Current Hardware Specifications and the Needs at NIST

Tim Blattner

NIST National Institute of Standards and Technology
U.S. Department of Commerce

ITL

Disclaimer

No approval or endorsement of any commercial product by NIST is intended or implied. Certain commercial software, products, and systems are identified in this report to facilitate better understanding. Such identification does not imply recommendations or endorsement by NIST, nor does it imply that the software and products identified are necessarily the best available for the purpose.

Outline

- Hardware
- Software
- Containers

Hardware Trends (Disks and RAM)

- Disk (sequential)
  - HDD ~ 100-200 MB/s, $
  - SATA3 SSD ~ 550 MB/s, $$
  - PCIe NVMe SSDs (peak), $$$
    - Gen 3 ~ 3500 MB/s, year ~2013
    - Gen 4 ~7000 MB/s, year ~2019
    - Gen 5 ~14000 MB/s, year ~2021
    - Gen 6 ~28000 MB/s, year ~2024
- RAM (per channel)
  - DDR3-1600 ~12800 MB/s
  - DDR4-3200 ~25600 MB/s
  - DDR5-4800 ~38400 MB/s
- Transformative impact on storage speed, within very short timeline

Hardware Trends CPUs

- Intel, AMD, ARM CPUs all packing more cores per socket
  - Workstation with 64 physical cores
  - Servers with 192+ physical cores (growing year-by-year)
- Clock speeds increasing
  - Dynamic overclocking (increases clock speed given enough cooling)
    - Increases the number of instructions per second
- Hardware-accelerated vector instructions
  - AVX512, packing 512-bits into a single instruction
- High memory bandwidth
  - 12-channel memory controllers

Hardware Trends GPU Compute Capabilities

- NVIDIA's specifications for capabilities of a GPU
- Each generation of GPU falls in a version of the compute capability
  - P100: compute capability: 6.0, H100 compute capability 9.0
- Define capabilities of hardware
  - TensorCore requires compute capabilities 7.x or greater

- **Compute capabilities map to GPU hardware architecture (NVIDIA)**
  - CUDA Software is compiled with specific versions of compute capabilities
    - nvcc -arch=compute_60,compute_70
  - Typically supports forward compatibility (older compute capabilities work on newer GPUs)
    - Some software this is not the case
      - PyTorch, Tensorflow requires recompilation to target newer GPUs

Software

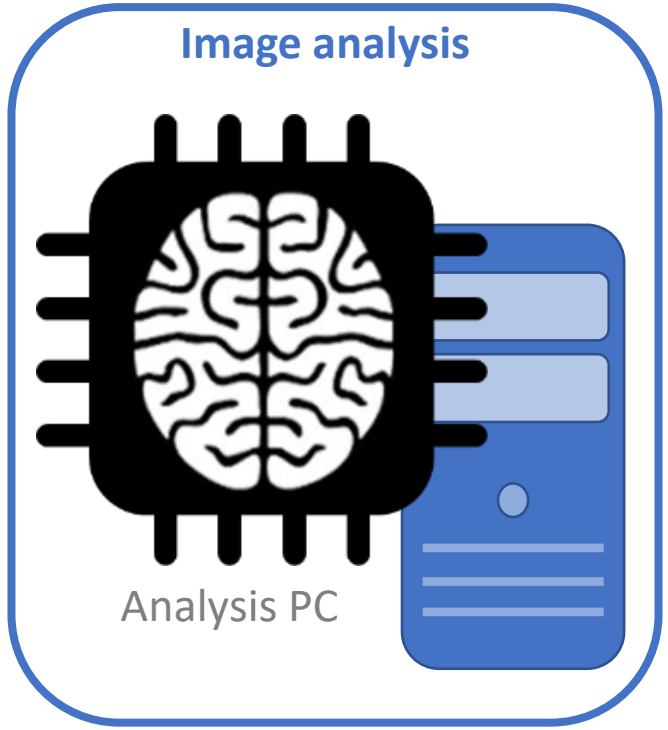
- Traditional image processing workflows
  - Whole image analysis
    - Require analysis on every pixel in an image
    - Might require fitting entire image in RAM
  - Partial/Tile-based image analysis
    - Targeted towards working on regions of an image
      - Region contains enough signal to analyze
    - Requires loading only a regions of an image at a time
  - Output: metadata or transformed image(s)
    - Types of outputs could cause failure points if not careful
      - Example: Outputting several large images

- Hardware Impacts:
  - Memory implications
    - Whole image versus tile-based
  - Execution time
    - Multi-core solutions versus GPU
    - Loading image for disk
      - SSD vs HDD vs NVMe SSD

- Failure Points:
  - GPU hardware requirements
    - Compute capabilities
  - Out-of-memory
  - Processing taking too long
  - No disk space for output
  - Others?

Software – AI

- Utilize Artificial Intelligence to analyze images
  - Tiling approaches for large resolution images
  - N-dimensional analysis (3D convolutional networks)
  - Outputs
    - Semantic segmentation, object detection, classification, …
  - AI training? Or only Inference?
- Hardware impacts:
  - Inference-only
    - Excellent mechanism for image analysis
  - Training
    - Compute, time, and memory intensive



What does this mean for containers?

- Resource Requirements are specified
  - Useful to identify minimum requirements to run container
- Ensures proper execution if container has specific requirements
  - CPU features, GPU features
  - Instruction set for CPU, GPU compute capabilities
- Challenge:
  - Legacy containers might not run on new hardware
    - Example: Upgrade GPUs to next generation
    - Requires mechanism to rebuild container
      - Recompilation of software or dependencies
        - Target next generation hardware

Property	Type
ramMin	number
coresMin	number
cpuAVX	boolean
cpuAVX2	boolean
gpu	boolean
cudaRequirements	object

https://github.com/usnistgov/fair-chain-compute-container/blob/master/docs/manifest-properties-computational-tool-resource-requirements.md

cudaRequirements:

Property	Type
deviceMemoryMin	number
cudaComputeCapability	String or array

https://github.com/usnistgov/fair-chain-compute-container/blob/master/docs/manifest-properties-computational-tool-resource-requirements-properties-gpu-cuda-related-requirements.md

What Hardware is Needed?

- Heterogenous system
  - CPU + GPU + variety of architectures
    - What kind of support do we want to have for legacy containers?
    - Is it reasonable to have access to source code to recompile legacy containers?
  - Different instruction set architectures?
- AI workloads will consume a tremendous amount of compute
  - Especially if you want to include training
    - May require multi-GPU or even multi-node
- How many compute nodes?
  - How many concurrent users?
  - What execution time for workflows?

Missing anything in the container spec?

- Estimating execution requirements
  - Memory usage could be based on input
  - Execution time estimation to ensure reasonable execution (avoiding starvation)
- Do we need a more dynamic structure that can be used?
  - Or is this too much to ask for container developers?
- Possible to track this metadata as containers are executed
  - Keep track of statistics to measure averages

Thank you