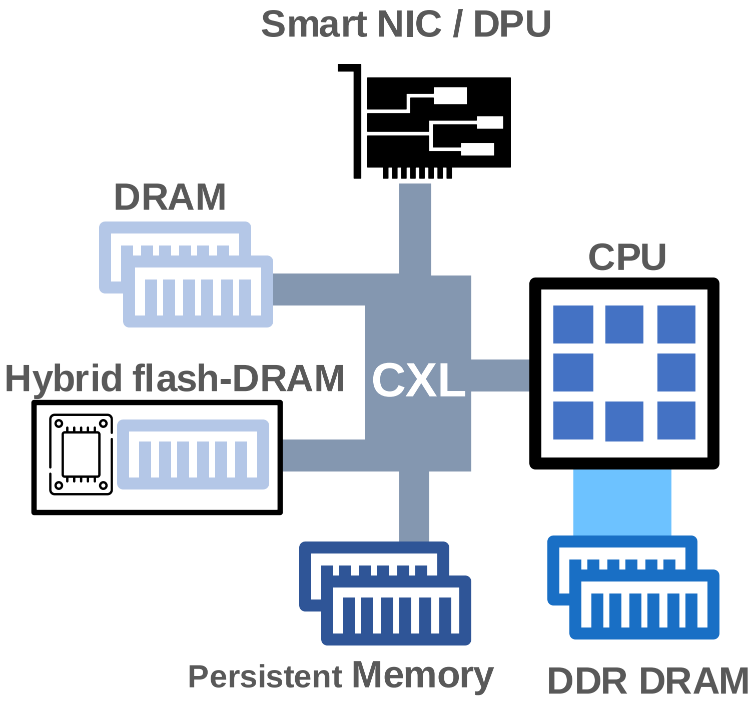
Leaky Memory Abstractions: Hidden Performance Loss on Modern Hardware

Hamish Nicholson

www.nicholson.ai

hamish.nicholson@epfl.ch

The Future of Memory is More than Local DRAM



CXL is an *interface*

Media agnostic

Performant memory mapped IO

- i.e. byte-addressable load-store transactions

CXL extends load-store semantics to the rest of the server

Problem: CPUs Can be Surprising

Hardware

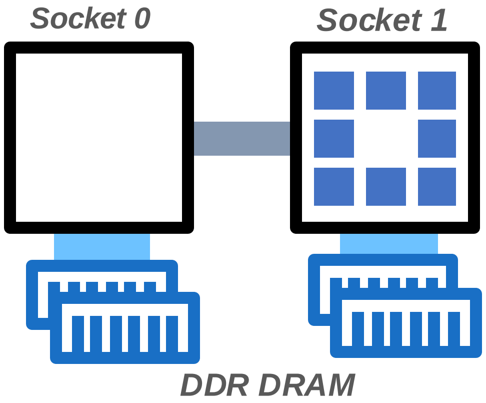
- 2x 24-core AMD EPYC 7413.

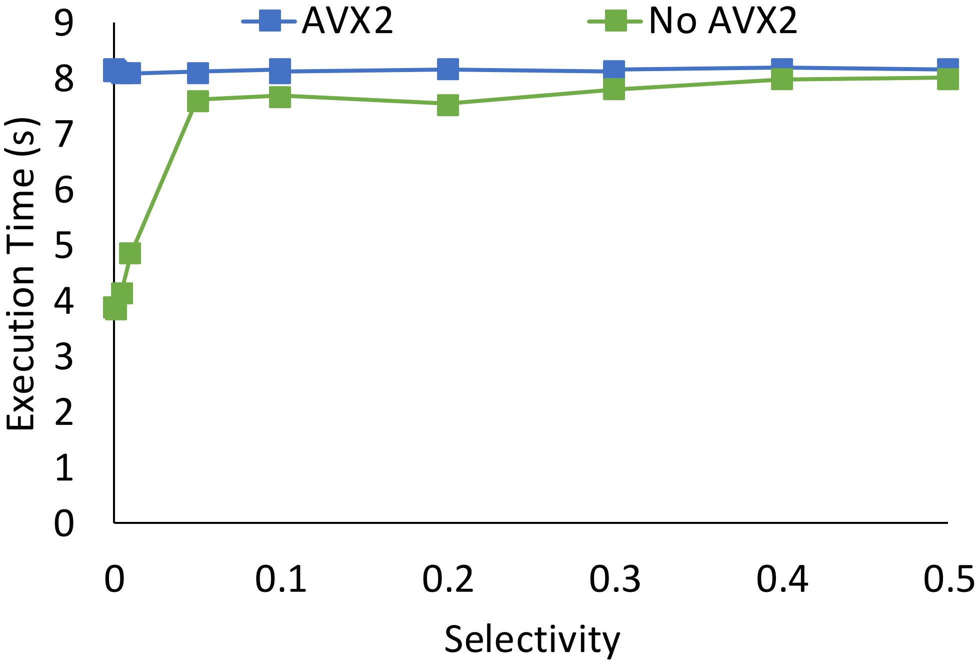
Software

- 2x 100GiB Int32 columns on **socket 0**

- Scan-filter-sum query on **socket 1**

- LLVM 14 w & w/o “-avx2 -avx"





Up to 2x penalty from compiler auto-vectorization

What if Everything was Memory?

- NUMA on steroids
  - Heterogenous media mapped in the same address space with **vastly** different properties
  - What is the right level of abstraction to handle this?
- Compilers tune to the CPU, not the memory hierarchy
  - Memory is increasingly a **dynamic** resource
  - (JIT) compile for available memory hierarchy?
- CPU cache/memory hierarchy is a notoriously leaky abstraction
  - Unstandardized, **rarely specified**, and now it’s flooding beyond the CPU
  - Should HW vendors provide us performance interfaces?

www.nicholson.ai
hamish.nicholson@epfl.ch

Thank you!