Table of Contents	
10.1g: Large Language Models	2
10.1.1 LLMs and Retrieval Augmented Generation	2
10.1.2 Google Generative AI API	2
10.1.3	2
10.1.4 Retrieval Augmented Generation application	2
10.1.5 Document loading	2
10.1.6 Document searching	3
10.1.7 Document querying	5
10.1.8	5
10.1.9 LLM agents	6
10.1.10 Open API agent	6
10.1.11	6
10.2g: CDN	8
10.2.1 Part 1: Networks and VMs	8
10.2.2 Deployment specification	8
10.2.3 Network deployment specification	8
10.2.4 Subnetwork deployment specification	8
10.2.5 Virtual machine deployment specification	8
10.2.6 Deployment	8
10.2.7 Firewall deployment specification	10
10.2.8 Update deployment	11
10.2.9 Latency measurements	11
10.2.10 Part 2: Scaling via Instance Groups and Load Balancing	12
10.2.11 Firewall rule for HTTP	12
10.2.12 Instance templates	12
10.2.13 Health check	12
10.2.14 Managed Instance Group (europe-west1-mig)	12
10.2.15 Managed Instance Group (us-east5-mig)	12

10.2.16 Test groups	12
10.2.17 HTTP load balancer	14
10.2.18 HTTP load balancer	14
10.2.19 Test load balancer	14
10.2.20 Siege! (Part 1)	15
10.2.21 Siege! (Part 2)	17
10.2.22 Clean-up	17

10.1g: Large Language Models

10.1.1 LLMs and Retrieval Augmented Generation

10.1.2 Google Generative AI API

10.1.3 -

10.1.4 Retrieval Augmented Generation application

10.1.5 Document loading

• Explain what the transformer does to the HTML retrieved by the loader

The transformer (BeautifulSoupTransformer) processes the loaded HTML documents. It extracts specific tags (in this case, the <article> tag) to isolate relevant content from the HTML structure. This helps filter out non-relevant parts of the HTML such as navigation bars, sidebars, or advertisements.

• Examine the document cleaning code. What kinds of characters are removed when the text is cleaned?

The clean_text function performs the following operations:

- Converts non-ASCII characters to their closest ASCII equivalents using unidecode.
- Replaces multiple whitespace characters (like tabs, newlines, and spaces) with a single space.
- Strips leading and trailing whitespace. These steps ensure that the text is consistent and free of extraneous formatting.
- What size chunks is the content split into?

The content is split into chunks of size 10,000 characters each.

What amount of overlap is there between chunks?

There is an overlap of 1,000 characters between chunks. This overlap ensures continuity of context between chunks

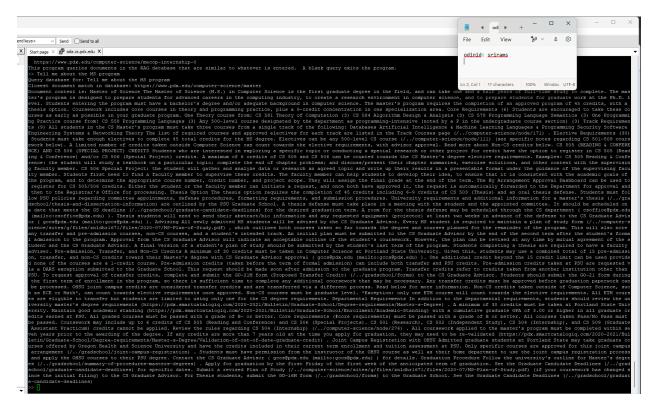
How many documents are loaded at a time into the vector database?

The function add_documents loads documents into the vector database in batches of 300 at a time. This batching process is controlled by the variable n set to 300 in the script.

10.1.6 Document searching

Show the document URLs that are returned for the following queries:

