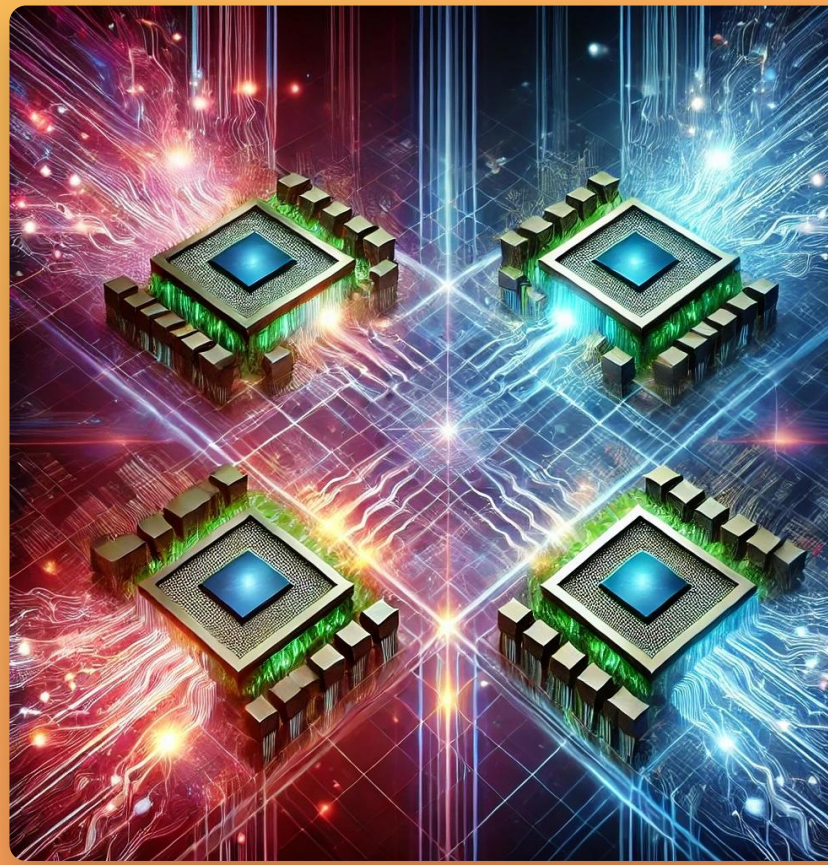
Grafana Labs

Never Underestimate Memory Architecture



Bryan Boreham

@bboreham.bsky.social

Grafana Labs

N U
M A

ヌマ

Presenter

Bryan Boreham
Distinguished Engineer
Grafana Labs

