Evening session: Introduction to Cloud and GPU computing

Author: Martin Morgan, Nitesh Turaga 21 June 2022

Cloud computing Why? What? How? – AnVIL Benefits Challenges Bioconductor resources (examples)

Intro to GPUs

What is a GPU? GPU for Machine learning GPGPU - General Purpose GPU Packages in Bioconductor using deep learning Using GPU on AnVIL

Cloud computing

Why?

- 1. Access more (CPUs, memory, storage) or different (e.g., GPUs) compute resources than easily available locally
- 2. Run workflows that apply the same steps (e.g., bulk RNAseq (pseudo)-alignment; scRNAseq preprocessing) across many different samples
- 3. Access 'consortium' data already stored in the cloud, perhaps requiring authentication, e.g., GTEx
- 4. Exploit novel cloud-based services, e.g., Google Bigtable (highly scalable database)

What?

Do-it-yourself

- Google, Amazon, or Azure (Microsoft) cloud resources
- (Virtually) unlimited access to compute power CPUs, memory, storage
- Highly flexible, but requires familiarity with cloud provider tools for managing resources

Pre-configured

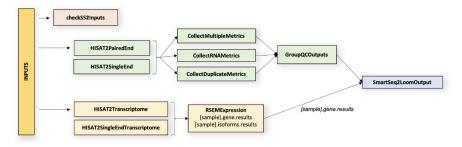
- E.g., In the US, Seven Bridges or NIH NHGRI AnVIL
- A 'higher level' interface requiring less technical knowledge of cloud provider tools for resource management
- Emphasizing particular use cases
 - Interactive analysis in Jupyter notebooks (Python, R) or RStudio

\leftarrow \rightarrow C $($ anvil.terra.bio/#workspaces/bioconductor-rpci-anvil/mtmorgan-bio	Û	☆ O Z	🖈 🖬 🌍 E	
Workspaces	> bioconductor-rpci-anvil/mtmorgan-bioc >			
DASHBOARD DATA ANALYSES WORKFLOWS	JOB HISTORY			:
R File Edit Code View Plots Session Build Debug Profile Tools Help ○ - ③ ④ - ○ □ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○)		rstudio 🔹	
helic.R × ODSCRIPTION × NAMESPACE × 1 # Generated by roxygen2: do not edit by hand 2 3 export(apodbyce)	Environment History Connections Build Cit DIff Commit - 1 -	Tutorial	Se main -	
4 export(hello) 5:18 NAMESPACE 0	Files Plots Packages Help Viewer	Delas Co	Rename 🏼 🎂 More	
Console Terminal × Jobs ×	New Polder V New Blank File V Upload V	Delete 🍝	Rename 🛛 🥨 More	191
(R 4.2.0 · ~/b/MetaPhOR/ ∞		Size	Modified	····
 checking 53 generic/method consistency (626ms) checking replacement functions checking foreign function calls checking R code for possible problems (2.9s) checking Ad files 			Jun 16, 2022, 7:25 Jun 21, 2022, 2:54	
 checking Rd metadata checking Rd line widths checking Rd cross-references 			Jun 21, 2022, 2:54 Jun 21, 2022, 2:54	

- Large-scale workflows described using Workflow Description Language
 - RNAseq pseudo-alignment (Kayla Interdonato)

```
task salmon_quant {
19
20
         File fastq1
         File fastq2
21
         File transcriptome_index
22
         String transcriptome_index_name
23
24
         String quant_out = basename(fastq1, ".fastq.gz")
25
26
         command {
27
28
             tar -xvzf ${transcriptome_index}
             salmon quant -i {transcriptome_index_name} -1 A \
29
             -1 ${fastq1} \
30
             -2 \{fastq2\} \setminus
31
             -p 8 --validateMappings -o ${quant_out}
32
             tar -cvzf ${quant_out}.tar.gz ${quant_out}
33
         }
34
35
         runtime {
36
             docker: "combinelab/salmon:1.3.0"
37
38
         }
```

- Human Cell Atlas <u>Smart-seq2 single-cell pipeline</u>



- 'Single sign on', including access to restricted data resources (if appropriate!)

