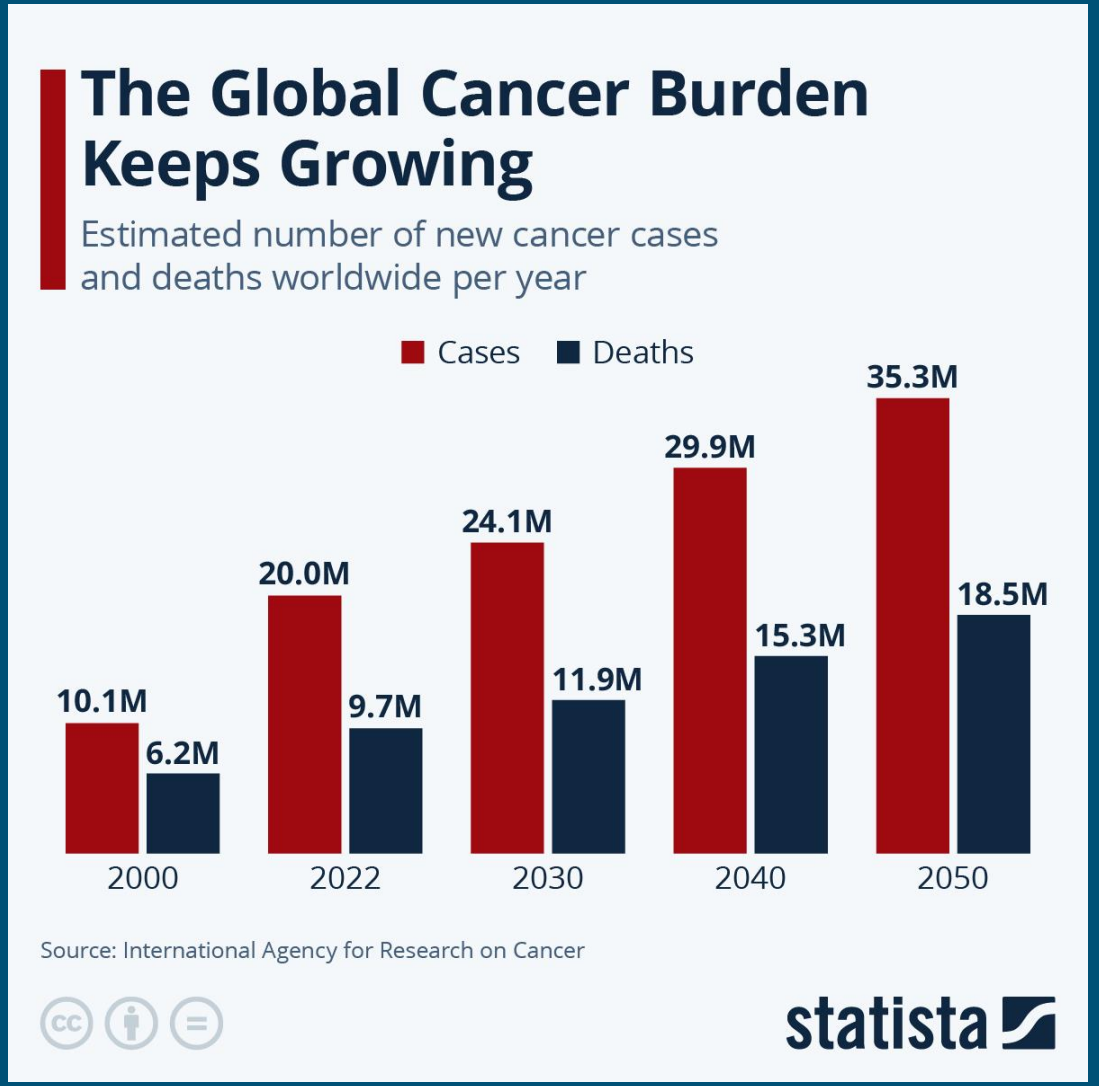
UNIVERSITY
OF COLOGNE

# EVALUATING TRUSTED EXECUTION ENVIRONMENT PERFORMANCE FOR GENOME SEQUENCE ALIGNMENT: AN AMD SEV CASE STUDY

Robert Keßler, Lech Nieroda, Simon Volpert, Moritz Gräf, Viktor Achter, Laslo Hunhold, Stefan Wesner

**Acknowledgement**
This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projects SFB1399 (ID-413326622) and GHGA (ID-441914366).