@bboreham.bsky.social

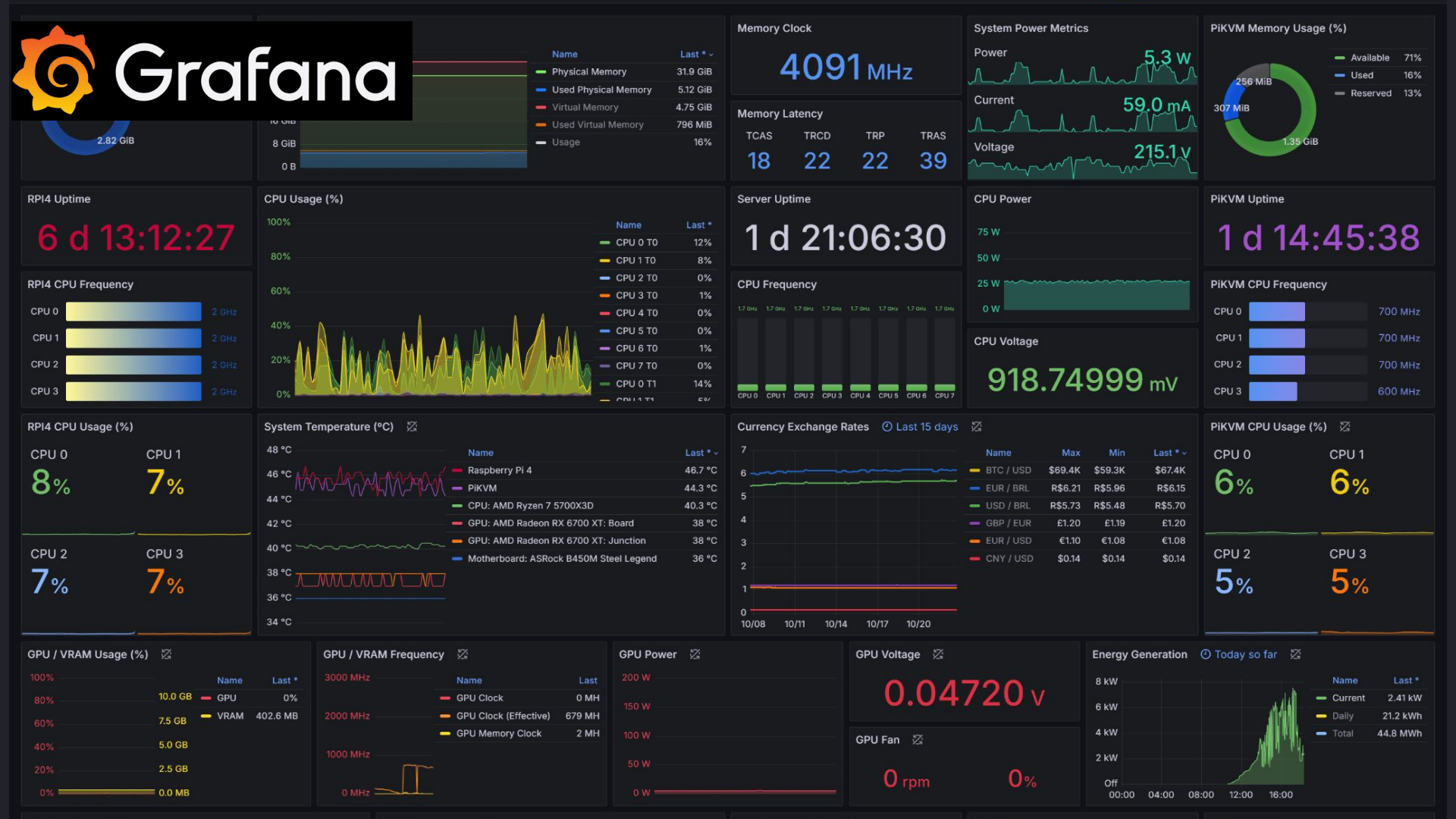
Outline

Why am I interested in NUMA?

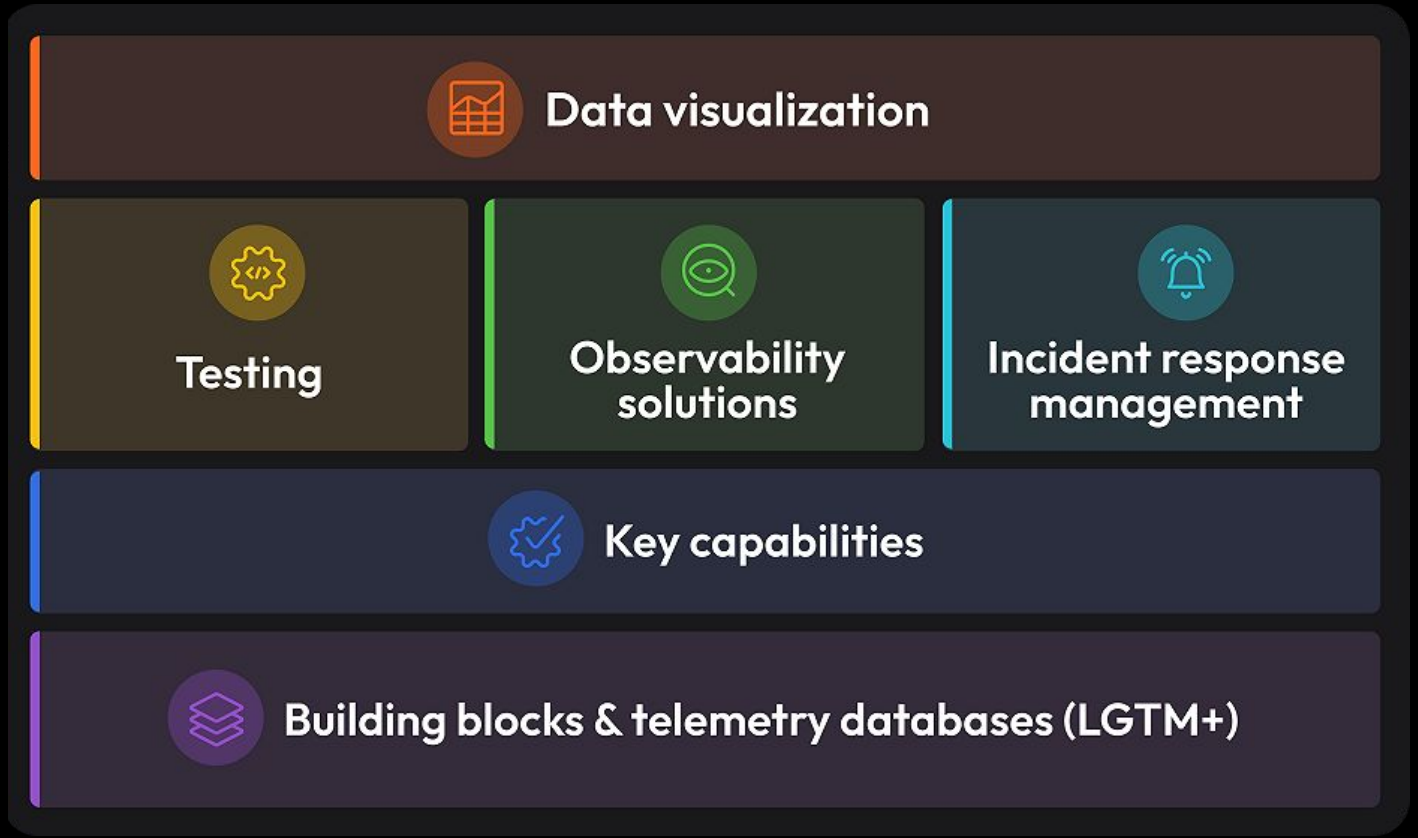
What does NUMA mean?

What can you do about it?

More memory architecture?