How? - AnVIL

Google account + credit card

- Yes, cloud computing costs money, and a payment method has to be established!
- Not likely your credit card, but perhaps tied to an account associated with your institution & grant

Use of unrestricted data

- 'Workspace' defining a project
- 'Cloud environment' describing resources for interactive computation CPU,

RStudio Cloud Environment A cloud environment consists of applica disk(s).	ation configuratio	n, cloud compute and persistent	
Running cloud compute cost Paused clou \$0.25 per hr \$0.01	id compute cost per hr	Persistent disk cost \$2.00 per month	
Application configuration ①	5, Python 3.8.10)	~	
What's installed on this environment?		Updated: May 24, 2022 Version: 3.15.1	
Cloud compute profile			
CPUs 4 V Memory (GB)	26 V		
CPUs 4 V Memory (GB)			

memory, disk space

- 'Workflows' for large-scale computation
- 'Bucket' for storing static (e.g., FASTQ) data

Use of restricted data

- E.g., suppose dbGaP has provided me with access to GTEx data
- Link my AnVIL account with dbGaP
- My AnVIL account grants access to cloud-based resources associated GTEx
- Restrictions on use may be imposed, e.g., data must remain in AnVIL environment

Benefits

Flexibility

- Today I need a very large computer for my interactive analysis, but tomorrow I do not

- For this project, one step involves summarizing 100's of FASTQ files from a single cell experiment to count matrices; I want to do this summary quickly so that I can get to the more interesting biological questions in an interactive analysis
- For some data sets, restricted access is only available through AnVIL or other clouds

Independence

- To some extent, we are free from the constraints of our local system administrators and IT departments to get a bigger computer or submit a large job, we just start a larger instance or launch a workflow with appropriate resources.
- 'Someone else' (usually an expert) has taken care of technical things, e.g., the ability to install any Bioconductor package; configuration of salmon for pseudo-alignment.

Costs

- General purpose compute nodes (e.g., 4 cores, 32 MB memory) are not expensive, and are only billed while in use. Useful in workflows and interactive analysis
- Data movement (e.g., from the bucket where the data is stored to the disk of the compute node is inexpensive and very fast.

Challenges

New concepts

- Buckets for data storage
- Formal workflow description languages
- Understanding cloud provider resources, e.g., billing, security, ...

Costs

- Cloud providers are making a lot of money; as a corollary, we're paying a lot for cloud services!
- Cost control unintentional consumption of a large amount of resources, either accidentally (oops, I meant to submit 100 FASTQ files to a workflow, not the same file 100 times!) or because the resources are too easily accessible (yes, I want to do 500,000 simulations; no, I didn't realize that would cost \$60,000).
- Data ingress (uploading data to the cloud) is usually 'free', but data egress is costly vendor lock-in

Bioconductor resources (examples)

- AnVIL package for working with AnVIL & cloud resources
- <u>hca</u> and <u>cellxgenedp</u> packages for Human Cell Atlas data access (not really restricted to the cloud, but data access is particularly fast on the cloud)
- Rcwl for workflow management within R
- Bioconductor docker images for easy deployment to cloud environments

Intro to GPUs

What is a GPU?

- The graphics processing unit, or GPU, has become one of the most important types of computing technology to dramatically accelerate workloads in high-performance computing (HPC), deep learning, and more.

- Designed for parallel processing

- GPUs were originally designed to accelerate the rendering of 3D graphics

- While CPUs have continued to deliver performance increases through architectural innovations, faster clock speeds, and the addition of cores, GPUs are specifically designed to accelerate computer graphics workloads.

- GPUs are more programmable than ever before, affording them the flexibility to accelerate a broad range of applications that go well beyond traditional graphics rendering.

GPU for Machine learning

- Some of the most exciting applications for GPU technology involve AI and machine learning.

- GPUs incorporate an extraordinary amount of computational capability, they can deliver incredible acceleration in workloads that take advantage of the highly parallel nature of GPUs, such as image recognition.

- Many of today's deep learning technologies rely on GPUs working in conjunction with CPUs.

GPGPU - General Purpose GPU

Applications in Bioinformatics -

https://en.wikipedia.org/wiki/General-purpose_computing_on_graphics_processing_units#Bi oinformatics

Packages in Bioconductor using deep learning

- These packages can be sped up using GPUs

ttgsea - Tokenizing Text of Gene Set Enrichment Analysis

DeepPINCS - Protein Interactions and Networks with Compounds based on Sequences using Deep Learning

VAExprs - Generating Samples of Gene Expression Data with Variational Autoencoders