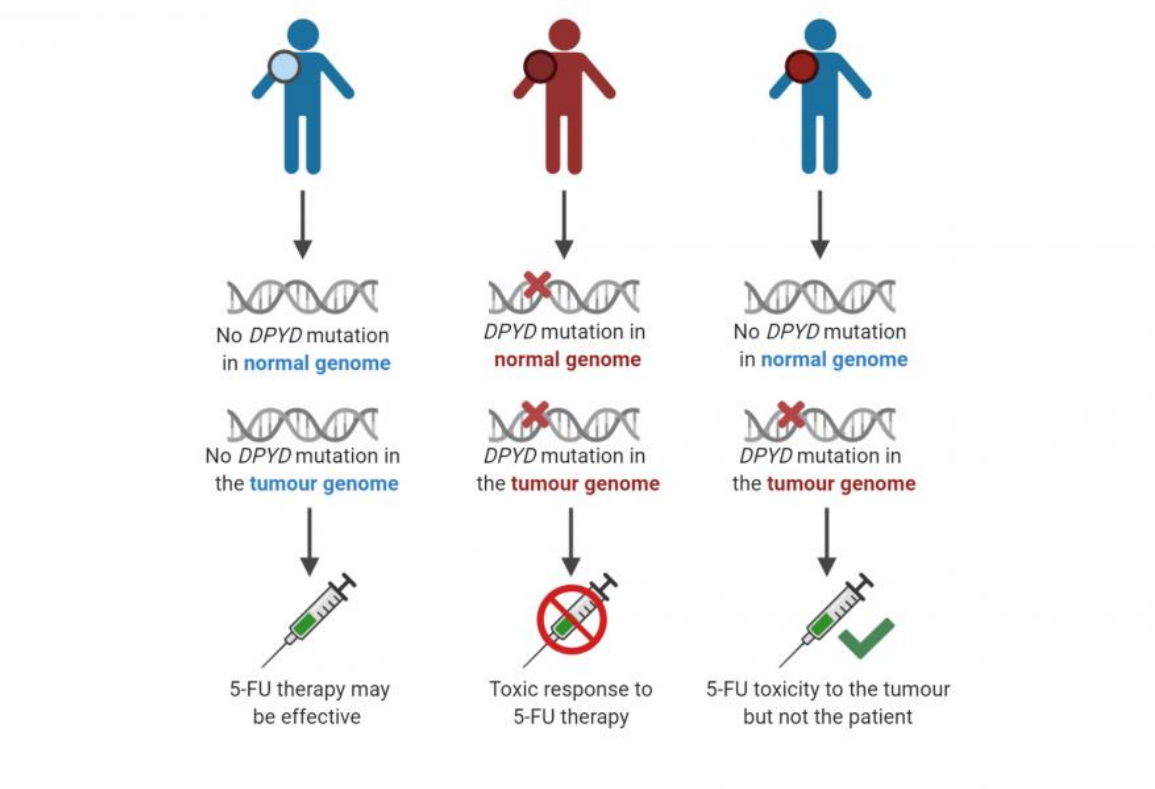
- The number of cancer diagnoses is steadily rising
- Nowadays, it is no longer a question of whether you will get cancer, but rather when
- Cancer has become the primary driver of healthcare cost increases




