An Investigation on Public Cloud Performance Variation for an RNA Sequencing Workflow

David Perez, Ling-Hung Hong, Sonia Xu,
Ka Yee Yeung, Wes Lloyd
daperez@uw.edu, wlloyd@uw.edu

September 21, 2020

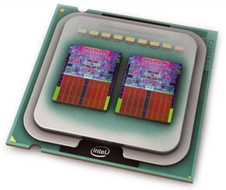
School of Engineering and Technology
University Of Washington, Tacoma

9th International Workshop on Parallel and Cloud-based Bioinformatics and Biomedicine (ParBio)
11th ACM Conference on Bioinformatics, Computational Biology and Health Informatics *(ACM BCB 2020)*

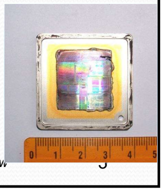
Outline

- Background
- Research Questions
- Use Case
- Experimental Implementation
- Experimental Results
- Summary
- Conclusions

CPU Heterogeneity

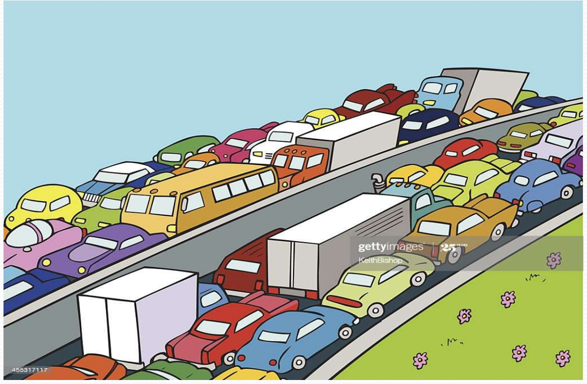


- Public cloud providers offer distinct **VM types** to simplify resource allocation to users
- **VM types:**
    - Have distinct configurations: (e.g. # of virtual CPUs (vCPUs), memory/storage capacity, and network bandwidth)



Resource Contention

- Resource Contention is when there is a competition over shared resources on a shared server



Research Questions

RQ1: What is the performance variation of running genomics data analytical tasks on the public cloud?

How much do factors such as provisioning variation, CPU heterogeneity, and resource contention contribute to performance variation?

RQ2: Over a 24-hour period, how does performance of individual cloud VMs vary for repeated runs of analytical tasks?

Provisioning Variation

- Provisioning variation is the random nature of VM placement across physical servers that occurs when cloud providers load balance VM launch requests.
- Where these VMs are hosted on public clouds is abstracted and not easily inferable in real time.

Outline

- Background
- Research Questions
- Use Case
- Experimental Implementation
- Experimental Results
- Summary
- Conclusions

Research Questions

RQ1: What is the performance variation of running genomics data analytical tasks on the public cloud?

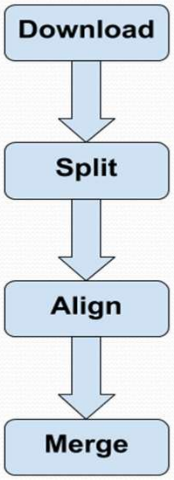
How much do factors such as provisioning variation, CPU heterogeneity, and resource contention contribute to performance variation?

RQ2: Over a 24-hour period, how does performance of individual cloud VMs vary for repeated runs of analytical tasks?

Outline

- Background
- Research Questions
- Use Case
- Experimental Implementation
- Experimental Results
- Summary
- Conclusions

Use Case: UMI RNA Sequencing Workflow (Xiong, Yuguang, et al)



Downloads input FASTQ files for the workflow using Amazon Simple Storage Service (Amazon S3).

Sequence tag in the first read and append it to the sequence identifier in the second read.

Align reads to the human reference sequence. In our case study, we use the BWA (Burrows Wheeler Aligner) [Li, Durbin 2009].