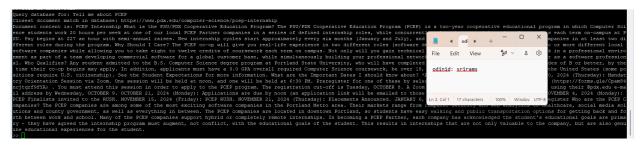
Tell me about the MS program



Tell me about the Cybersecurity certificate



Tell me about PCEP



Tell me about the capstone



10.1.7 Document querying

10.1.8 -

How many credits does it take to complete the MS program?

```
https://www.pdx.edu/computer-science/master
https://www.pdx.edu/computer-science/master
https://www.pdx.edu/computer-science/research-areas
https://www.pdx.edu/computer-science/research-areas
https://www.pdx.edu/computer-science/geaduate-internablp-credit
https://www.pdx.edu/computer-science/geaduate-internablp-credit
https://www.pdx.edu/computer-science/contact
lino> How many credits does take to complete the MS program 7

It takes 45 credits to complete the MS program in Computer Science at Portland State University. There is also a thesis option.

* https://www.pdx.edu/computer-science/master
* https://www.pdx.edu/computer-science/masters-track-courses
```

• How many core courses are there in the Cybersecurity certificate?



How many hours per week does a student work in PCEP?



How long is the capstone?



10.1.9 LLM agents

10.1.10 Open API agent

10.1.11 -

Take a screenshot showing the results of execution

What is the current xkcd?

```
Action Engors 1. OHT /Annth-0.jeem

> Encergage may AgentExector chain...

**Station : Grant John John

**Thought: I meed to fetch the current coast information using the '/info.d.jeen' empoint. I'll use 'requests_get' for this.

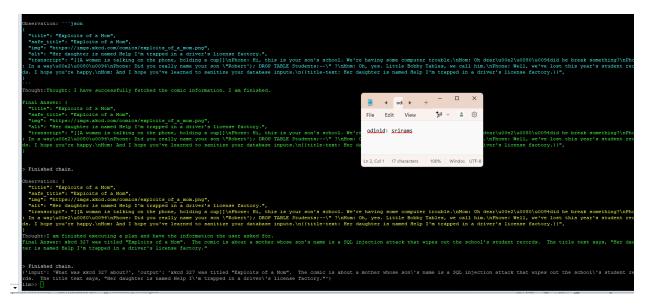
Action : requests_get

**subject : requests_get

*
```

What is the image link of the current xkcd?

What was xkcd 327 about?



- 10.2g: CDN
- 10.2.1 Part 1: Networks and VMs
- 10.2.2 Deployment specification
- 10.2.3 Network deployment specification
- 10.2.4 Subnetwork deployment specification
- 10.2.5 Virtual machine deployment specification

10.2.6 Deployment

• Take a screenshot of the output to include in your lab notebook. How many networks, subnetworks, and VM instances have been created?

```
Martine (constant): (notes to the deglosses as to bring pitterStort): dispray the constant of the deglosses as bring pitterStort): dispray the constant of the deglosses as bring pitterStort): dispray the constant of the deglosses as bring pitterStort): dispray the constant of the deglosses as bring pitterStort): dispray the constant of the degloss as the bring pitters of the constant of the degloss as the bring pitters of the constant of the degloss as the bring pitters of the constant of the degloss as the bring pitters of the constant of the degloss as the bring pitters of the constant of the degloss as the bring pitters of the constant of the degloss as the bring pitters of the constant of the degloss as the bring pitters of the degloss as the degloss
```

```
NAME: us-west-s1
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: us-west-s2
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: w1-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

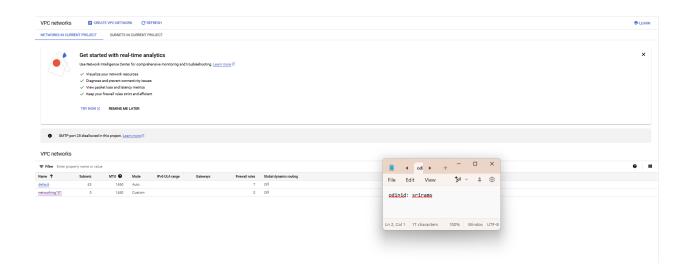
NAME: w2-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

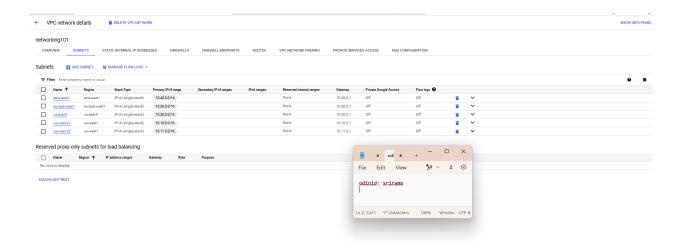
NAME: w2-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

STATE: COMPLETED
ERRORS: []
INTENT:
STATE: COMPLETED
ERRORS: []
INTENT:
STATE: COMPLETED
ERRORS: []
INTENT:
STATE: STATE: COMPLETED
ERRORS: []
INTENT: STATE: COMPLETED
ERRORS: []
ERRORS
```