https://www.statista.com/chart/20706/new-cancer-cases-per-year/

# Human Genome Sequencing
# The Enabler of Personalized Treatment

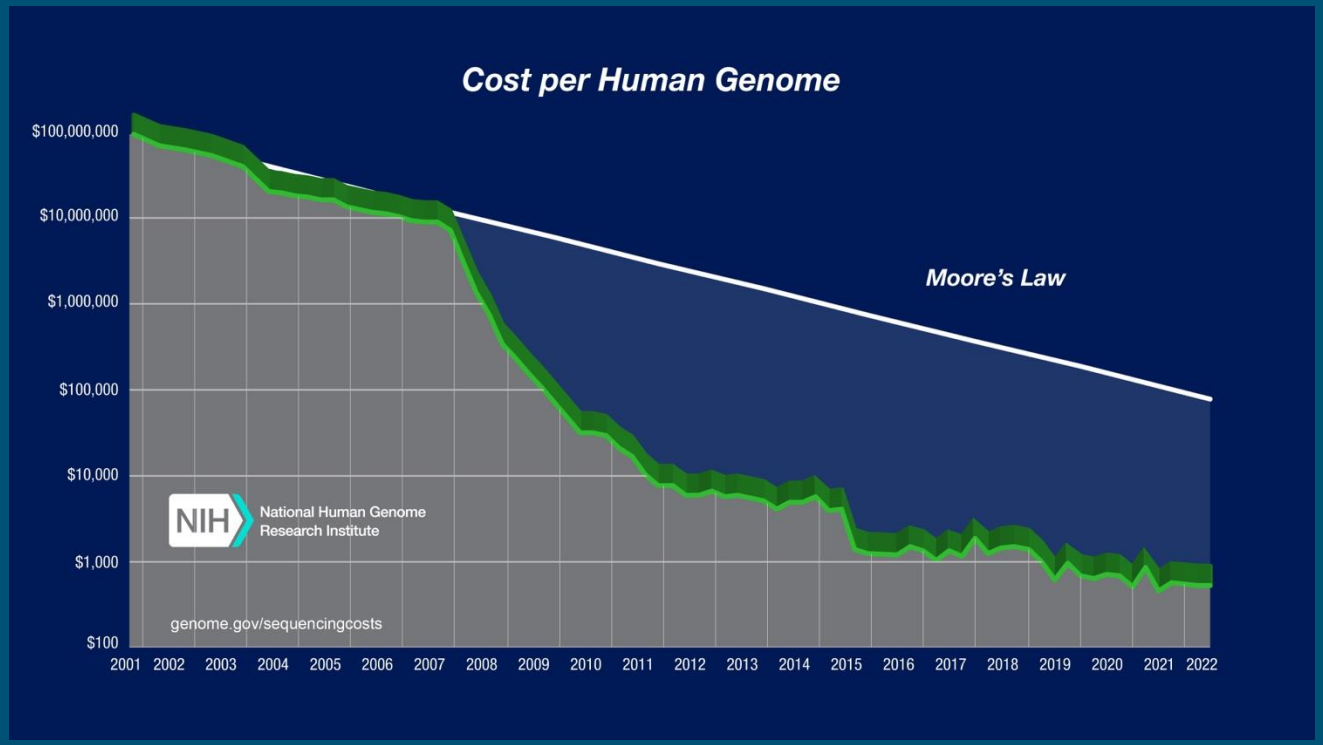
- Cancer genomics aims to decode each patient's normal and cancer genomes to enable clinicians to determine the best individual treatment
- Immense potential of integrating whole genome sequencing
  - allows the identification of mutations that could be exploited




https://www.bcgsc.ca/news/genome-sequencing-helps-prioritize-cancer-treatment-options

# Human Genome Sequencing
# The Cost of Sequencing

- Sequencing costs dropped a lot over the past 2 decades, due to:
  - Development of Next-generation sequencing (NGS) in the early 2000s
  - Continuation of Moore's Law during this time
  - Advanced alignment algorithms




https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

# Human Genome Sequencing
# The Threat Scenario

- Problem: Patient data are highly sensitive
- Electronic patient records contain personal data on current and past illnesses
- In fact, this is even more critical in the case of human genome sequencing
  - Previously unknown predispositions to diseases, which may even be hereditary, could be discovered


Image created with Perplexity AI

# Human Genome Sequencing
# The Threat Scenario

- Should unauthorized third parties gain access, this may lead to serious implications
- Consequence: Digital patient data must be secured
  - at rest (stored),
  - in motion (transiting),
  - and in use (during processing)