Grafana
2.82 GiB
8 GiB
0 B
Name
Last *
Physical Memory
31.9 GiB
Used Physical Memory
5.12 GiB
Virtual Memory
4.75 GiB
Used Virtual Memory
796 MiB
Usage
16%
Memory Clock
4091 MHz
Memory Latency
TCAS
TRCD
TRP
TRAS
18
22
22
39
System Power Metrics
Power
5.3 W
Current
59.0 mA
Voltage
215.1 V
PiKVM Memory Usage (%)
256 MiB
307 MiB
1.35 GiB
Available
71%
Used
16%
Reserved
13%
RPI4 Uptime
6 d 13:12:27
CPU Usage (%)
100%
80%
60%
40%
20%
0%
CPU 0 T0
12%
CPU 1 T0
8%
CPU 2 T0
0%
CPU 3 T0
1%
CPU 4 T0
0%
CPU 5 T0
0%
CPU 6 T0
1%
CPU 7 T0
0%
CPU 0 T1
14%
Server Uptime
1 d 21:06:30
CPU Power
75 W
50 W
25 W
0 W
PiKVM Uptime
1 d 14:45:38
RPI4 CPU Frequency
CPU 0
CPU 1
CPU 2
CPU 3
2 GHz
CPU Frequency
1.7 GHz
CPU 0
CPU 1
CPU 2
CPU 3
CPU 4
CPU 5
CPU 6
CPU 7
CPU Voltage
918.74999 mV
PiKVM CPU Frequency
CPU 0
CPU 1
CPU 2
CPU 3
700 MHz
600 MHz
RPI4 CPU Usage (%)
CPU 0
8%
CPU 1
7%
CPU 2
7%
CPU 3
7%
System Temperature (°C)
48 °C
46 °C
44 °C
42 °C
40 °C
38 °C
36 °C
34 °C
Name
Last *
Raspberry Pi 4
46.7 °C
PiKVM
44.3 °C
CPU: AMD Ryzen 7 5700X3D
40.3 °C
GPU: AMD Radeon RX 6700 XT: Board
38 °C
GPU: AMD Radeon RX 6700 XT: Junction
38 °C
Motherboard: ASRock B450M Steel Legend
36 °C
Currency Exchange Rates
Last 15 days
Name
Max
Min
Last *
BTC / USD
$69.4K
$59.3K
$67.4K
EUR / BRL
R$6.21
R$5.96
R$6.15
USD / BRL
R$5.73
R$5.48
R$5.70
GBP / EUR
£1.20
£1.19
£1.20
EUR / USD
€1.10
€1.08
€1.08
CNY / USD
$0.14
$0.14
$0.14
10/08
10/11
10/14
10/17
10/20
PiKVM CPU Usage (%)
CPU 0
6%
CPU 1
6%
CPU 2
5%
CPU 3
5%
GPU / VRAM Usage (%)
100%
80%
60%
40%
20%
0%
10.0 GB
7.5 GB
5.0 GB
2.5 GB
0.0 MB
GPU
0%
VRAM
402.6 MB
GPU / VRAM Frequency
3000 MHz
2000 MHz
1000 MHz
0 MHz
GPU Clock
0 MH
GPU Clock (Effective)
679 MH
GPU Memory Clock
2 MH
GPU Power
200 W
150 W
100 W
50 W
0 W
GPU Voltage
0.04720 V
GPU Fan
0 rpm
0%
Energy Generation
Today so far
8 kW
6 kW
4 kW
2 kW
Off
00:00
04:00
08:00
12:00
16:00
Current
2.41 kW
Daily
21.2 kWh
Total
44.8 MWh

Grafana Labs



Caution, hot material!

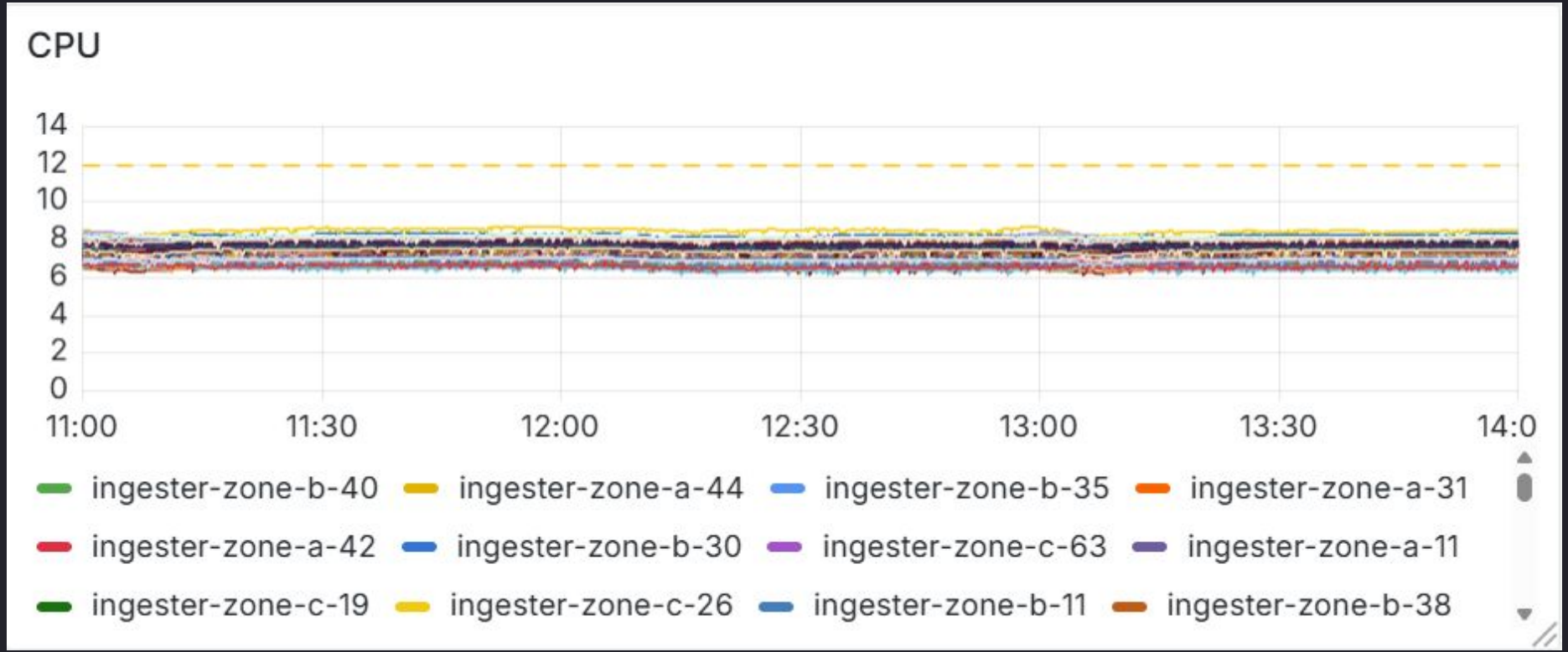
There is a lot of detail in this talk.

