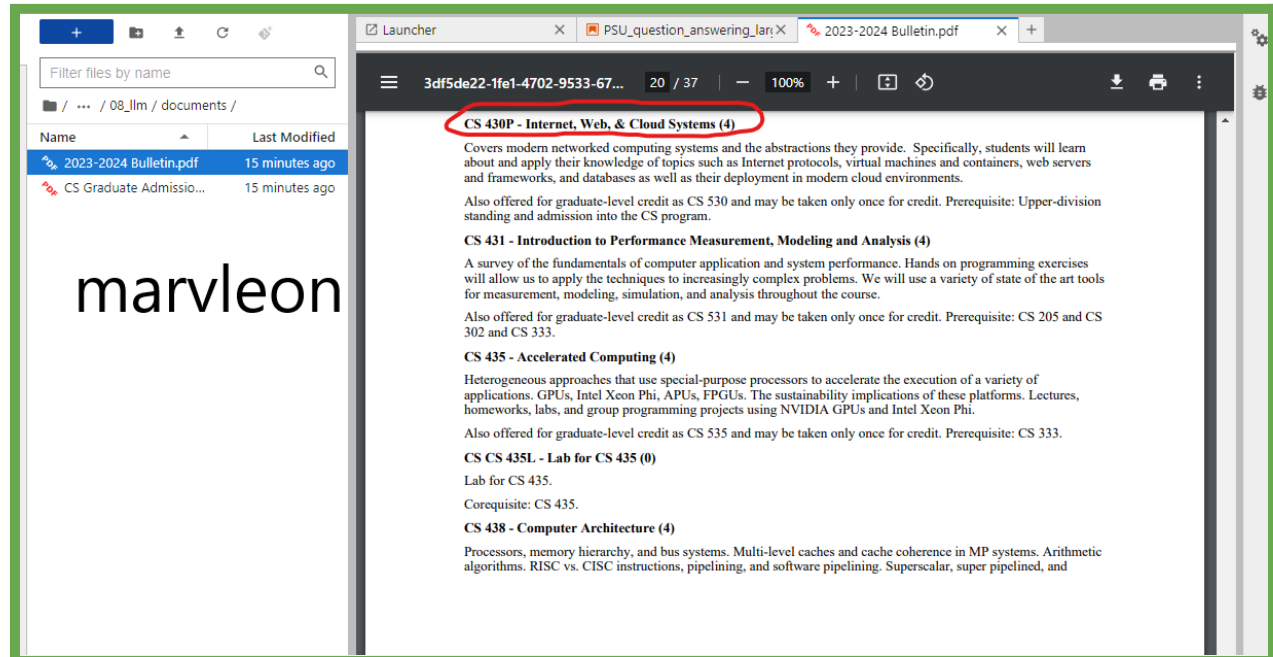


# Lab Week 10 — marvleon

<b>10.1g: LLMs</b>	<b>1</b>
10.1.4 Walk though notebook	1
10.1.5 Final questions	4
<b>10.2g CDN</b>	<b>5</b>
10.2.6 Deployment	5
10.2.9 Latency measurements	7
10.2.9 Latency measurements	7
10.2.16 Test Groups	7
10.2.19 Test load balancer	8
10.2.20 Siege (Part 1)	8
10.2.21 Siege! (Part 2)	9

## 10.1g: LLMs

### 10.1.4 Walk though notebook



```
[9]: try:
      print("PaLM Predicted:", generation_model.predict(prompt).text)
    except Exception as e:
      print(
        "The code failed since it won't be able to run inference on such a huge context and throws this exception: ",
        e,
      )
```

The code failed since it won't be able to run inference on such a huge context and throws this exception: 400 The request cannot be processed. The most likely reason is that the provided input exceeded the model's input token limit.

