NCATS

COLLABORATE. INNOVATE. ACCELERATE.

Hardware Heterogeneity at NCATS

more hardware, more problems

December 7, 2023

Nick Schaub, Ph.D.
Senior Data Scientist[1]
Associate Director of Deep Learning[2]

[1]Information Technology Resources Branch, National Center for Advancing Translational Sciences, National Institutes of Health
[2]Axle Research and Technology, 6116 Executive Blvd, Rockville, MD 20850

NIH National Center for Advancing Translational Sciences

Diverse Users

Programmers
- SDKs and libraries

Technical
- Command line tools & terminal
- Direct edit YAML/JSON

Advanced Domain Experts
- Complex GUIs
- Configure/optimize hyperparameters
- Some "Macro" scripting

Domain Experts
- Simple GUIs
- Single button if possible!

Diverse Needs

Language Support

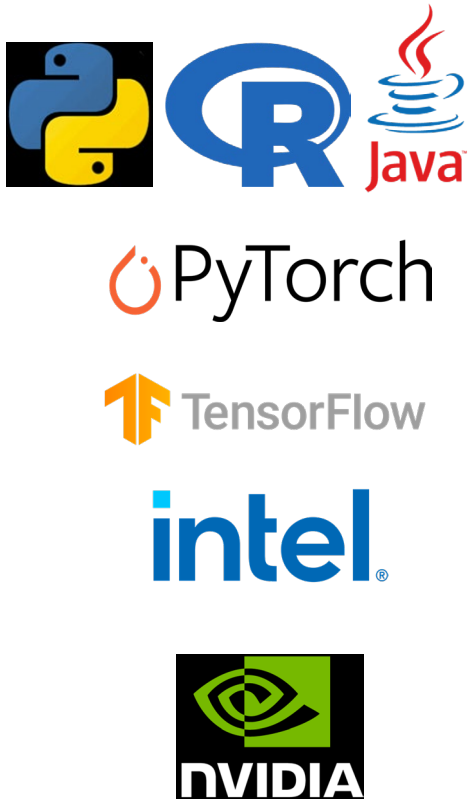
- Python, Java, R, C/C++, and Rust

Diverse Libraries

- Traditional, Tensorflow, Pytorch, Jax, etc.
- Diverse driver requirements i.e. CUDA

Diverse Hardware

- CPU - x86 and ARM
  - AVX/AVX2
- GPU - Nvidia
  - Historical code written across multiple Nvidia compute capabilities
- RAM - 8GB → ∞



Scalable Algorithm Development

Things we need to do:

1. Process data larger than memory
2. Distributed computation

Things we get from the process:

1. More abstraction
2. Run on consumer grade hardware

Things we need to know:

- Number and type of CPUs (x86/ARM)
- RAM
- Storage (NAS, cloud, etc)
- GPU type and minimum driver/toolkit version