I settled on this sign, for the parts where you really have to pay attention.

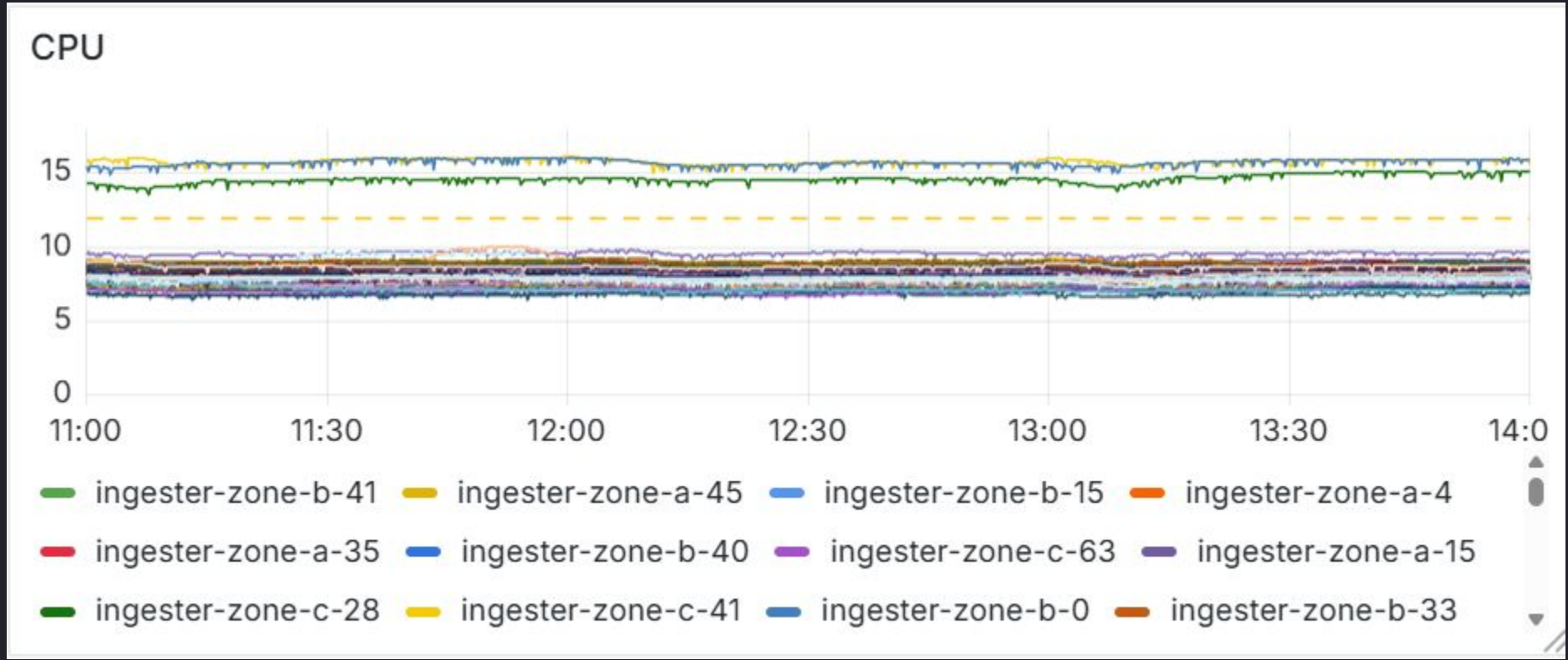
Grafana Labs

OK, let’s get on with the talk

Story begins here



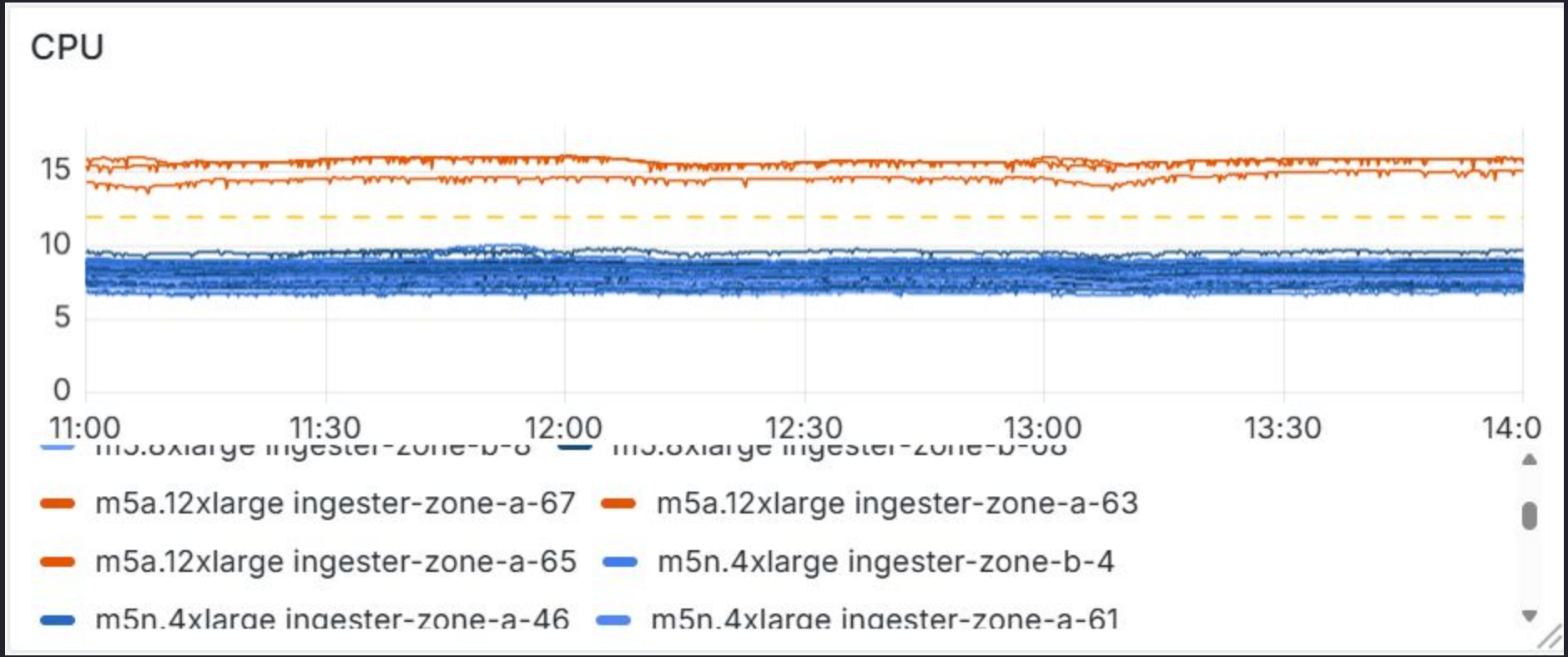
One day, some pods are using more CPU



Grafana Labs

What’s going on?