Filters outs duplicates and consolidates the transcript counts.

https://www.nature.com/articles/s41598-017-14892-x.pdf

Outline

- Background
- Research Questions
- Use Case
- Experimental Implementation
- Experimental Results
- Summary
- Conclusions

Container Profiler

The Container Profiler measures and records resource utilization of any containerized task capturing over 50+ Linux system metrics to characterize CPU, memory, disk, and network utilization at the VM, container, and process levels.

These metrics are important as they can help identify what system resources your workflow is consuming the most.

Controlling provisioning variation with AWS EC2 Placement Groups

- **Standard Placement:** No strategy – standard VM launch
- **Spread Placement:** Instances placed on distinct servers located on different server racks.
- **Cluster Placement:** Instances placed packed together inside an Availability Zone

AWS. 2020. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/placement-groups.html Last accessed July, 2020.

Experimental Setup

Using AWS EC2, we provisioned 30 x ec2 c5.2xlarge instances, 10 VMs for each placement strategy:

	Total Runs	Standard	Cluster	Spread
Intel 8124M	**16**	**4**	**4**	**8**
Intel 8275CL	**14**	**6**	**6**	**2**
AMD EPYC 7R32	**30**	**10**	**10**	**10**

C5.2xlarge/c5a.2xlarge CPU comparison

	Intel Xeon(R) Platinum 8124M CPU @ 3.00 GHZ	Intel Xeon(R) Platinum 8275CL @ 3.00 GHZ	AMD EPYC 7R32 CPU @ 2.80 GHZ
EC2 Instance Type	C5.2xlarge	C5.2xlarge	C5a.2xlarge
Family/microns/yr	Skylake/14nm/2017	Cascade Lake/14nm/2019	Rome/7 & 14nm/2019
Virtual CPU cores/host	72	96	96
Physical CPU cores/host	36	48	48
Burst clock MHz (Single/all)	3400/3500	3600/3900	3300/3400
L1 Cache (Per core)	32K (½ data, ½ instruction)	64k (½ data, ½ instruction)	64k (½ data, ½ instruction)
L2 Cache (Per core)	1024K	1024K	512K
L3 Cache (Per core)	25344K	36608K	16384K
Total Freq.	53%	47%	100%
Standard Freq.	13%	20%	100%
Cluster Freq.	13%	20%	100%
Spread Freq.	27%	7%	100%

https://en.wikipedia.org/wiki/List_of_Intel_Skylake-based_Xeon_microprocessors#Xeon_Platinum_8124M
https://en.wikipedia.org/wiki/List_of_Intel_Cascade_Lake-based_Xeon_microprocessors#Xeon_Platinum_8275CL

Outline

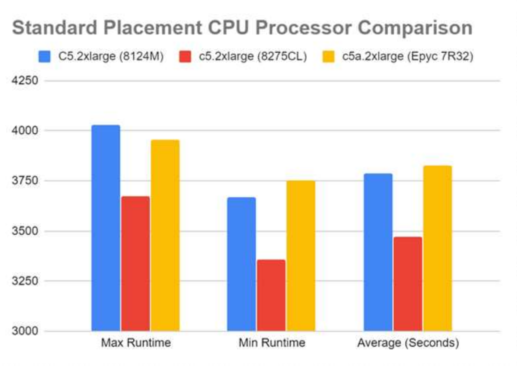
- Background
- Research Questions
- Use Case
- Experimental Implementation
- Experimental Results
- Summary
- Conclusions

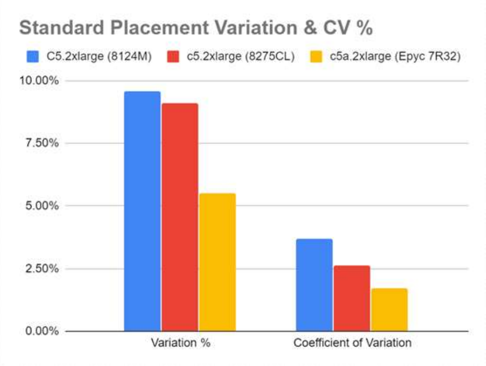
RQ-1: Performance Variation

- What is the performance variation of running genomics data analytical tasks on the public cloud?