- Take a screenshot that includes your OdinID showing the error that is returned for your lab notebook

- **Provide an explanation as to why the description is not returned for your lab notebook**
  - The description is not returned because the answer/result was not within the context.

```
[17]: question = "What is the course description for CS 530?"

import time
t0 = time.time()
pdf_data_sample["predicted_answer"] = pdf_data_sample.app
      get_answer, axis=1
    )
t1 = time.time()

print(f"Time elapsed {(t1-t0)}")
pdf_data_sample
```

Time elapsed 15.20050573348999

[17]:	file_name	file_type	page_number	conten
-------	-----------	-----------	-------------	--------

PORTLAND STA

- **How many chunks returned predictions?**
  - 5

```
print('PaLM Predicted: ', generation_model.predict(prompt).text)
```

the prompt: Answer the question as precise as possible using the provided context. If the answer is not contained in the context, say "answer not available in context"

Context:

['Internet, Web, Cloud Systems', 'Internet, Web, Cloud Systems', 'Covers modern networked computing systems and the abstractions they provide Specifically, students will learn about and apply their knowledge of topics such as Internet protocols, virtual machines and containers, web servers and frameworks, and databases as well as their deployment in modern cloud environments', 'Covers modern networked computing systems and the abstractions they provide Specifically, students will learn about and apply their knowledge of topics such as Internet protocols, virtual machines and containers, web servers and frameworks, and databases as well as their deployment in modern cloud environments Also offered for graduate -level credit as CS 430P and may be taken only once for credit Prerequisite: Graduate - standing and admission into CS program', 'Advanced software design patterns using Java as the presentation language Course is suitable to software architects and developers who are already well -versed in this language In addition, it offers continuous opportunities for learning the most advanced features of the Java language and understanding some principles behind the design of its fundamental libraries Also offered as CS 653 and may be taken only once for credit Prerequisite: programming in Java and CS 520']?

Question:

What is the course description for CS 530?

Answer:

the number of words in the prompt: 1623

PaLM Predicted: Covers modern networked computing systems and the abstractions they provide Specifically, students will learn about and apply their knowledge of topics such as Internet protocols, virtual machines and containers, web servers and frameworks, and databases as well as their deployment in modern cloud environments Also offered for graduate -level credit as CS 430P and may be taken only once for credit Prerequisite: Graduate - standing and admission into CS program

- Take a screenshot that includes your OdinID showing the result that is returned for your lab notebook

There are several pros and cons to this method of performing this task compared to Stuffify:

Then, run the cells below to get responses to common questions we get in Computer Science.

```
[29]: print(answer_my_question("Are international students eligible for grad prep?"))
```

Yes, international students are eligible for the postbaccalaureate Grad Prep program and can receive a 1-20 for the program.

```
[30]: question("If my undergraduate GPA is below 3.0, will it be possible to be admitted to the MS program?")
```

It is possible for an applicant to be recommended for admission whose undergraduate GPA is slightly below 3.0 if their overall application is very strong and the admissions committee determines that the applicant is a good fit for the program. It is recommended that an applicant's low GPA be addressed in their Statement of Purpose within their application.

```
[31]: print(answer_my_question("What are the requirements for the masters cybersecurity certificate?"))
```

The cybersecurity certificate program requires admission as a graduate student, similar to admission to the Master's program, in the Computer Science department. The program requires 21 total credits of graduate classes. There are two core classes for a total of 6 credits. In addition, five elective classes must be taken for the needed additional 15 credits. In summary, seven total graduate classes must be taken two are core and five are electives.

```
[32]: print(answer_my_question("What are the requirements for admission to the Computer Science major?"))
```

1. Completion of each of the following core CS courses with a C or better: CS 161 Introduction to Programming and Problem Solving 4
2. Completion of each of the following non-CS courses with a grade of C- or better: MTH 251 Calculus I MTH 252 Calculus II or MTH 261 Linear Algebra Three Approved Laboratory Science courses
3. Prior to admission, PSU students are expected to complete the Freshman and Sophomore Inquiry series. Similarly, transfer students are expected to complete the Maseeh College lower division general education requirements. Completing the general

```
[33]: print(answer_my_question("What are the requirements for admission to the Computer Science major?"))
```

1. Completion of each of the following core CS courses with a C or better: CS 161 Introduction to Programming and Problem Solving 4
2. Completion of each of the following non-CS courses with a grade of C- or better: MTH 251 Calculus I MTH 252 Calculus II or MTH 261 Linear Algebra Three Approved Laboratory Science courses
3. Prior to admission, PSU students are expected to complete the Freshman and Sophomore Inquiry series. Similarly, transfer students are expected to complete the Maseeh College lower division general education requirements. Completing the general