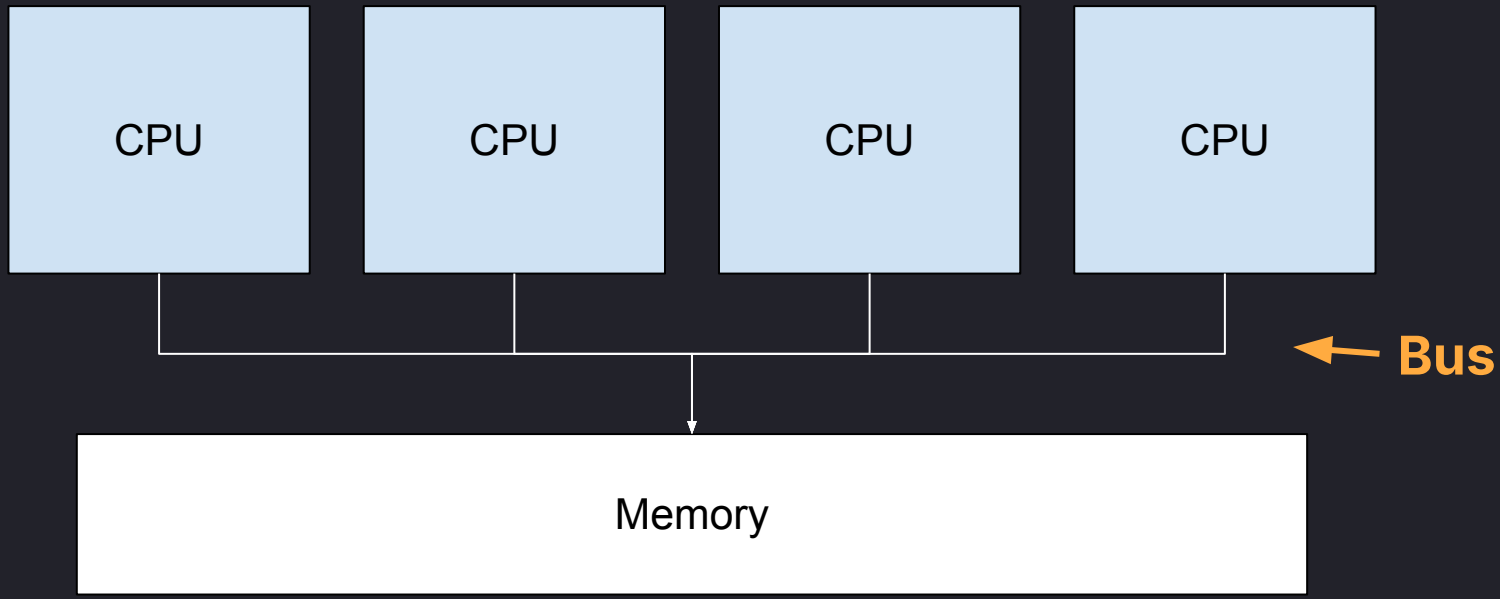
Showing instance type



Grafana Labs

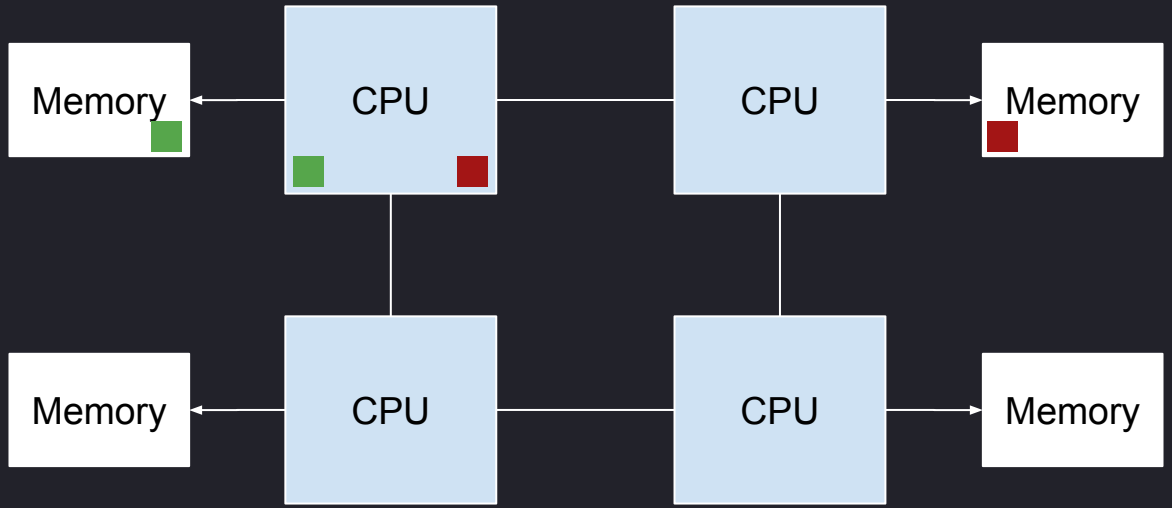
How do computers work?