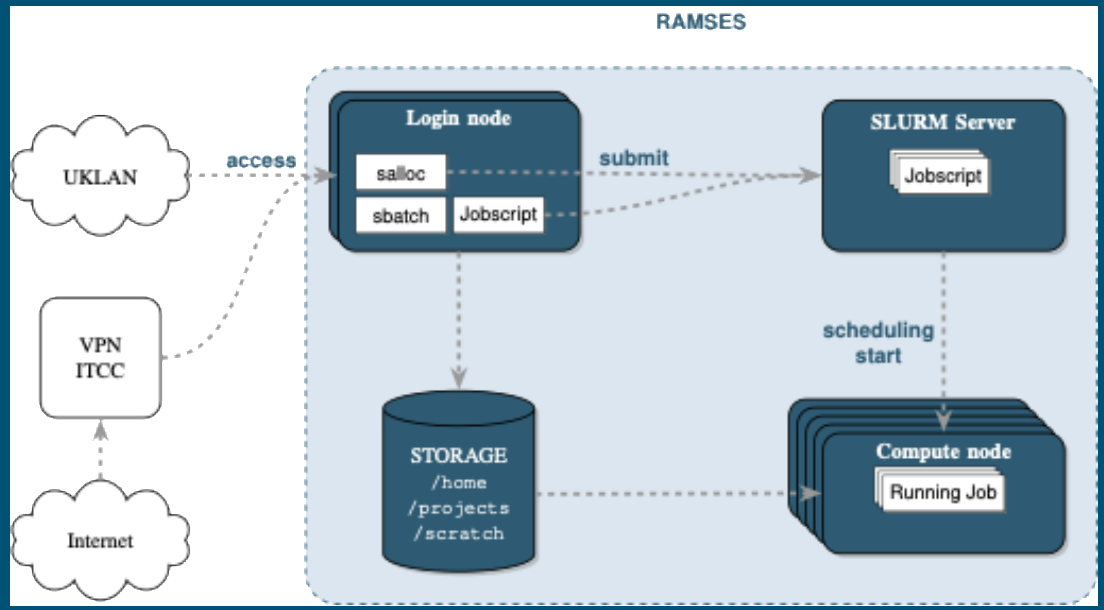

Image created with Perplexity AI

# RAMSES

- Research Accelerator for Modeling and Simulation with Enhanced Security
- Some specs:
  - 164 Compute nodes (AMD EPYC GENOA 9654 CPU)
  - 750 GB RAM per node
  - 40 nodes with NVIDIA Hopper H100 GPUs

# RAMSES
# Security Measurements

- Multi-factor authentication with personal SSH keys
- Parallel filesystem GPFS as well as RAM are encrypted
- AMD SEV-based Trusted Execution Environment (TEE)
  - So far, no support for multi-node TEEs via encrypted MPI



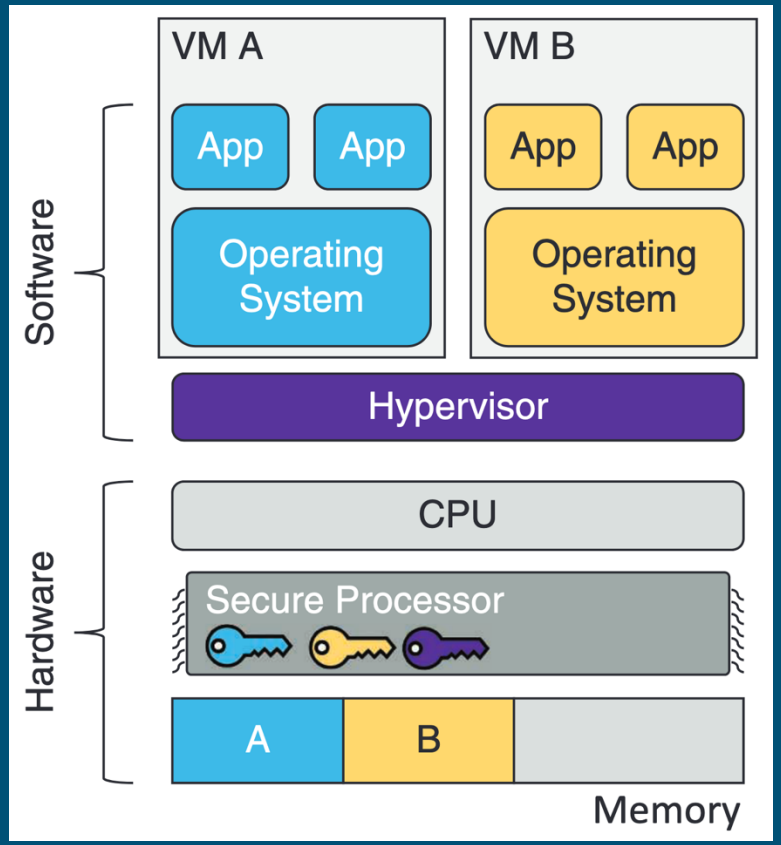

https://doi.org/10.1145/3731599.3767558

# Trusted Execution Environments

- Tamper-resistant processing environment that runs on a separation kernel
- Guarantees:
  - the authenticity of the executed code,
  - the integrity of the runtime states (e.g. CPU registers, memory and sensitive I/O),
- (a) Application- vs (b) virtualization-based TEEs
  - (a) isolates individual processes
  - (b) isolates entire virtual machine