GenProSeq - Generating Protein Sequences with Deep Generative Models

Using GPU on AnVIL

- Choose the cloud environment with 'Enable GPU' option
- Choose the GPGPU type you'd like to use
 - Choose the number of GPUs
- Start the environment

ersistent disk(s). Running cloud compute cost Persistent disk cost \$2.18 per hr \$0.01 per hr Application configuration • R / Bioconductor: (Python 3.7.12, R 4.1.3, Bioconductor 3.14, tidyverse 1 v What's installed on this environment? Updated: May 31, 2022 What's installed on this environment? Updated: May 31, 2022 Cloud compute profile • CPUs 16 Memory (GB) 60 © Enable CPUs mer Learn more about CPU cost and restrictions. gr CPU type NVIDIA Tesla T4 VR UR Compute type Standard VM Image: Standard VM	Cloud Environment A cloud environment consists of application configuration, cloud compute and					
R / Bioconductor: (Python 3.7.12, R 4.1.3, Bioconductor 3.14, tidyverse 1 v What's installed on this environment? Updated: May 31, 2022 Version: 2.1.3 Cloud compute profile CPUs 16 Memory (CB) 60 CPUs environment Learn more about GPU cost and restrictions. If GPU type NVIDIA Tesla T4 CPUs 4 v Startup script URI Compute type Standard VM 30 minutes of inactivity Location environment in the series of the seri	Running cloud compute cost					
R / Bioconductor: (Python 3.7.12, R 4.1.3, Bioconductor 3.14, tidyverse 1 v What's installed on this environment? Updated: May 31, 2022 Version: 2.1.3 Cloud compute profile CPUs 16 Memory (CB) 60 CPUs environment Learn more about GPU cost and restrictions. If GPU type NVIDIA Tesla T4 CPUs 4 v Startup script URI Compute type Standard VM 30 minutes of inactivity Location environment in the series of the seri						
What's installed on this environment? Updated: May 3], 2022 Version: 2.1.3 Cloud compute profile CPUs 16 Memory (CB) 60 Enable GPUs Enable GPUs Memory (CB) GPU type NVIDIA Tesla T4 GPUs 4 Startup script URI Compute type Standard VM Image: Standard VM 30 minutes of inactivity Location ere Image: I	Application configuration	on 🕄				
Version: 2.1.3 Cloud compute profile CPUs 16 Memory (GB) 60 © Enable CPUs ere Learn more about CPU cost and restrictions. If GPU type NVIDIA Tesla T4 GPUs 4 Startup script URI Standard VM Standard VM Standard VM 30 minutes of inactivity Location ere 1	R / Bioconductor: (Pyth	on 3.7.12, R 4.1.3, Biocond	luctor 3.14, tidyverse 1 🗸			
CPUs 16 Vermory (GB) 60 V Enable GPUs era Learn more about GPU cost and restrictions. C GPU type NVIDIA Tesla T4 V GPUs 4 V Startup script URI Compute type Standard VM V Enable autopause Learn more about autopause. C 30 minutes of inactivity Location era 3	What's installed on this o	environment?				
 Enable GPUs ETA Learn more about GPU cost and restrictions. If GPU type NVIDIA Tesla T4 GPUs 4 Startup script URI Compute type Standard VM Enable autopause Learn more about autopause. If 30 minutes of inactivity Location ETA 3 	Cloud compute pro	file				
GPU type NVIDIA Tesla T4 CPUs 4 CPUs 4 CPUs Startup script URI URI Standard VM Compute type Standard VM Compute sclope Learn more about autopause. C 30 minutes of inactivity Location EFTA 3	CPUs 16 V	Memory (GB) 60	~			
Startup script URI Compute type Standard VM Standard V	Enable GPUs	Enable GPUs Learn more about GPU cost and restrictions.				
URI Compute type Standard VM Enable autopause Learn more about autopause. 30 minutes of inactivity Location BETA 3	GPU type NVIDIA Te	esla T4 🗸 🗸	GPUs 4			
Compute type Standard VM ✓ Carable autopause Learn more about autopause. 30 minutes of inactivity Location BETA 3	Startup script					
Standard VM ✓ Enable autopause Learn more about autopause. ☑ 30 minutes of inactivity Location BETA	URI	URI				
 ✓ Enable autopause Learn more about autopause. ☑ 30 minutes of inactivity Location BETA 3 	Compute type					
30 minutes of inactivity	Standard VM 🗸					
Location BETA 3	Enable autopause	Learn more about autopa	use. 🗗			
	30 minutes of	of inactivity				
Select 🗸						
	Select	Select 🗸				