Nextflow and Singularity enabling heterogeneous CPU/GPU bioinformatics containers in workflows

Kevin Sayers

September, 2017

Overview

- Technologies
  - Nextflow
  - Containers
  - GPU
  - AWS
  - Machine learning
- SRAGPU-nf workflow
- Results
- Conclusions

Nextflow

- Domain specific workflow language for bioinformatics
- Deploy workflows locally, in an HPC cluster, or on Amazon web services (AWS)
- Supports Docker and Singularity containers
- GitHub integration
- Automatic parallelization

Nextflow script

```
process setup{
        container = "docker://sayerskt/samtools"
        publishDir './', mode: 'copy', overwrite: true

        output:
        file "Homo_sapiens.GRCh38.cdna.all.fa" into reference
        file "Homo_sapiens.GRCh38.cdna.all.fa.fai" into refindex

        """
        wget ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz
        gunzip Homo_sapiens.GRCh38.cdna.all.fa.gz
        samtools faidx Homo_sapiens.GRCh38.cdna.all.fa
        """

}
process barracudaIndex{
        container = 'shub://KevinSayers/BarraCUDA_Singularity'

        storeDir 'index/'
        input:
        file ref from reference

        output:
        file "${reference.baseName}.*" into indexOut, indexFiles

        """
        barracuda index -p ${reference.baseName} ${ref}

        """

}
```

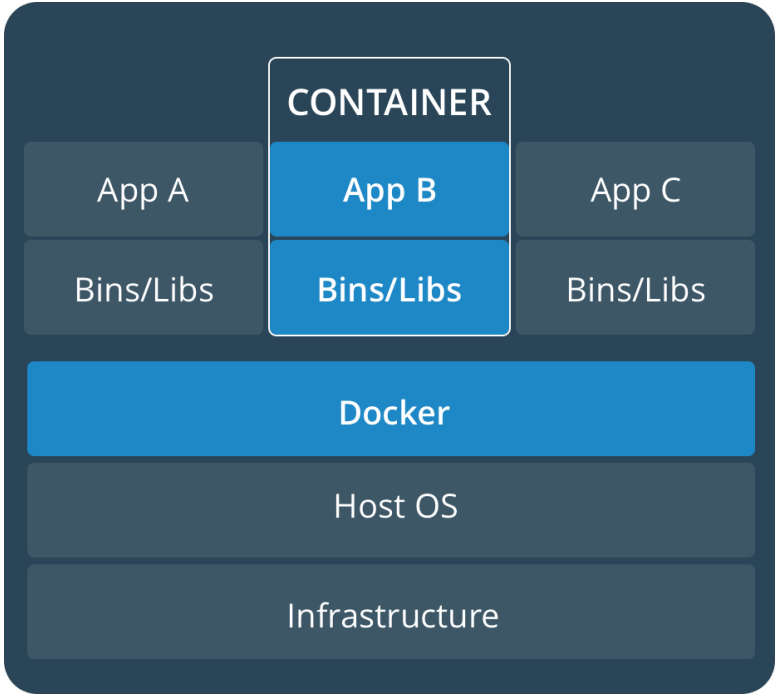
```
singularity{
        enabled = true
        runOptions = '--nv'
}

cloud {
    imageId = 'ami-f379948a'
    instanceType = 'p2.xlarge'
    spotPrice = 0.40
    bootStorageSize = '800 GB'
}
```

Containers

- Software portability
  - Packaged together
  - No conflicting dependencies
- Reproducibility
  - Created from an image
  - Control over software versions
- Bioinformatics respositories

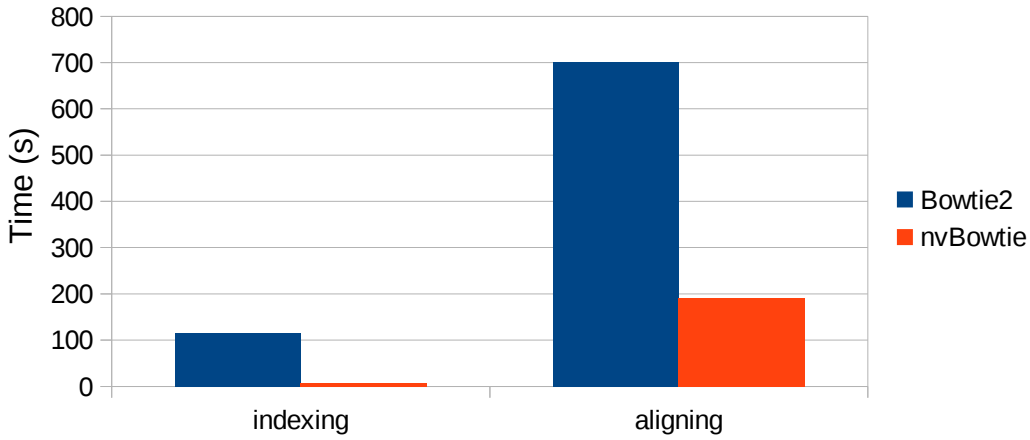
Biocontainers



https://www.docker.com/what-container

GPUs in Bioinformatics

- Thousands of parallel cores
- Can substantially reduce processing time
- Bioinformatics tools
  - Short read aligners (BarraCUDA, nvBowtie, SOAP3)
  - GPU-BLAST
  - Molecular modeling
- Machine learning