# AMD TEEs

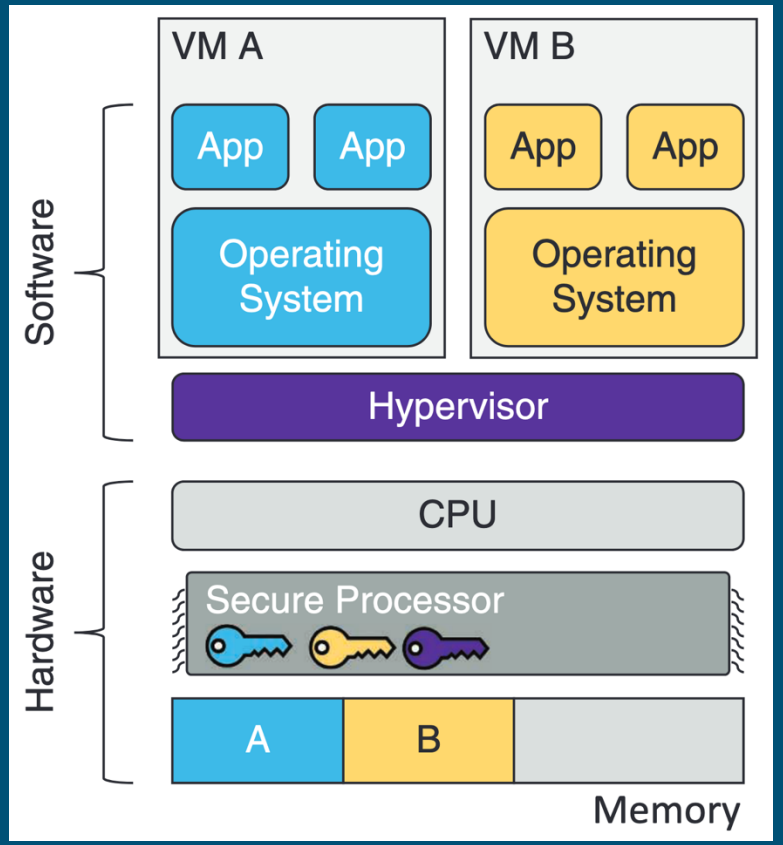
- AMD Secure Memory Encryption (SME)
  - Uses a single key to encrypt system memory
  - Memory encryption is transparent and can be run with any operating system
- AMD Secure Encrypted Virtualization (SEV)
  - Single key per virtual machine to isolate guests and the hypervisor from one another
  - Keys are managed by the AMD Secure Processor
  - VM indicates which pages in memory should be encrypted




https://doi.ieeecomputersociety.org/10.1109/MSEC.2019.2947124

# AMD TEEs

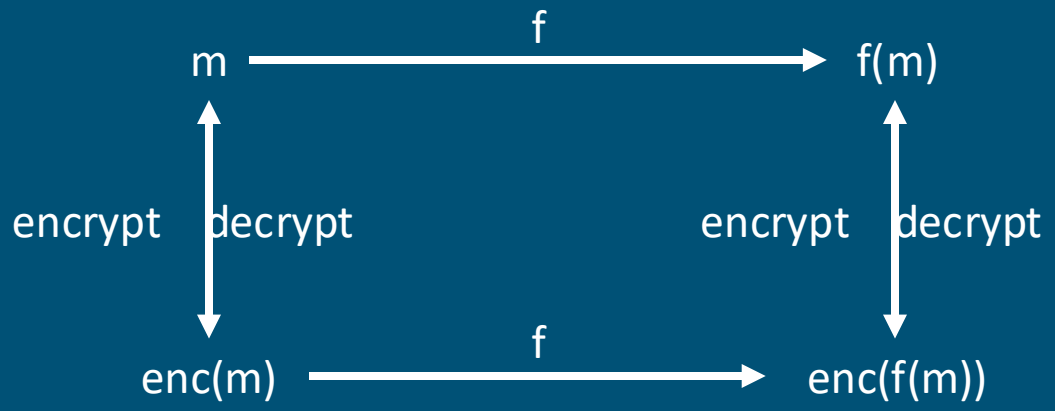
- AMD Secure Encrypted Virtualization-Encrypted State (SEV-ES)
  - Encrypts all CPU register contents when a VM stops running
  - Prevents the leakage of information in CPU registers to components like the hypervisor
- AMD Secure Encrypted Virtualization-Secure Nested Paging (SEV-SNP)
  - Strong memory integrity protection to help prevent malicious hypervisor-based attacks like data replay, memory re-mapping
  - Stronger protection around interrupt behavior




https://doi.ieeecomputersociety.org/10.1109/MSEC.2019.2947124

# TEE Alternative

- Fully Homomorphic Encryption (FHE)
  - Data stays encrypted even during processing
  - Result for unencrypted and encrypted data processing is the same
- BUT: significant performance penalty and in HPC performance is the key metric