Conceptually

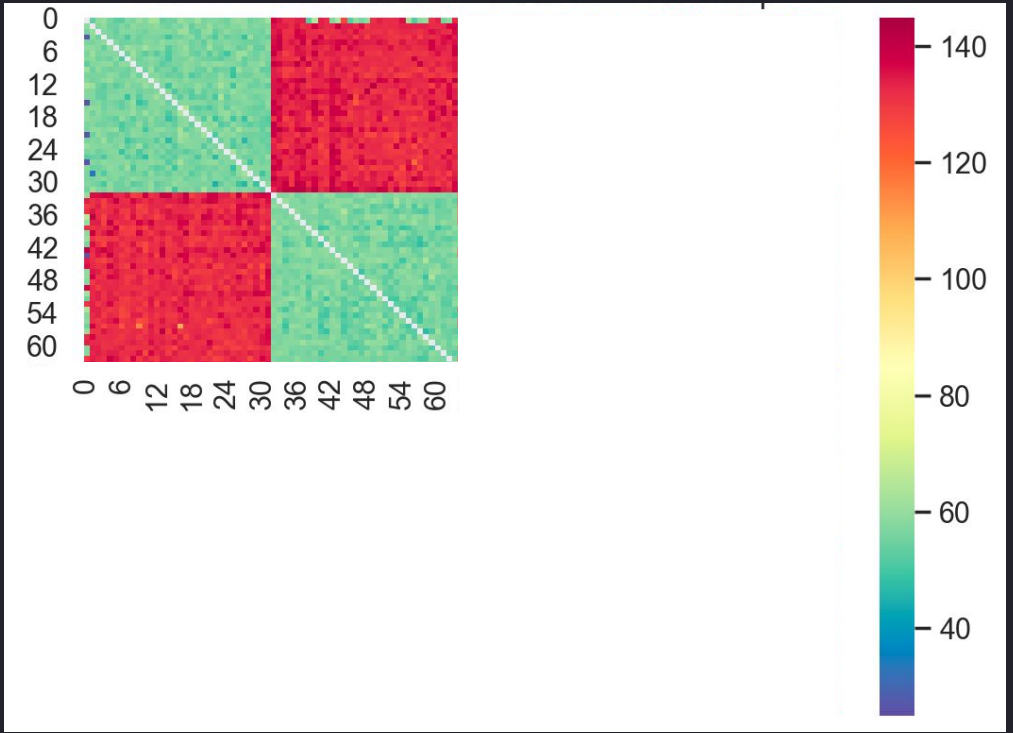


Reality

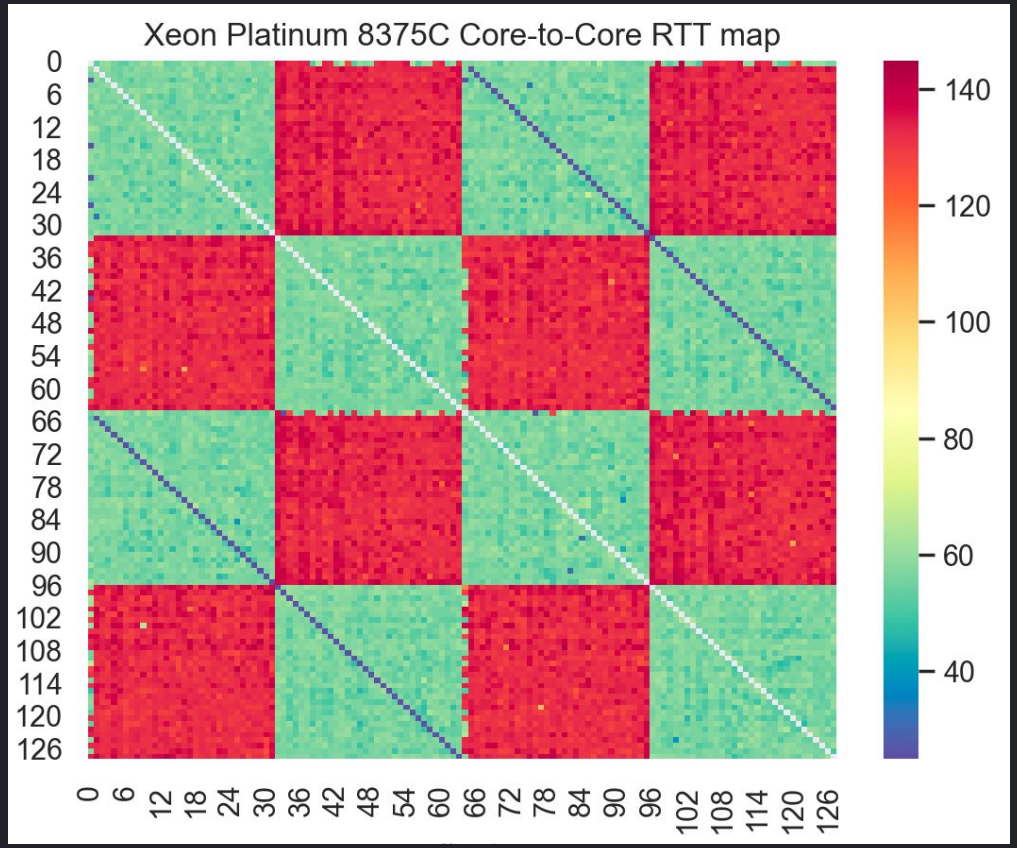
Non-Uniform Memory Access (NUMA)



Memory latency



Memory latency, with hyperthreading



Grafana Labs

Sidenote: workloads vary

Grafana Labs

How do I know if I'm suffering from NUMA?

Prometheus metrics

From node-exporter (with `zoneinfo` collector enabled):

`node_zoneinfo_numa_other_totalvs`

`node_zoneinfo_numa_local_total`

Grafana Labs

How do I find a server's NUMA characteristics?

On the CSP Website?

Instance Size	vCPU	Memory (GiB)	Instance Storage (GB)	Network Bandwidth (Gbps)	EBS Bandwidth (Gbps)
m7i.large	2	8	EBS-Only	Up to 12.5	Up to 10
m7i.xlarge	4	16	EBS-Only	Up to 12.5	Up to 10
m7i.2xlarge	8	32	EBS-Only	Up to 12.5	Up to 10
m7i.4xlarge	16	64	EBS-Only	Up to 12.5	Up to 10
