Tools issues for emerging hardware

Working group at Scalable Tools Workshop 2023

https://bit.ly/stw-ehw

Existing and upcoming hardware technologies

- Different internal architectures per hardware thread on CPUs (E- & P-cores, one with AVX512, the other not)
- Acceleration engines integrated into CPUs:
  - Some with ISA changes (AVX512, AMX, …)
  - Some provide additional features (Data streaming engine, …)
  - Changes the way how to measure
- Hybrid == APU (integrated CPU & GPU)
  - Unified memory between CPU and GPU
  - OS on CPUs, (main) compute on GPU
- Deep Learning/ML/AI processors
  - Different architectures and way to program
    - Data flow
    - Stored-program computer
    - Mix of both
  - Examples: Cerebras WSE-2, Habana Gaudi, Sambanova SN30, GroqChip, Fraunhofer STX, NextSilicon

Existing and upcoming hardware technologies

- Memory technologies
  - On-chip HBM (flat or cache mode)
  - Off-chip DDR
  - Integrated SRAM
  - Optane/Flash/NVMe

- Quantum
  - Tools for the link between control system and quantum “processor”

Tool concerns

- Different scope of tools:
  - Correctness
  - Debugging
  - Performance measurement tools
  - And more
- What do we need?
  - Device utilization, resource usage, what's happening?
  - Data movement /Communication problems?
  - Opportunities for better (data flow) mapping or scheduling?
  - Inefficiencies? Imbalances? Underutilization of functional units?

High complexity in devices with exploding architectural variety
Complexity in the HW market seems not to be doable for single tool groups

How to proceed

- White paper from this community?
  - What is required and why?
  - Can be used as checklist in procurements (or bids need to be graded by tool experts)
  - List of successful guidance
    - HPCToolKit and PAPI sent feature requests to vendors
      → Some successfully
  - (Maybe) sparks collaboration between
    - HW designers
    - Vendors
    - Standards committee
    - National labs
    - Academic partners
- Collaboration with a HW architect to estimate costs for a feature and the potential usage for tools?
- We need to be more visible also in other communities:
  Game developers, show successful perf. optimizations to ML/AI community