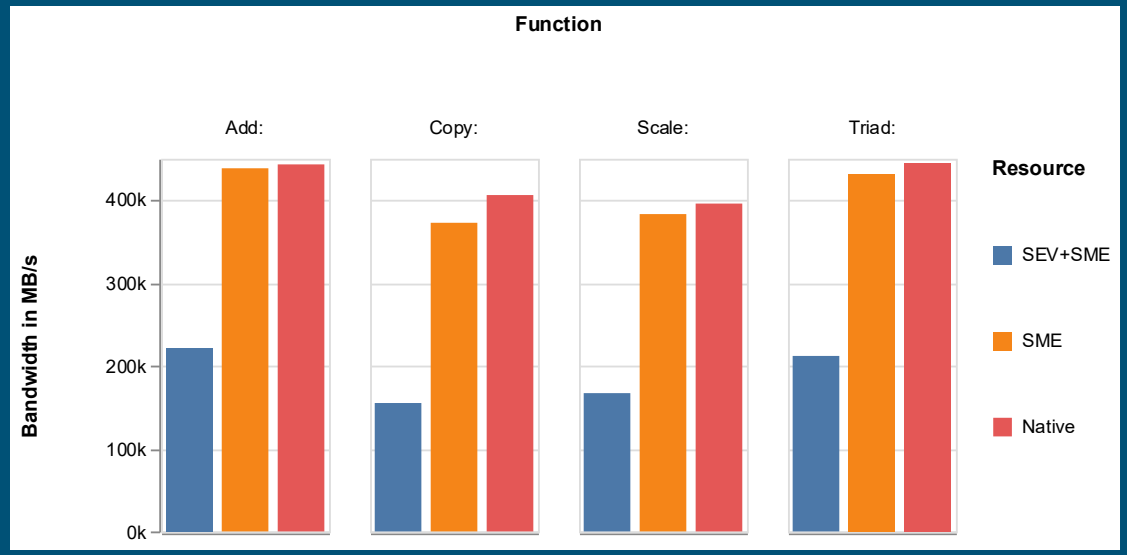
# Methodology

- To benchmark the RAMSES supercomputing system, we applied the following evaluation matrix
  - Resource set R={Native, SME, SME+SEV}
  - Thread counts T={32, 64, 96, 128, 168}
  - Storage location S={encrypted, local, scratch, shm}
  - Benchmarks B={STREAM, NPB, BWA-MEM2}
- Each combination of the evaluation matrix has been carried out with a sampling rate of 10
- For the BWA-MEM2 benchmark we used the *human_g1k_v37.fasta* human reference genome [1]

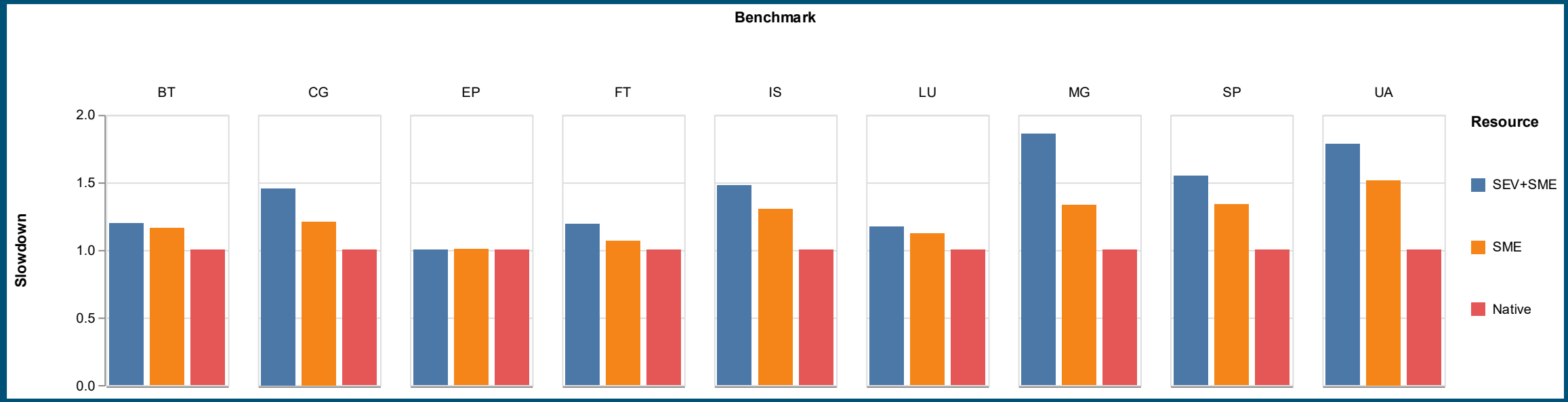
[1] https://gatk.broadinstitute.org/hc/en-us/articles/360035890711-GRCh37-hg19-b37-humanG1Kv37-Human-Reference-Discrepancies#grch37