Performance Variation: Standard Placement

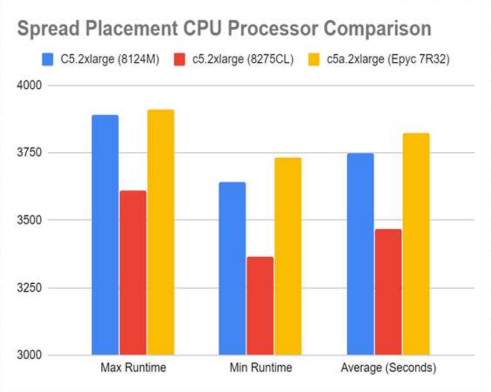
RNA-Seq Alignment runtime variation – c5a/c5.2xlarge

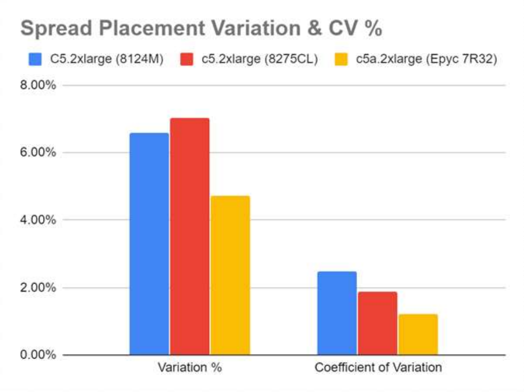




Performance Variation:
Spread Placement

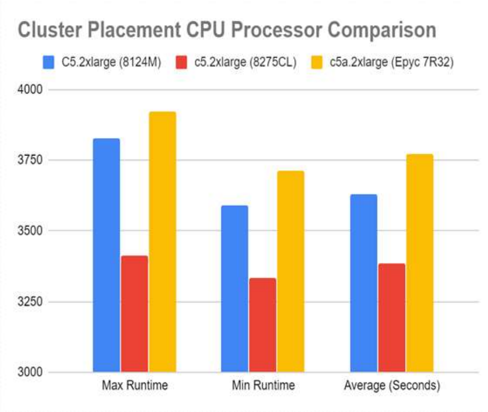
CPU runtime variation - c5.2xlarge, Spread placement:

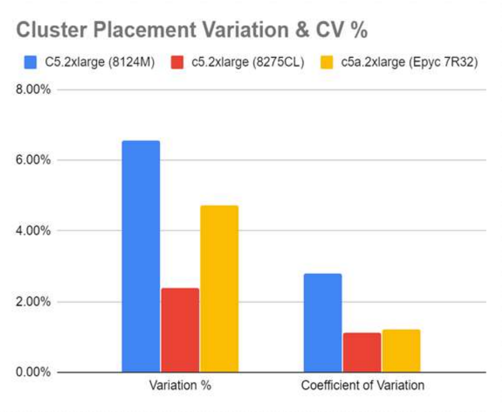




Performance Variation:
Cluster Placement

RNA-Seq Alignment runtime variation – c5a/c5.2xlarge





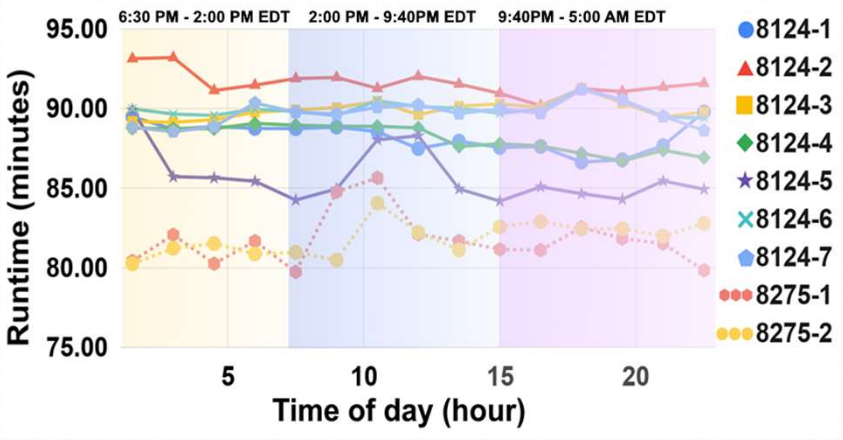
RQ-2: Inferring performance from resource utilization metrics

