

Lab 10

Stuart Rimel

CS530: Internet Web & Cloud, Fall 2023

Odin: srimel

** In all the terminal screenshots my Odin name is in the terminal prompt **

Table of contents:

[10.1g: LLMs](#)

[10.1g.4: Walk through notebook](#)

[10.1g.5: Final questions and clean-up](#)

[10.2g: CDN](#)

[10.2g.6: Deployment](#)

[10.2g.8: Update deployment](#)

[10.2g.9: Latency measurements](#)

[10.2g.16: Test groups](#)

[10.2g.19: Test load balancer](#)

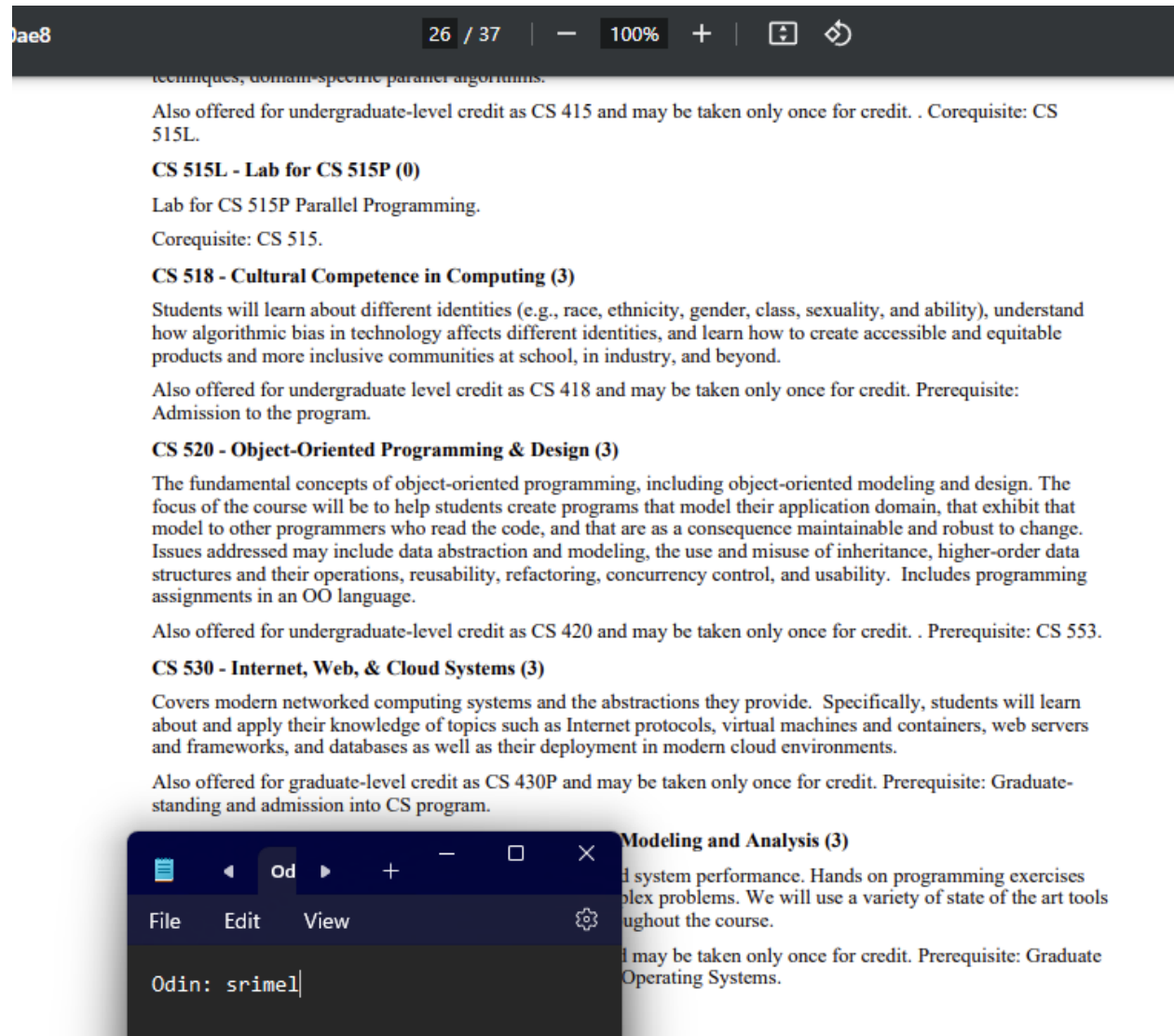
[10.2g.20: Siege! \(Part 1\)](#)