# Results
# STREAM Benchmark

- Peak memory bandwidth of 443GB/s for the *native* resource type
- Pure memory encryption (*SME*) only causes marginal performance loss
- However, *SEV+SME* causes 55.4% reduction in memory bandwidth



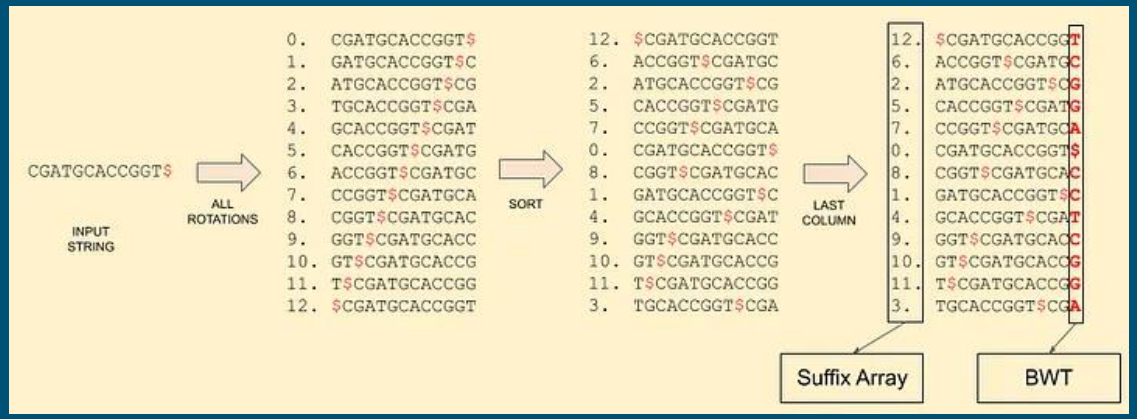
# Results
# NPB Benchmark




- Memory-bound workloads have remarkable performance drops for *SME* and *SEV+SME*
  - e.g. MG and UA
- Compute-bound workloads only suffer from minor performance overhead
  - e.g. BT, LU and EP

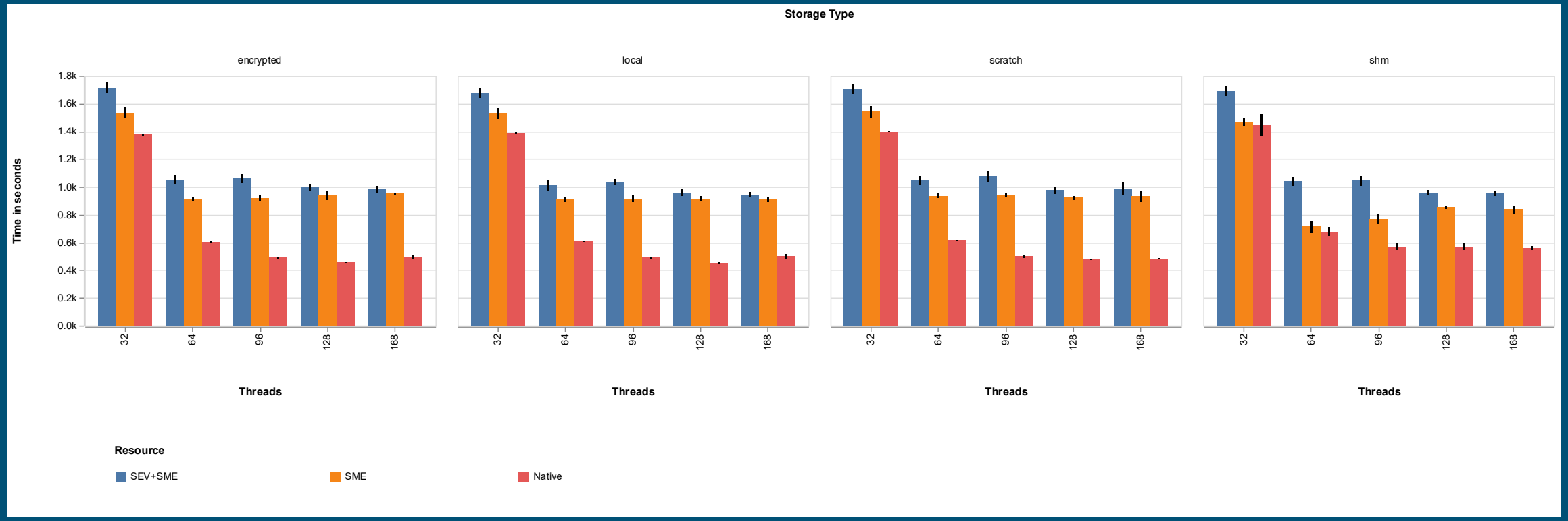
# Burrows-Wheeler Aligner (BWA)