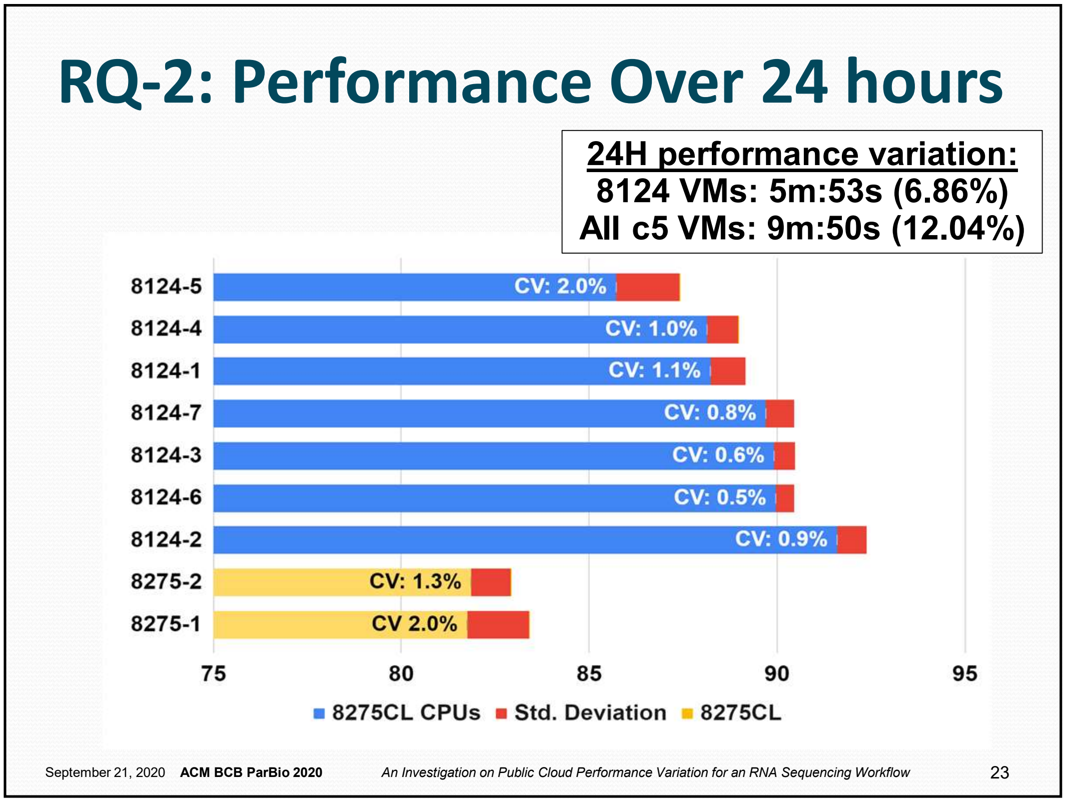
Over a 24-hour period, how does performance of individual cloud VMs vary for repeated runs of analytical tasks?

- What relationships exist between Linux resource utilization metrics (CPU, memory, disk, and network) and task runtimes?

RQ-2: Performance Over 24 hours

Performance of all 9 VMs randomly provisioned using c5.2xlarge instances over a 24-hour period (initial warmup run removed).