Case Study: FTL Label Tool

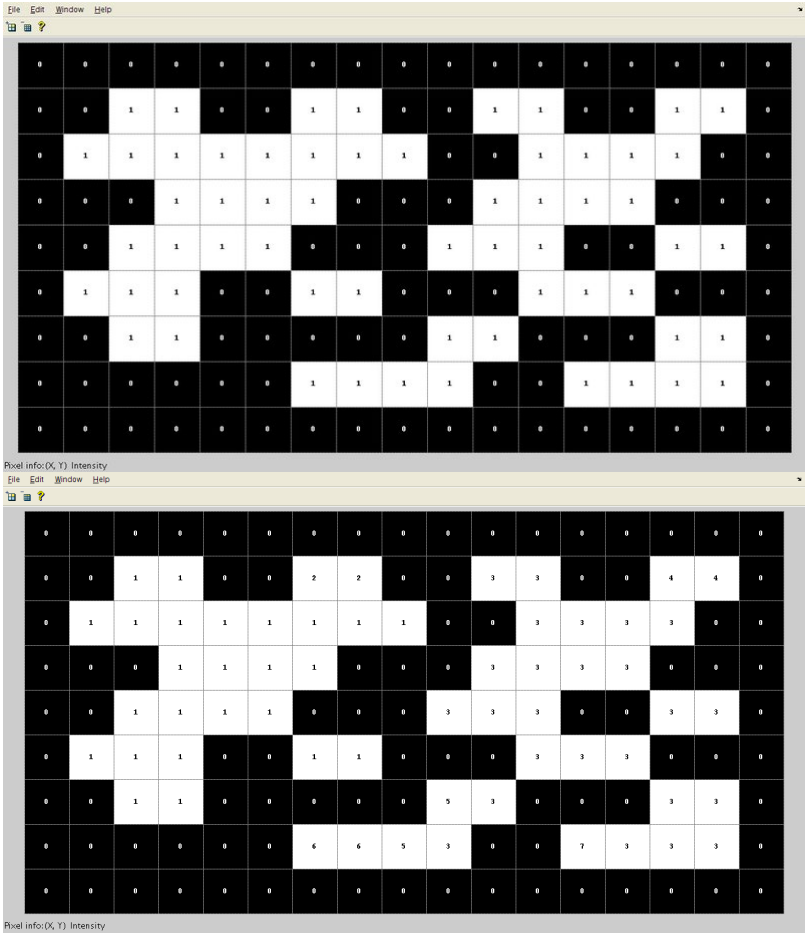
Connected component labeling is used to find discrete regions in a binary image

Developed a compressed encoding (run length encoding)

For increased speed, we used SIMD (AVX2)

“Just put it in Docker and it will work”

What about ARM? (Apple Silicon)

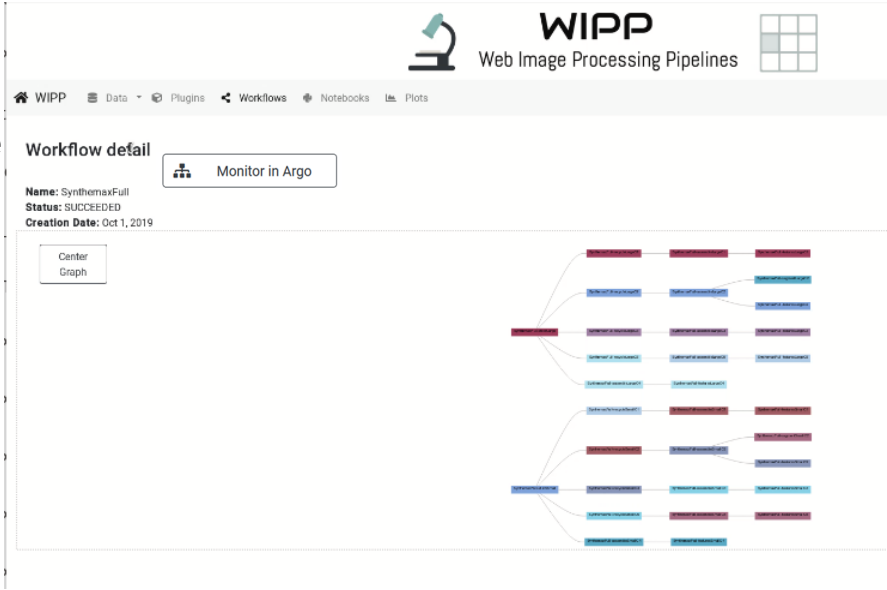


Distributed Workflows on Argo

Design philosophy: Highly specialized tools to maximize composability

What resources are available to me on the node?

Can we take advantage of tool information to reduce costs? (run low resource jobs on cost efficient nodes)



GPUs

What kind of GPU/toolkits are required?

Basic information is required

1. Compute score (nvidia)
2. Minimum driver version
3. Cuda toolkit version
4. Vulkan/opencl?

Still need more information:

1. Many unmodified LLMs require a minimum of 16GB of GPU ram
2. Many large models require multiple GPUs
3. Some tools can run on CPU if a GPU is not present

Questions?