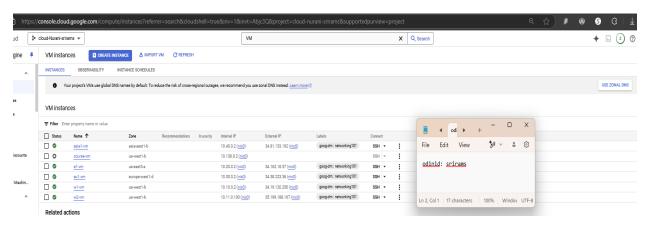
5-subnetworks 1-network 5-instances

 Visit the web console for VPC network and show the network and the subnetworks that have been created. Validate that it has created the infrastructure in the initial figure. Note the lack of firewall rules that have been created.





 Visit the web console for Compute Engine and show all VMs that have been created, their internal IP addresses and the subnetworks they have been instantiated on. Validate that it has created the infrastructure shown in the initial figure.



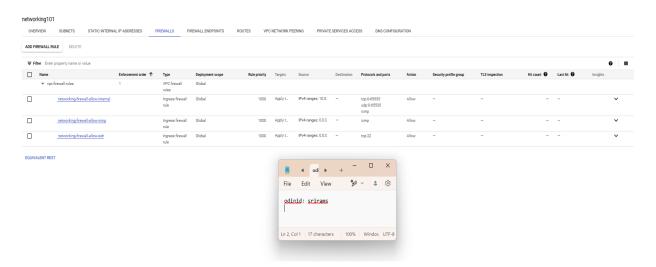
 Click on the ssh button for one of the VMs and attempt to connect. Did it succeed?

No it did not connect. ERROR: (gcloud.compute.ssh) [/usr/bin/ssh] exited with return code [255]

10.2.7 Firewall deployment specification

10.2.8 Update deployment

Take a screenshot that indicates the new rules have been deployed



10.2.9 Latency measurements

 Given this, fill in the table with the measured latencies between the 6 pairs and include it in your lab notebook. Use the shortest latency measured for each pair.

Location pair	Ideal latency	Measured latency
us-west1 us-east5	~45 <u>ms</u>	49.5 <u>ms</u>
us-west1 europe-west1	~93 <u>ms</u>	133 <u>ms</u>
us-west1 asia-east1	~114 <u>ms</u>	118.3 <u>ms</u>
us-east5 europe-west1	~76 <u>ms</u>	87.4 <u>ms</u>
us-east5 asia-east1	~141 <u>ms</u>	166.4 <u>ms</u>
europe-west1 asia-east1	~110 <u>ms</u>	251.5 ms

10.2.10 Part 2: Scaling via Instance Groups and Load Balancing

10.2.11 Firewall rule for HTTP

10.2.12 Instance templates

10.2.13 Health check

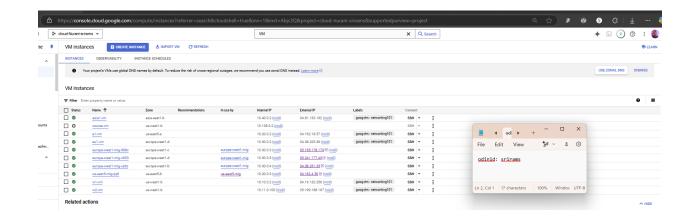
10.2.14 Managed Instance Group (europe-west1-mig)

10.2.15 Managed Instance Group (us-east5-mig)

10.2.16 Test groups

Are the instances in the same availability zone or in different ones?
 They are all in different zones

List all availability zones that your servers show up in for your lab notebook.