- Take a screenshot including your OdinID that shows the results of the queries

### 10.1.5 Final questions

- ***Which of the approaches described would have issues with token limits on LLMs?***
  - Definitely the method of Stuffing!
- ***Which of the approaches would result in the most queries for the LLM to handle? How many LLM requests are performed from a single user query in this approach?***
  - Map reduce would probably result in the most queries.
  - 41 requests
- ***Which of the approaches requires one to search a vector database for an appropriate context that is then sent to the LLM?***
  - Map Reduce Embedding

# 10.2g CDN

## 10.2.6 Deployment

```
marvleon@cloudshell:~/code/networking101 (cloud-leon-marvleon)$ gcloud deployment-manager deployments create networking101 --config networking-lab.yaml
The fingerprint of the deployment is b'icFjss0npgvBRVedfSUvxQ=='
Waiting for create [operation-1701727482666-60bb64e1bdc8e-a97e7f14-829d72e3]...done.
Create operation operation-1701727482666-60bb64e1bdc8e-a97e7f14-829d72e3 completed successfully.
NAME: asia-east1
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: asia1-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: e1-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

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

NAME: europe-west1
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: networking101
TYPE: compute.v1.network
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: us-east5
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

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:
marvleon@cloudshell:~/code/networking101 (cloud-leon-marvleon)$
```

- **How many networks, subnetworks and VM instances have been created?**
  - 11 (5 subnetworks, 5 instances, 1 network)

cloud-Leon-marvleon
vpc network
Search

VPC networks
CREATE VPC NETWORK
REFRESH

NETWORKS IN CURRENT PROJECT
SUBNETS IN CURRENT PROJECT

Select the VPC networks for which you want to view subnets. If no networks are selected, the table shows the subnets in the current project.

VPC networks

### Subnets

Filter Enter property name or value

Name	Region	VPC network	Internal IP ranges	External IP ranges	Secondary IP
<a href="#">asia-east1</a>	asia-east1	<a href="#">networking101</a>	10.40.0.0/16	None	None
<a href="#">europe-west1</a>	europe-west1	<a href="#">networking101</a>	10.30.0.0/16	None	None
<a href="#">us-east5</a>	us-east5	<a href="#">networking101</a>	10.20.0.0/16	None	None
<a href="#">us-west-s1</a>	us-west1	<a href="#">networking101</a>	10.10.0.0/16	None	None
<a href="#">us-west-s2</a>	us-west1	<a href="#">networking101</a>	10.11.0.0/16	None	None

- **Did it succeed?**
  - No, stuck on establishing connection to SSH server...