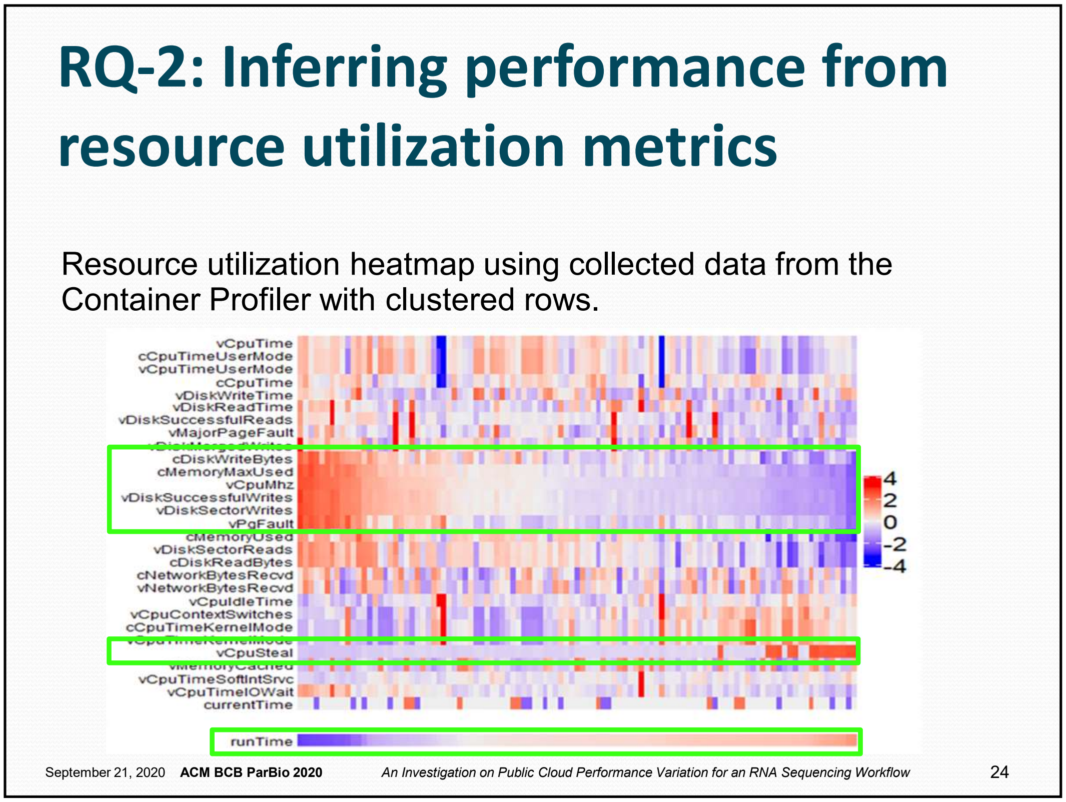




RQ-2: Performance Over 24 hours

<u>24H performance variation:</u>
8124 VMs: 5m:53s (6.86%)
All c5 VMs: 9m:50s (12.04%)

September 21, 2020 **ACM BCB ParBio 2020** *An Investigation on Public Cloud Performance Variation for an RNA Sequencing Workflow* 23



RQ-2: Inferring performance from resource utilization metrics

Resource utilization heatmap using collected data from the Container Profiler with clustered rows.

September 21, 2020 **ACM BCB ParBio 2020** *An Investigation on Public Cloud Performance Variation for an RNA Sequencing Workflow* 24

Outline

- Background
- Research Questions
- Use Case
- Experimental Implementation
- Experimental Results
- Summary
- Conclusions

Summary

- RQ-1 Performance variation:

  Performance variance of long running compute-bound tasks on were found to be as high as 20.04% (c5.2xlarge) and as low as 4.71% (c5a.2xlarge).

- RQ-2 Metric relationships with performance:

  A subset of metrics gathered by the Container profiler have been shown to exhibit a strong inverse relationship with runtime.