Networking 101 Lab

Client IP

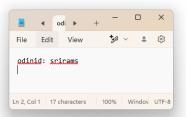
Your IP address: 131.252.218.119

Hostname

Server Hostname: us-east5-mig-rjq9

Server Location

Region and Zone: us-east5-b



← C **(\Lambda** Not secure | 35.195.178.173

Networking 101 Lab

Client IP

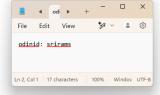
Your IP address: 131.252.218.119

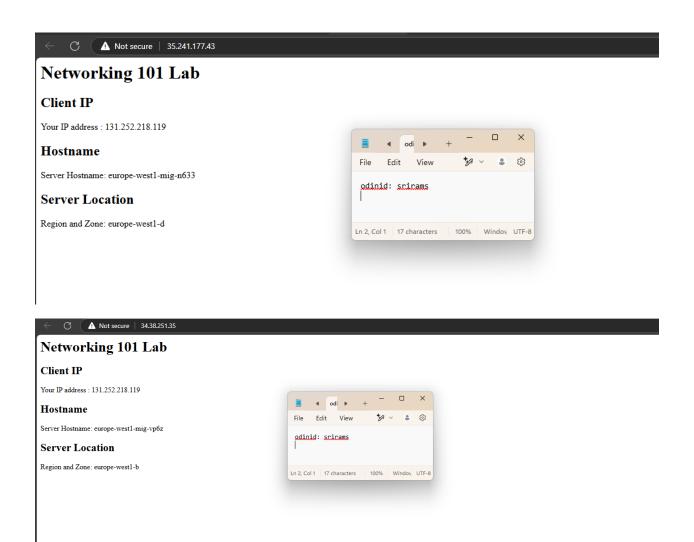
Hostname

Server Hostname: europe-west1-mig-988c

Server Location

Region and Zone: europe-west1-c

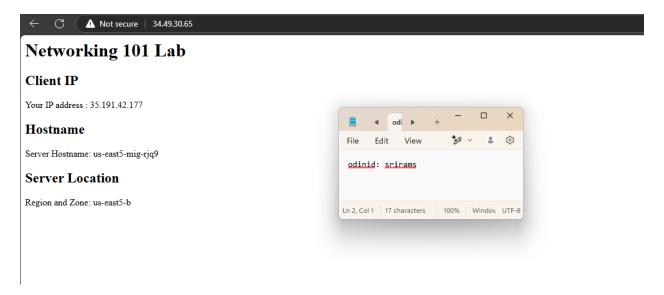




us-east5-b, europe-west1-c, europe-west1-d, europe-west1-b

- 10.2.17 HTTP load balancer
- 10.2.18 HTTP load balancer
- 10.2.19 Test load balancer

Show a screenshot of the page that is returned.

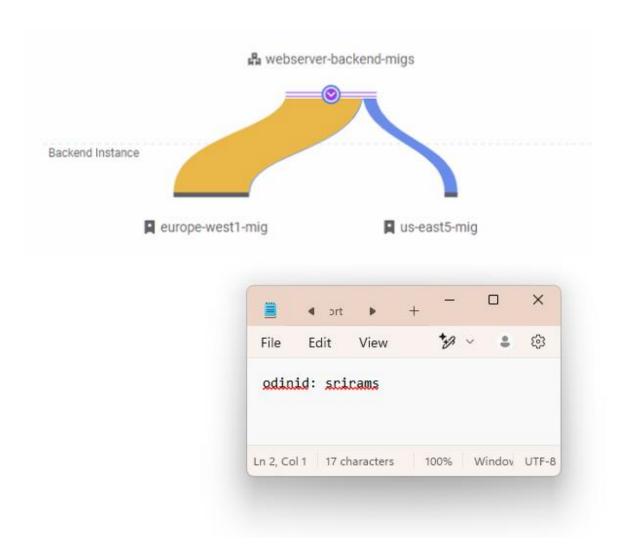


• Which availability zone does the server handling your request reside in?

us-east5-b

10.2.20 Siege! (Part 1)

• Take a screenshot of the initial traffic distribution

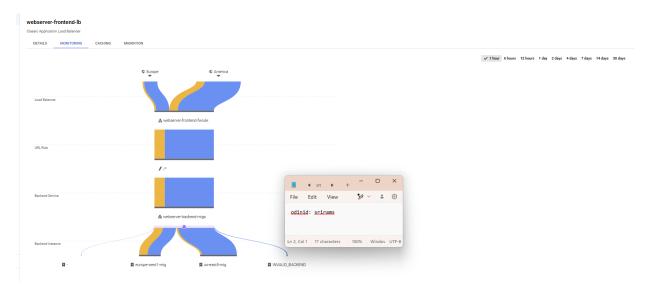


• Take a screenshot of the UI as additional instances are brought up and show that the traffic distribution shifts



10.2.21 Siege! (Part 2)

Show a screenshot of the final traffic distribution.



10.2.22 Clean-up