[10.2g.21: Siege! \(Part 2\)](#)

10.1g: LLMs

10.1g.4: Walk through notebook

Document loader:



Context error:

```
[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

If you restrict the context to first 5000 words or something which is less than the token limit for the PaLM API, the call will succeed, but only if the result is within the context. Attempt to do so.

[ ]: as possible using the provided context. If the answer is
      say "answer not available in context" \n\n

Odin: srime1
```

Description not returned reason: I think it's because the information we want is further down in the document than what was loaded into the context we passed to PaLM.

Time elapsed on running prediction across all chunks:

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

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

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

Time elapsed 14.566023826599121

[17]: type page_number
      .pdf 1 PC
      .pdf 2 166 P

Odin: srime1
```

How many chunks returned predictions?

- A. 5 chunks returned predictions

Prediction on map reduced context:

```
[19]: prompt = f"""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" \n\n
      Context: \n {context_map_reduce}? \n
      Question: \n {question} \n
      Answer:
      """

print("the prompt: ", prompt)
print("the number of words in the prompt: ", len(prompt))

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
```

Queries run with map reduce with embeddings:

```
[24]: 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 an I-20 for the program.

[25]: print(answer_my_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.

[26]: 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.

[27]: 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

[28]: 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
```

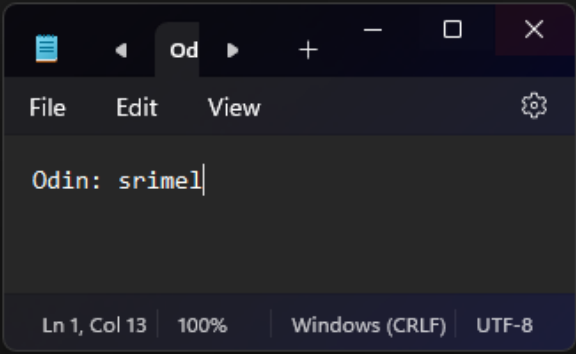
10.1g.5: Final questions and clean-up

1. Stuffing has issues with token limits since it sends all data at once
2. Map-reduce results in most queries for LLM to handle
3. Map-reduce with embedding uses vector db to find similar chunks

10.2g: CDN

10.2g.6: Deployment

```
Create operation operation-1700375119466-60a7b6efb106d-5ac2f466-702d9df7 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: eu1-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:

srinel@cloudshell:~/networking101 (cloud-rimel-srimel)$

```

```

# networks = 1
# subnetworks = 5
# instances = 5

```

VPC Network

Subnets
+ ADD SUBNET
≡ FLOW LOGS

Filter
Enter property name or value

<input type="checkbox"/>	Name ↑	Region	Stack Type	Internal IP ranges	External IP ranges	Secondary IPv4 ranges	Gateway	Private Google
<input type="checkbox"/>	asia-east1	asia-east1	IPv4	10.40.0.0/16	None	None	10.40.0.1	Off
<input type="checkbox"/>	europe-west1	europe-west1	IPv4	10.30.0.0/16	None	None	10.30.0.1	Off
<input type="checkbox"/>	us-east5	us-east5	IPv4	10.20.0.0/16	None	None	10.20.0.1	Off
<input type="checkbox"/>	us-west-s1	us-west1	IPv4	10.10.0.0/16	None	None	10.10.0.1	Off
<input type="checkbox"/>	us-west-s2	us-west1	IPv4	10.11.0.0/16	None	None	10.11.0.1	Off

Od
+
-
□
×

File
Edit
View
⚙

Odin: srinel

balancing

ranges	Gateway	Role	Purpose
--------	---------	------	---------

Compute Engine:

VM instances

Filter Enter property name or value

<input type="checkbox"/>	Status	Name ↑	Zone	In use by	Internal IP	External IP	Network
<input type="checkbox"/>	✓	asia1-vm	asia-east1-b		10.40.0.2 (nic0)	34.80.181.7 (nic0)	networking101