m7i.8xlarge	32	128	EBS-Only	12.5	10
m7i.12xlarge	48	192	EBS-Only	18.75	15

`lscpu` command - m5a.12xlarge

```
Architecture:                    x86_64
  Model name:                    AMD EPYC 7571
    Thread(s) per core:          2
    Core(s) per socket:          24
    Socket(s):                   1
NUMA:
  NUMA node(s):                  3
  NUMA node0 CPU(s):             0-7,24-31
  NUMA node1 CPU(s):             8-15,32-39
  NUMA node2 CPU(s):             16-23,40-47
```

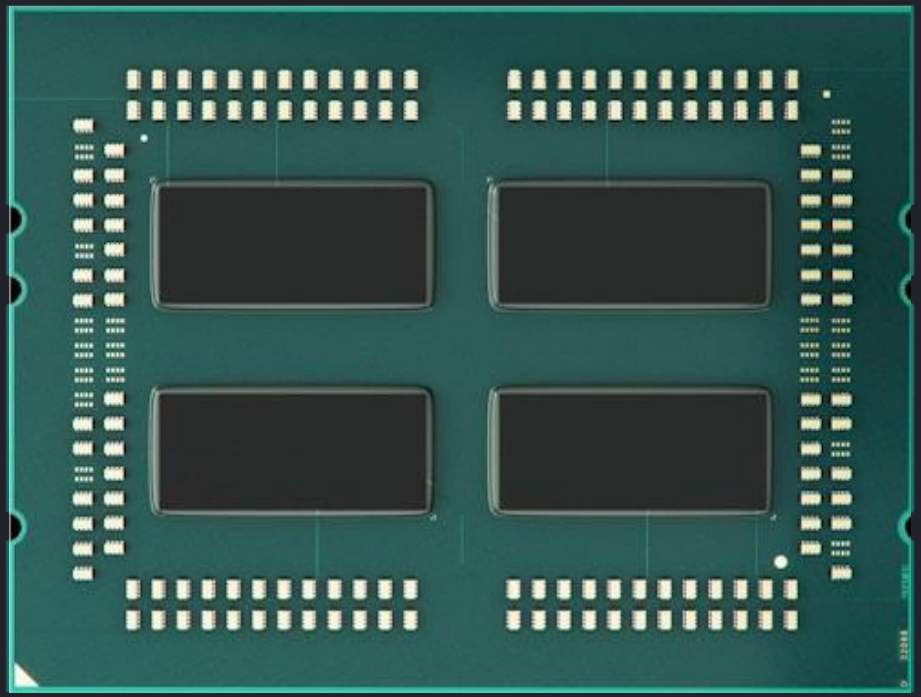
`lscpu` command - m7g.16xlarge

```
Architecture:                    aarch64
  Model name:                    Neoverse-V1
    Thread(s) per core:          1
    Core(s) per cluster:         64
    Cluster(s):                  1
NUMA:
  NUMA node(s):                  1
  NUMA node0 CPU(s):             0-63
```