- BWA is a software package for mapping low-divergent sequences against a large reference genome
- For the benchmark in our paper, we used the latest version BWA-MEM2
  - Index size on disk is down by 8 times
  - Required memory down by 4 times



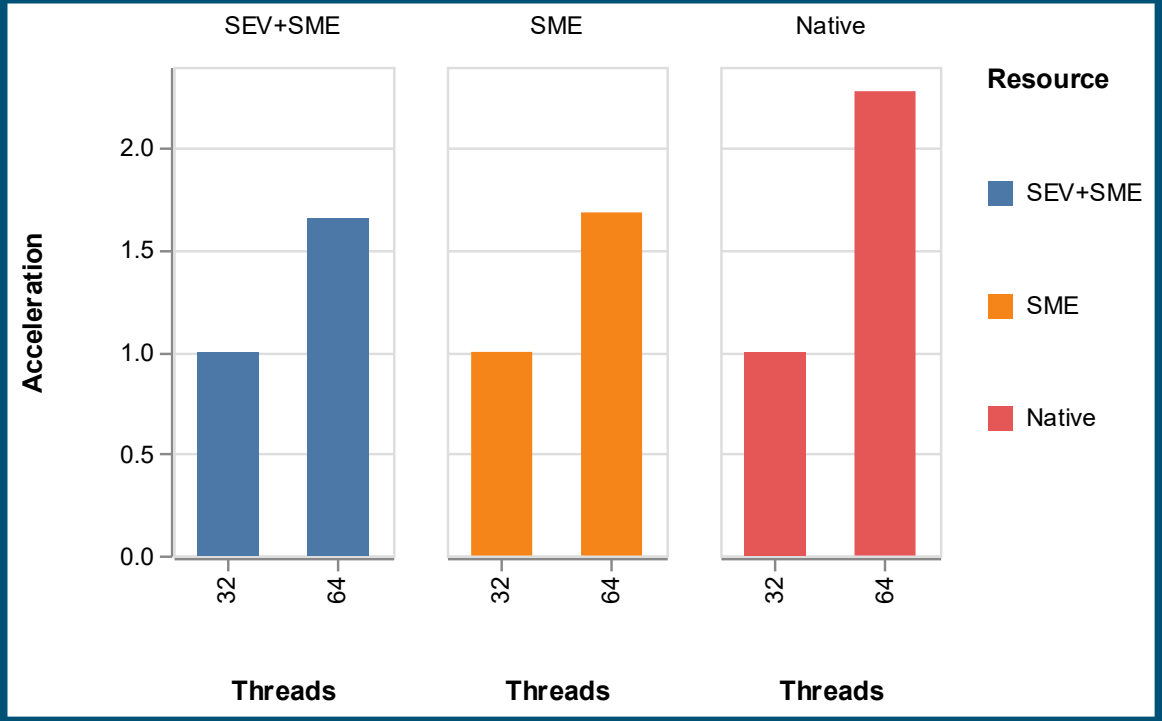

https://github.com/bwa-mem2/bwa-mem2

# Results
# BWA-MEM2



# Results
# BWA-MEM2

- Figure on the right shows the acceleration with input data from the *local* file system
  - Values are normalized to the 32-thread performance for each resource type
- *Native* environment demonstrates the most efficient scaling



Feel free to contact me:

kessler@uni-koeln.de