<input type="checkbox"/>	○	course-vm	us-west1-b		10.138.0.2 (nic0)		default
<input type="checkbox"/>	○	course-vm-image-1	us-west1-a		10.138.0.17 (nic0)		default
<input type="checkbox"/>	✓	e1-vm	us-east5-a		10.20.0.2 (nic0)	34.162.46.115 (nic0)	networking101
<input type="checkbox"/>	✓	eu1-vm	europa-west1-d		10.30.0.2 (nic0)	34.140.137.104 (nic0)	networking101
<input type="checkbox"/>	✓	w1-vm	us-west1-b		10.10.0.2 (nic0)	35.199.151.250 (nic0)	networking101
<input type="checkbox"/>	✓	w2-vm	us-west1-b		10.11.0.100 (nic0)	35.185.249.3 (nic0)	networking101

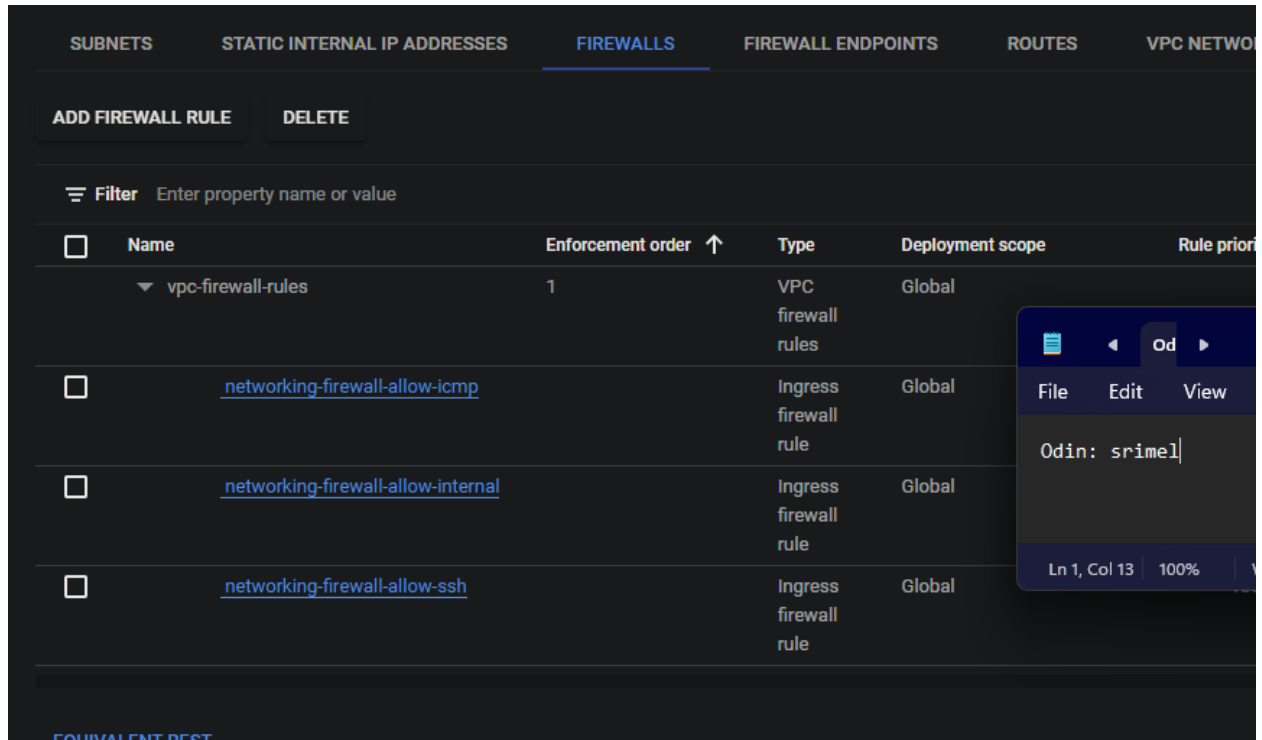
Related links

- [View billing report](#)
- [Monitor VMs](#)
View outlier VMs across metrics like CPU and network
- [View, stop, or delete instances](#)

Odin: srime1

When I try to connect via the 'ssh' button on any of the VMs I just get an endlessly spinning icon that says "Establishing connection to SSH server". It does not seem like it will go through.

10.2g.8: Update deployment



10.2g.9: Latency measurements

Us-west1 us-east5	45	49.8
Us-west1 europe-west1	93	134
Us-west1 asia-east1	114	119