GPU containers

- Portability of difficult to compile GPU programs
- Recent work
- Implementations
  - Nvidia-docker
  - Shifter
  - Singularity
- Containerized single tools

Portable, high-performance containers for HPC

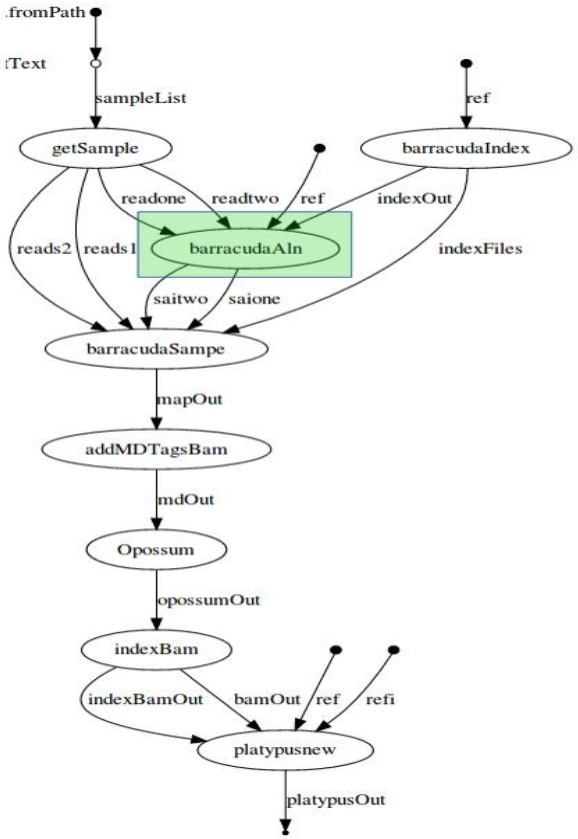
Lucas Benedicic*, Felipe A. Cruz, Alberto Madonna, Kean Mariotti
Systems Integration Group
CSCS, Swiss National Supercomputing Centre
Lugano, Switzerland
Email: *lucas.benedicic@cscs.ch

Experimental data

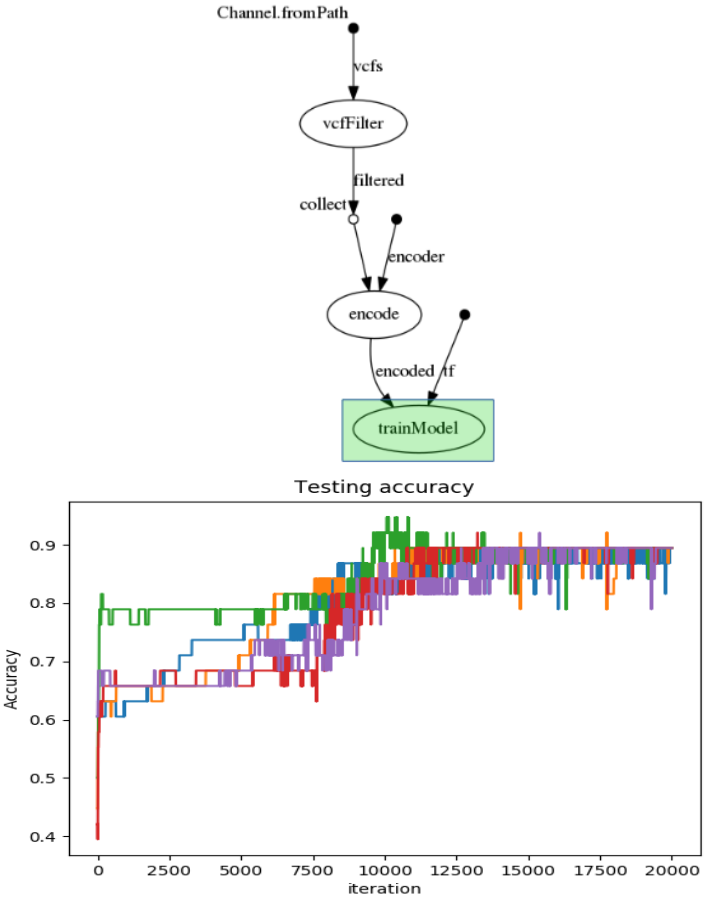
- Single cell RNA-seq
- 4 classes
  - Estrogen receptor positive (ER+)
  - Human epidermal growth factor receptor 2 (HER2+)
  - Triple-negative breast cancer (TNBC)
  - ER2+ and HER2+
- 381 samples
- Reference transcriptome
- BarraCUDA GPU enabled short read aligner
- TensorFlow machine learning classifier