Outline

- Background
- Research Questions
- Use Case
- Experimental Implementation
- Experimental Results
- Summary
- Conclusions

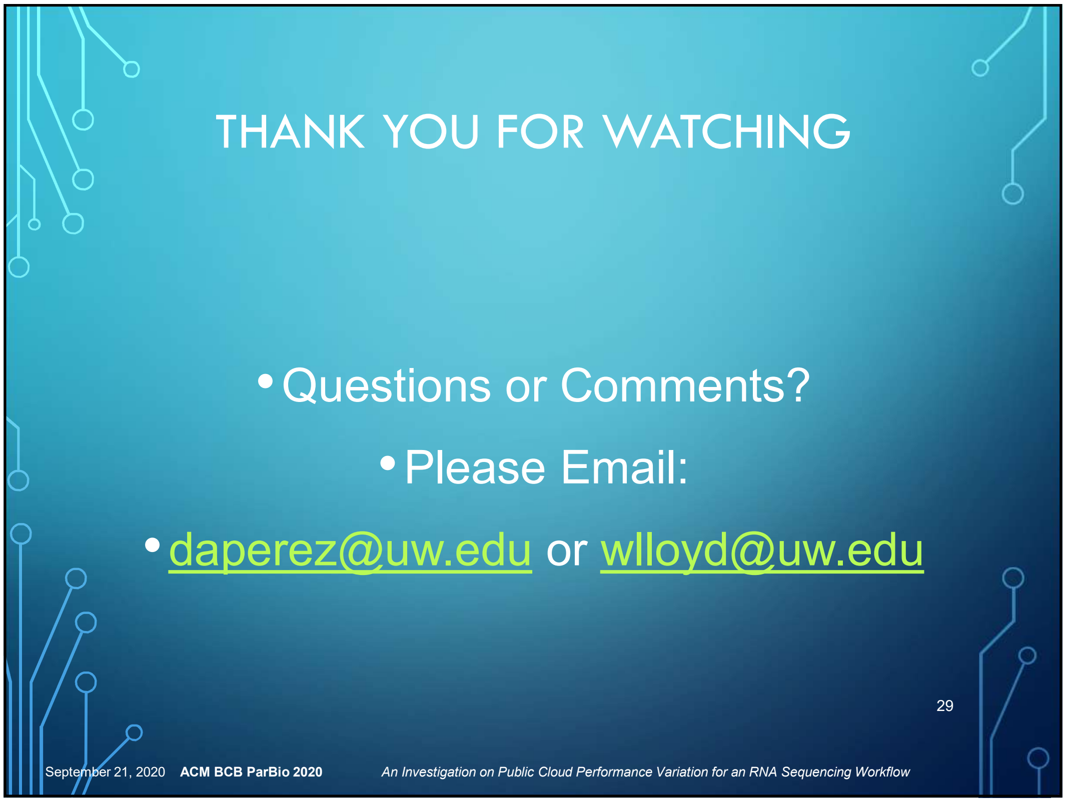
Conclusions

From RQ-1 we determined for RNA-sequence alignment that:

- Spread VM placement had the fastest runtime performance for both c5a/c5 VM types, with the least variation for c5a.2xlarge (AMD), and the most variation for c5.2xlarge (INTEL)
- Standard VM placement had the slowest runtime performance for both VM types, with the most variation for c5.2xlarge (INTEL)
- Cluster VM placement had “middle of the pack” runtime performance, with the least variation for c5a.2xlarge (AMD)

From RQ-2 we determined for our RNA-seq workflow that:

- Over 24 hours, c5 VMs (8124M CPU) had a performance spread of 5m:53s (6.86%), and for *all* c5 VMs the spread was 9m:50s (12.04%)
- cDiskWriteBytes, cMemoryMaxUsed, vCpuMhz, vDiskSuccessfulWrites, vDiskSectorWrites, vPgFaults have an inverse relationship to runtime



THANK YOU FOR WATCHING

- Questions or Comments?
- Please Email:
- daperez@uw.edu or wlloyd@uw.edu