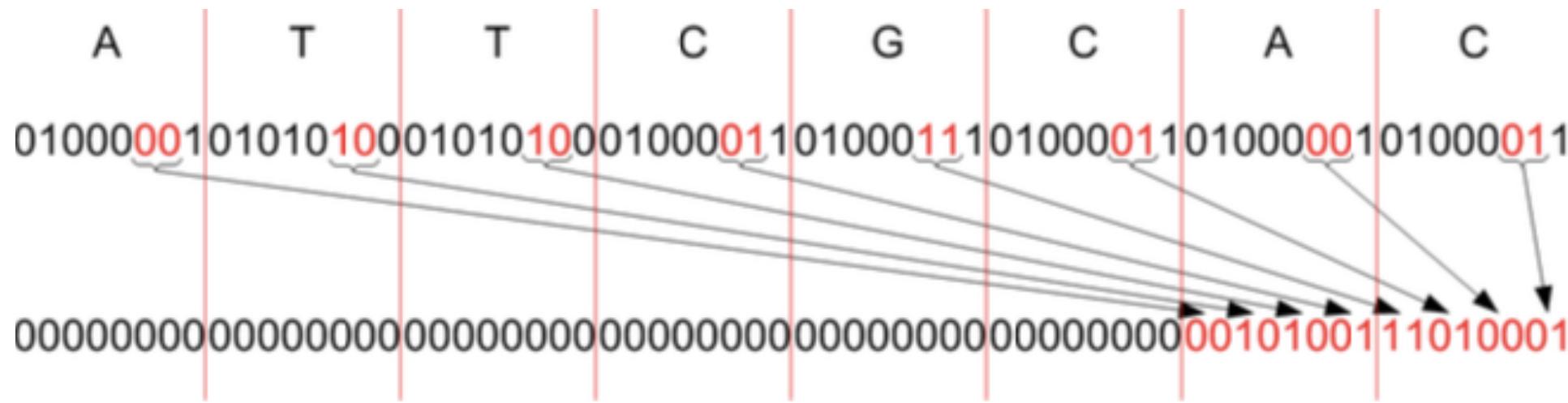


Arm SVE Accelerates Smith-Waterman Further



Bit Extraction Instruction for Sequence Compression

- DNA bases ACGT have a 2-bit representation using bits 2 and 1 of their ASCII codes. A **BEXT** instruction is used to select and compress bits 2 and 1 of each byte of a word.

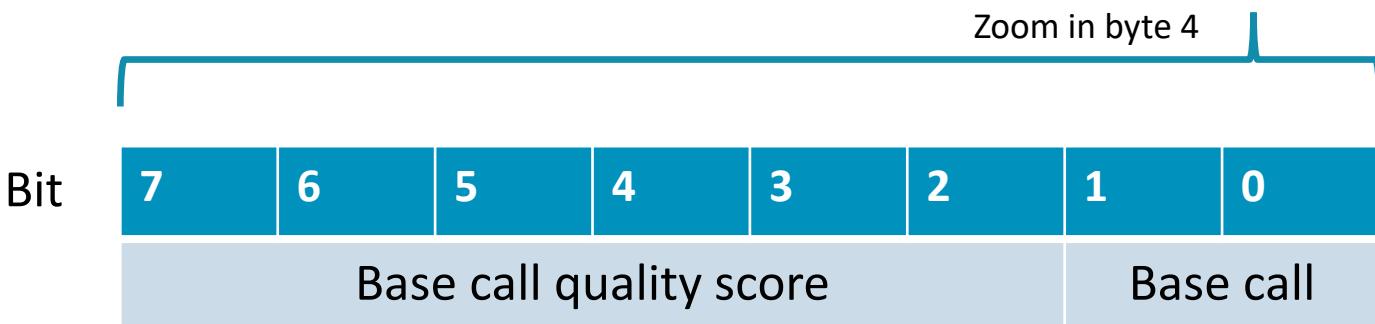


Compression of read sequence ATTCGCAC. **BEXT** to extract bit[2:1] from ATCG ASCII representations.

Read base calls - Bcl binary file format

(N+4) bytes, lowest 4 bytes as size N, each N byte contains quality score and base call

Byte	N+3	...	5	4	3	2	1	0
	Each byte as uint8_t has base call and quality score					Size N as uint32_t (little endian)		



Bits 0-1 are the bases, respectively [A, C, G, T] for [0, 1, 2, 3]:

Bits 2-7 are shifted by two bits and contain the quality score.

All bits '0' in a byte is reserved for no-call [N].

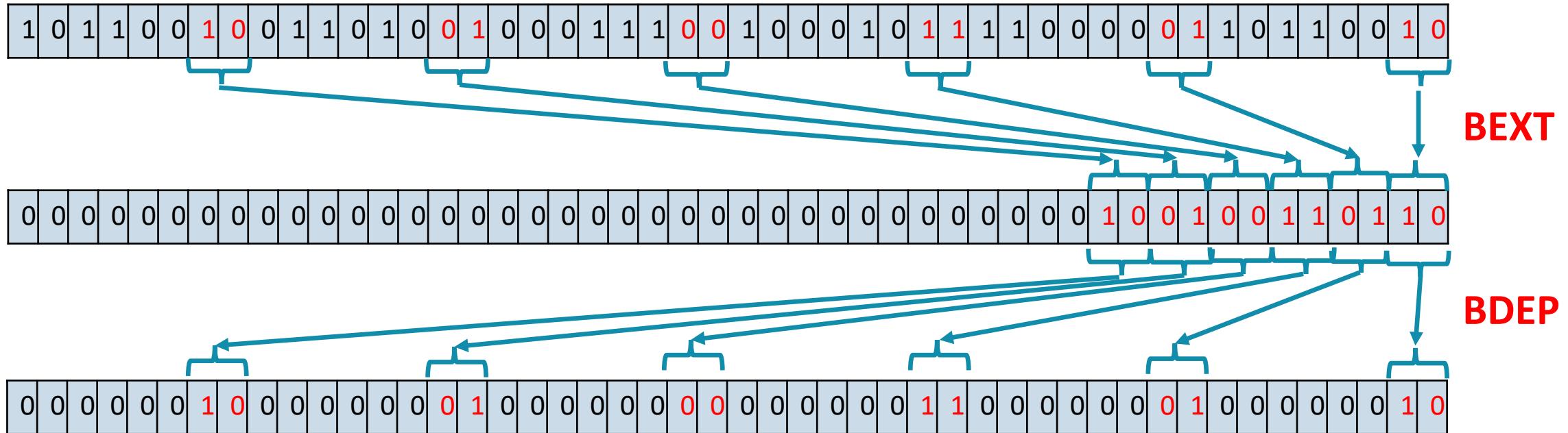
Bcl file format is the base call output from Illumina sequencers.