SRAGPU-nf

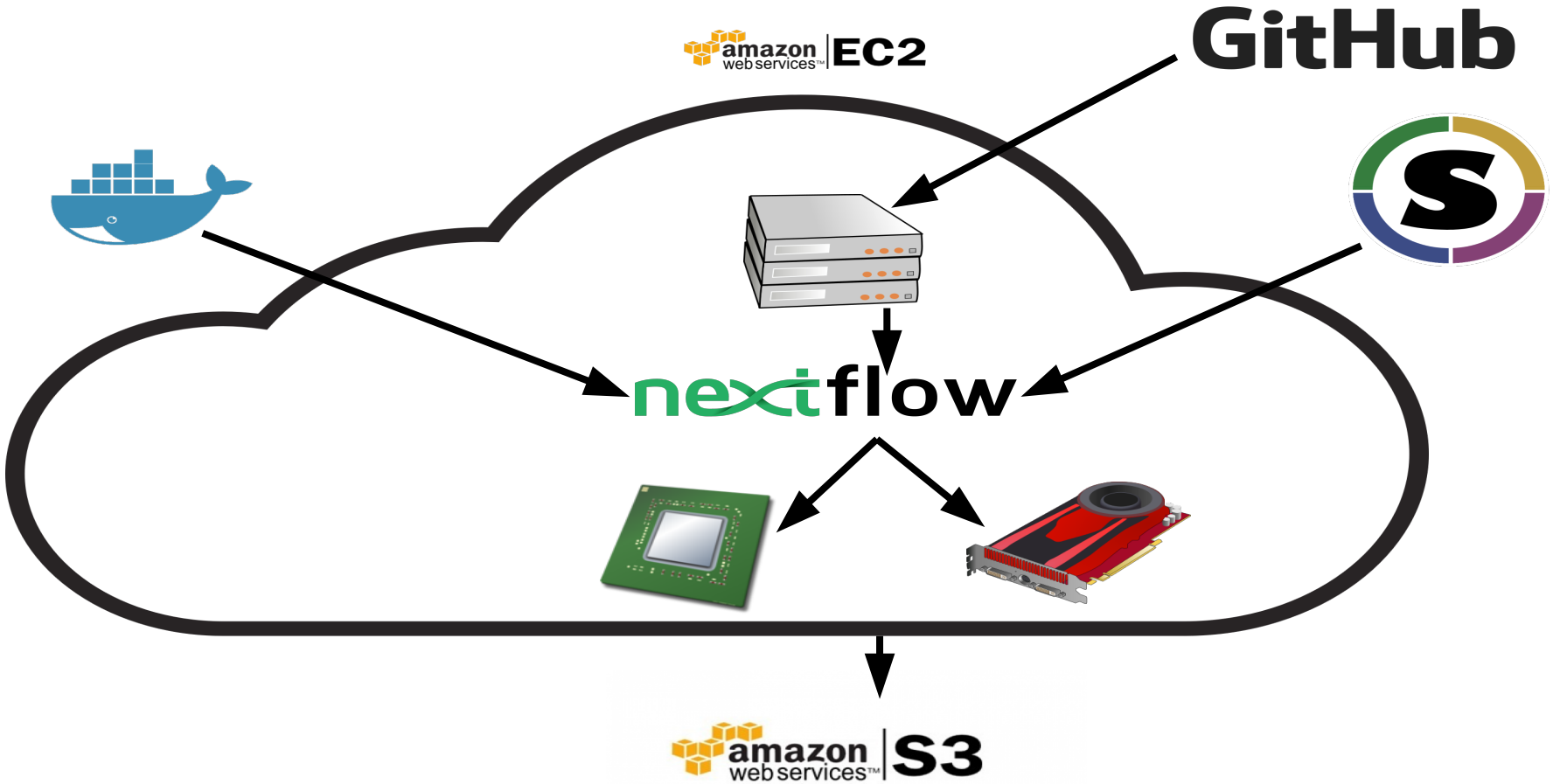
Sample processing



Classification



Cloud configuration



Conclusions

- GPU bioinformatics tools can decrease processing times
- Nextflow and Singularity can be used to deploy heterogeneous container workflows
- Publicly published nvBowtie and BarraCUDA GPU containers
- Machine learning steps can be incorporated into Nextflow using GPU containers
- Opening up new workflow possibilities that rely heavily on GPU based tools

Future works

- Add support for each process being a different instance type
- Singularity still actively developing GPU features
- Improve the parameters for the SRAGPU-nf workflow
- Improve the machine learning models
- Validate with other RNA-seq data

Thank you for listening!

1. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), 316-319.
2. Benedicic, L., Cruz, F. A., Madonna, A., & Mariotti, K. (2017). Portable, high-performance containers for HPC. arXiv preprint arXiv:1704.03383.
3. Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. PloS one, 12(5), e0177459.
4. Oikkonen, L., & Lise, S. (2017). Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. Wellcome open research, 2.
5. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Ghemawat, S. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

Questions?

- BarraCUDA: shub://KevinSayers/BarraCUDA_Singularity
- nvBowtie: shub://KevinSayers/nvBowtie_Singularity
- SRAGPU-nf: https://github.com/KevinSayers/SRAGPU-nf
- OneHotVCF: https://github.com/KevinSayers/OneHotVCF
- Paper in the works!
- I presented portions of this work at the 2017 Bioinformatics Open Source Conference (BOSC) in Prague as a lightning talk.