Us-east5 europe-west1	76	91.7
Us-east5 asia-east1	141	166
Europe-west1 asia-east1	110	247

10.2g.16: Test groups

1. Different availability zones, europe-west1-c and us-east5-a
2. Europe-west1-d, europe-west1-c, europe-west1-b, us-east5-a

10.2g.19: Test load balancer

webserver-frontend-lb
Classic Application Load Balancer

Faster web performance and

DETAILS MONITORING

Frontend

Protocol ↑	IP:Port
HTTP	34.36.240.52:80

Host and path rules

Hosts ↑	Paths
All unmatched (default)	All unmatched

Backend

Backend services

1. webserver-backend-migs

Networking 101 Lab

Client IP

Your IP address : 35.191.20.184

Hostname

Server Hostname: us-east5-mig-7h9v

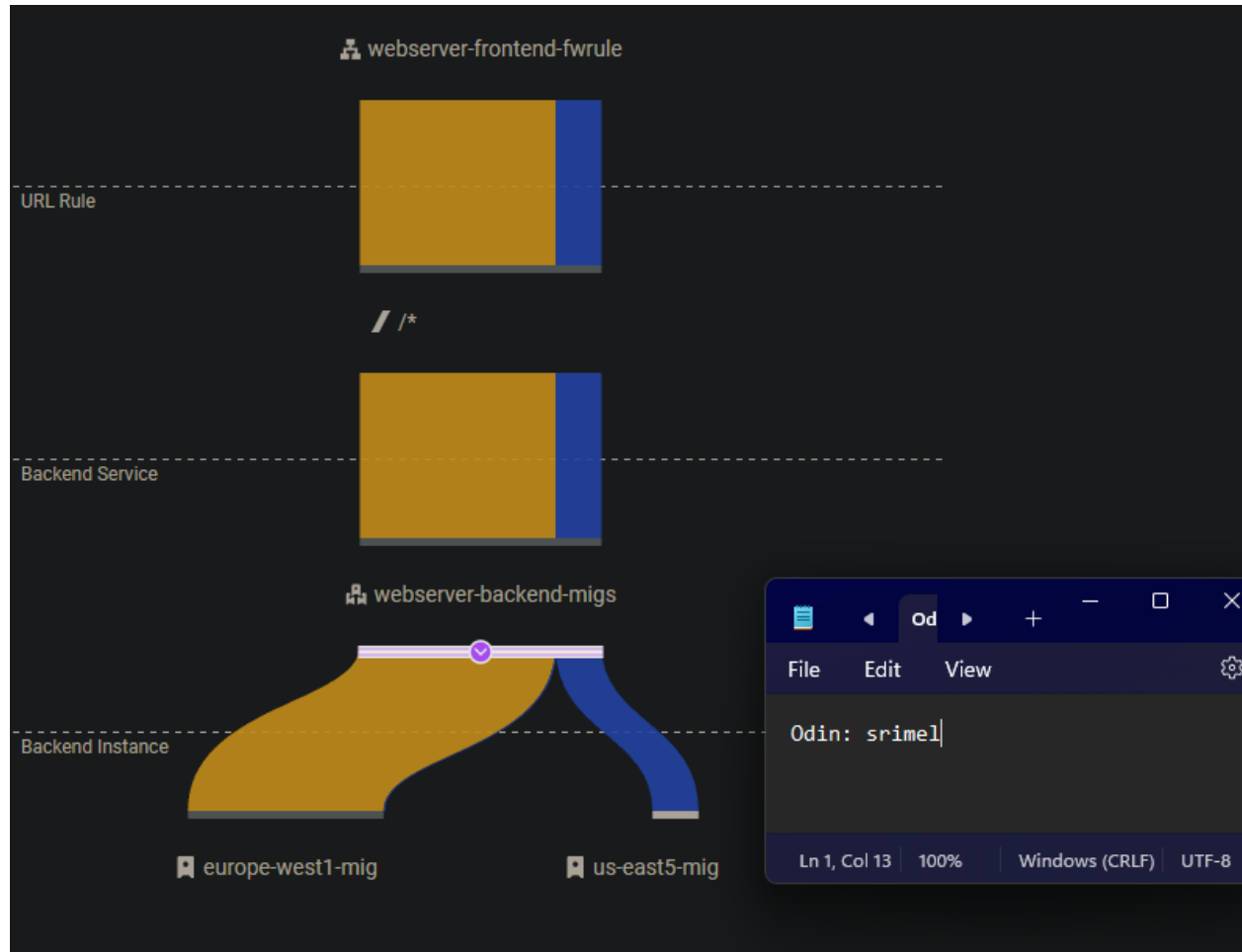
Server Location

Region and Zone: us-east5-a

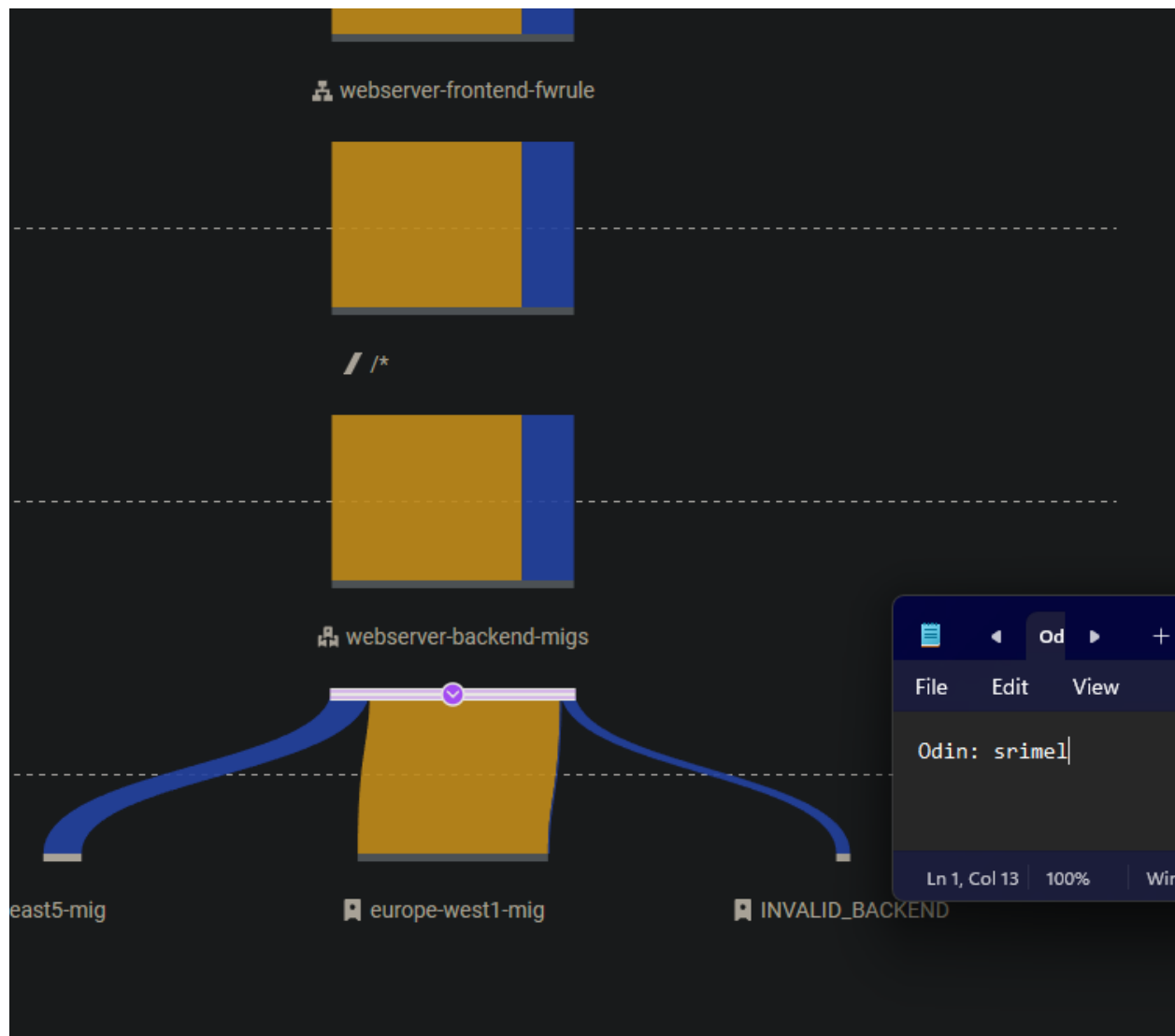
Odin: srime1

Availability zone = us-east5-a

10.2g.20: Siege! (Part 1)



Siege completed, and this is what I got. Not sure if I didn't refresh the monitoring page enough and missed it? Google Cloud is extremely slow right now....



I'm scared to run it again for the sake of my credits...

10.2g.21: Siege! (Part 2)

These were my final traffic results