cloud-Leon-marvleon
vpc network
Search

VM instances
CREATE INSTANCE
IMPORT VM
REFRESH
LEARN

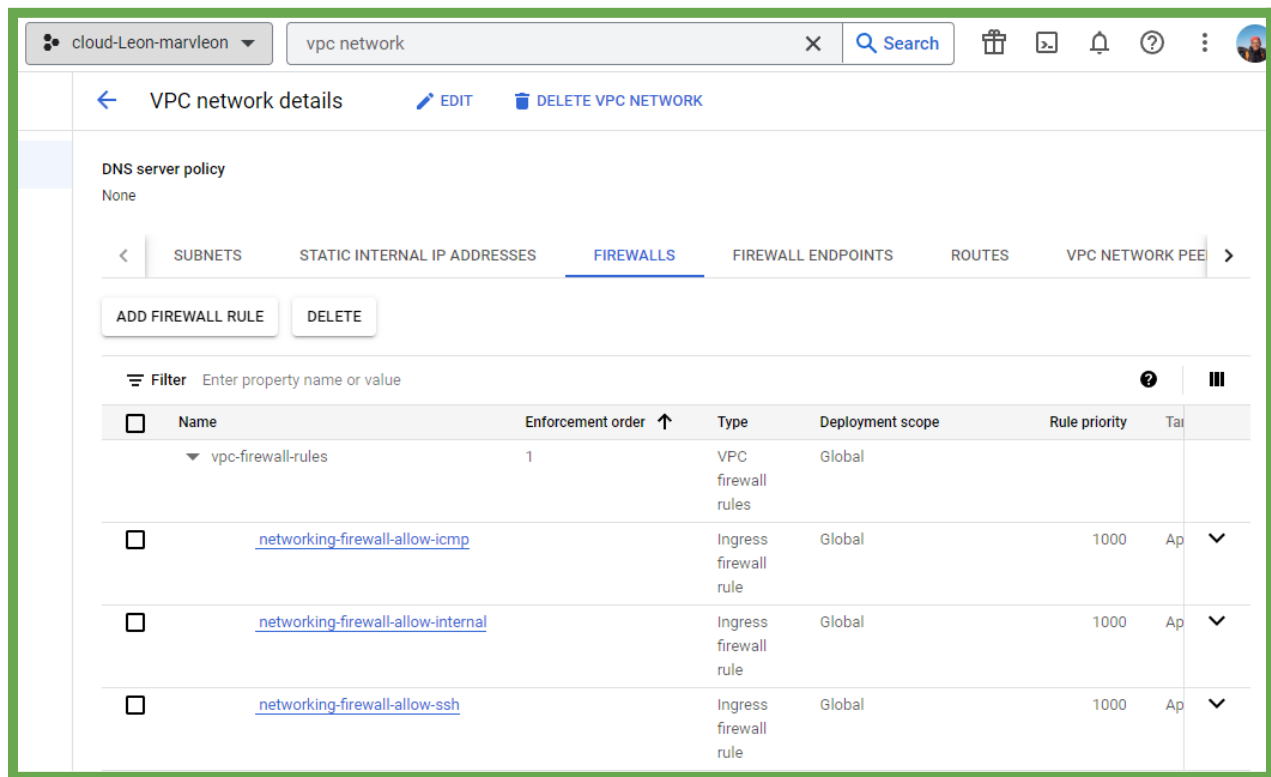
INSTANCES
OBSERVABILITY
INSTANCE SCHEDULES

### VM instances

Filter Enter property name or value

<input type="checkbox"/>	Status	Name	Zone	Internal IP	External IP	Network	Connect
<input type="checkbox"/>	✓	<a href="#">asia1-vm</a>	asia-east1-b	10.40.0.2 ( <a href="#">nic0</a> )	35.229.221.16 ( <a href="#">nic0</a> )	<a href="#">networking101</a>	SSH
<input type="checkbox"/>	✓	<a href="#">e1-vm</a>	us-east5-a	10.20.0.2 ( <a href="#">nic0</a> )	34.162.89.132 ( <a href="#">nic0</a> )	<a href="#">networking101</a>	SSH
<input type="checkbox"/>	✓	<a href="#">eu1-vm</a>	europe-west1-d	10.30.0.2 ( <a href="#">nic0</a> )	34.140.201.139 ( <a href="#">nic0</a> )	<a href="#">networking101</a>	SSH
<input type="checkbox"/>	✓	<a href="#">w1-vm</a>	us-west1-b	10.10.0.2 ( <a href="#">nic0</a> )	35.233.254.128 ( <a href="#">nic0</a> )	<a href="#">networking101</a>	SSH
<input type="checkbox"/>	✓	<a href="#">w2-vm</a>	us-west1-b	10.11.0.100 ( <a href="#">nic0</a> )	35.230.85.176 ( <a href="#">nic0</a> )	<a href="#">networking101</a>	SSH

## 10.2.9 Latency measurements



## 10.2.9 Latency measurements

Location pair	ideal latency	measured latency
us-west1 us-east5	~45 ms	49 ms
us-west1 europe-west1	~93 ms	133 ms
us-west1 asia-east1	~114 ms	116 ms
us-east5 europe-west1	~76 ms	88 ms
us-east5 asia-east1	~141 ms	174 ms
europe-west1 asia-east1	~110 ms	249 ms