Illumina is the market leader for genome sequencing instruments and has over 70% market share.

isaac::alignment::Read::decodeBcl takes **3.64%** of runtime, in top **7** hottest functions.

Pack and unpack base calls

The base calls can also be packed and unpacked for processing with **BEXT** and **BDEP**



Vector table lookup can be used to turn bit[1:0] into [A, T, C, G] base calls.

Source: <https://github.com/Illumina/Isaac3/src/c++/include/oligo/Nucleotides.h>

Academic Collaborations

Trevor Mudge

- Arm works with UoM to explore and enhance SVE for genomics

Ignacio Medina

- Arm works with UoC HPC on openCB HPG-lbs support for Arm, i.e., Smith-Waterman

Onur Mutlu

- Arm works with ETHZ on processing in memory (via SRC funded projects)



Future Research Directions

Compute

- Scale up (multicores) and out (cluster)
- Accelerator IP for genomics, new instructions for SVE

Memory

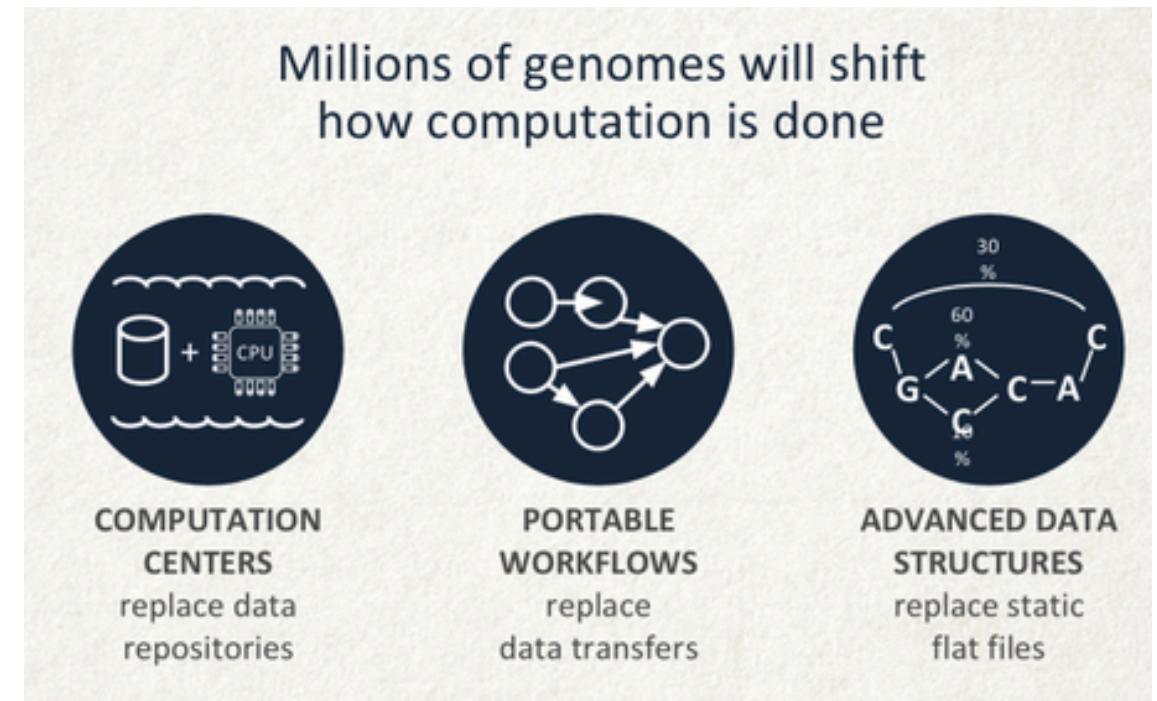
- Improve throughput of memory subsystems, i.e., memory compression, 3D stacked HBM
- Non-volatile memories, i.e., indexes
- Processing in memory, i.e., filtering in seeding stage

Storage

- Column major file format to improve data locality for parallel processing, i.e., aggregate genomic data format
- Compression algorithms and accelerators, i.e., CRAM

Compute vs. Store

- Compute is cheap, storage is expensive



Summary

Healthcare is an important emerging market for Arm

We started with genomics and we plan to do more

Look forward to working with the academic community

Thank You

The Arm trademarks featured in this presentation are registered trademarks or trademarks of Arm Limited (or its subsidiaries) in the US and/or elsewhere. All rights reserved. All other marks featured may be trademarks of their respective owners.

The Arm logo consists of the lowercase word "arm" in a bold, white, sans-serif font.

www.arm.com/company/policies/trademarks