## 10.2.16 Test Groups

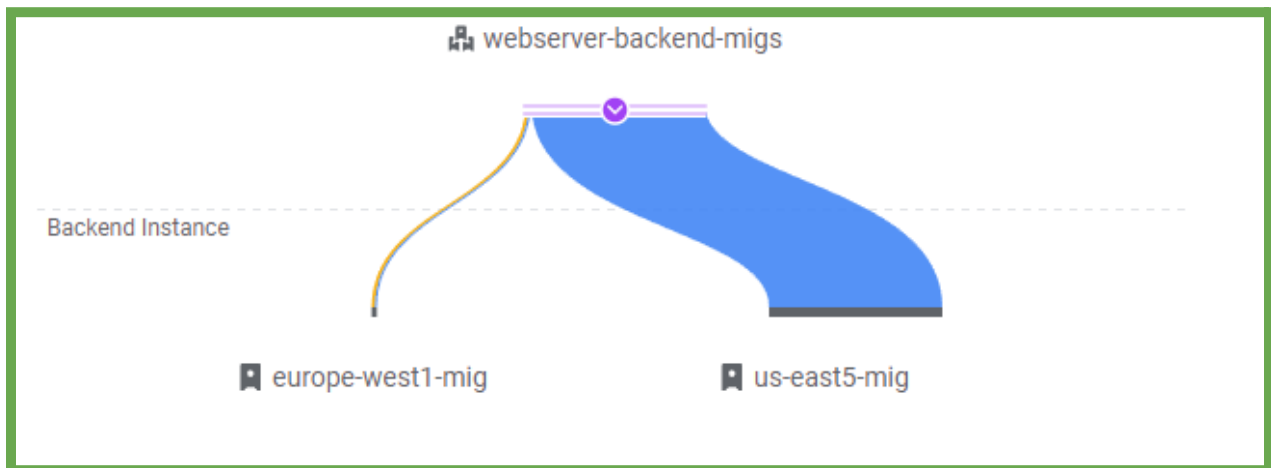
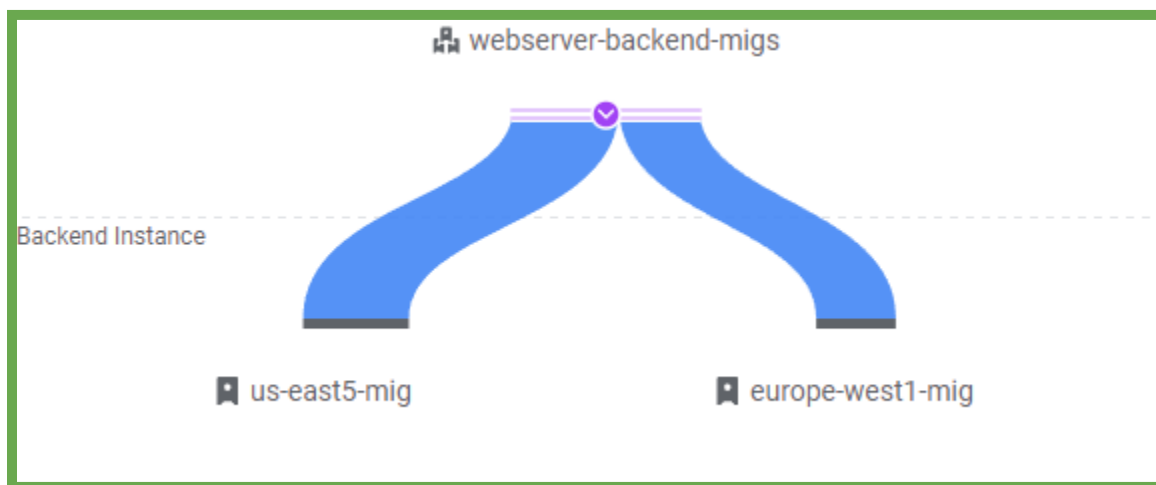
- **Are the instances in the same availability zone or in different ones?**
  - Different

- **List all availability zones that your servers show up in for your lab notebook.**
  - europe-west1-c
  - europe-west1-d
  - europe-west1-b
  - us-east5-b

### 10.2.19 Test load balancer

- **Which availability zone does the server handling your request reside in?**
  - us-east5-b

### 10.2.20 Siege (Part 1)





## 10.2.21 Siege! (Part 2)