`lscpu` command - m8g.metal-48xl

```
Architecture:                    aarch64
  Model name:                    Neoverse-V2
    Thread(s) per core:          1
    Core(s) per cluster:         96
    Cluster(s):                  2
NUMA:
  NUMA node(s):                  2
  NUMA node0 CPU(s):             0-95
  NUMA node1 CPU(s):             96-191
```

AMD Chiplets

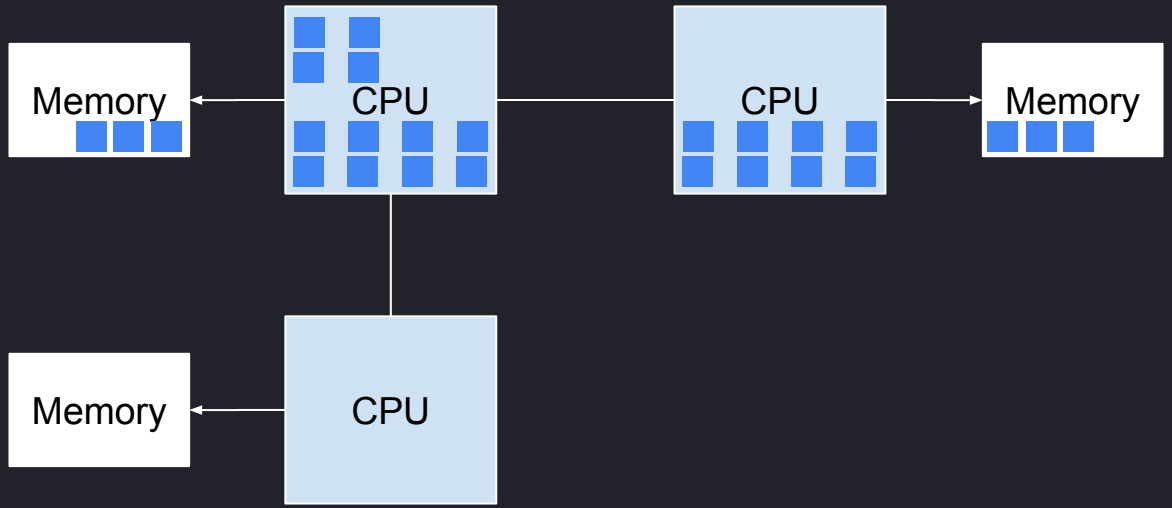


AMD EPYC 7xx1 32-core (64 with SMT)

Memory latency, m5a.12xlarge model

C->C	CPU0	CPU1	CPU2	CPU3	CPU4	CPU5	CPU6	CPU7	CPU8	CPU9	CPU10	CPU11	CPU12	CPU13	CPU14	CPU15	CPU16	CPU17	CPU18	CPU19	CPU20	CPU21	CPU22	CPU23
CPU0	X	32	33	34	182	182	183	182	220	220	220	220	220	220	220	219	222	222	223	223	222	223	223	223
CPU1	32	X	32	33	182	181	183	182	219	218	219	219	219	218	219	219	221	225	222	225	221	224	222	225
CPU2	33	32	X	33	183	183	183	183	220	220	221	220	220	220	221	220	223	223	223	223	223	223	223	223
CPU3	34	33	33	X	182	182	183	182	219	219	220	219	219	219	220	219	222	222	223	223	222	222	222	223
CPU4	182	182	183	182	X	32	33	34	220	219	220	220	220	220	220	220	222	222	223	222	222	222	223	223
CPU5	182	181	183	182	32	X	32	33	218	218	219	219	219	218	219	219	221	224	222	225	221	224	222	225
CPU6	183	183	183	183	33	32	X	33	221	220	221	220	220	220	221	220	223	223	223	223	223	223	224	223
CPU7	182	182	183	182	34	33	33	X	220	219	220	220	220	219	220	220	222	222	222	222	222	222	222	223
CPU8	220	219	220	219	220	218	221	220	X	32	33	34	182	182	183	182	222	224	222	225	222	224	223	224
CPU9	220	218	220	219	219	218	220	219	32	X	32	33	182	181	183	182	221	221	222	221	221	221	222	222
CPU10	220	219	221	220	220	219	221	220	33	32	X	33	183	182	183	183	222	226	223	227	223	226	223	227
CPU11	220	219	220	219	220	219	220	220	34	33	33	X	182	182	183	182	222	224	222	224	222	224	222	224
CPU12	220	219	220	219	220	219	220	220	182	182	183	182	X	32	33	34	222	224	222	224	222	224	223	224
CPU13	220	218	220	219	220	218	220	219	182	181	182	182	32	X	32	33	221	221	222	222	221	221	222	222
CPU14	220	219	221	220	220	219	221	220	183	183	183	183	33	32	X	33	222	226	223	226	223	226	223	227
CPU15	219	219	220	219	220	219	220	220	182	182	183	182	34	33	33	X	222	224	222	225	222	224	222	224
CPU16	222	221	223	222	222	221	223	222	222	221	222	222	222	221	222	222	X	32	33	34	192	193	192	194
CPU17	222	225	223	222	222	224	223	222	224	221	226	224	224	221	226	224	32	X	32	33	193	191	196	193
CPU18	223	222	223	223	223	222	223	222	222	222	223	222	222	222	223	222	33	32	X	33	193	196	194	197
CPU19	223	225	223	223	222	225	223	222	225	221	227	224	224	222	226	225	34	33	33	X	193	193	196	195
CPU20	222	221	223	222	222	221	223	222	222	221	223	222	222	221	223	222	192	193	193	193	X	32	33	34
CPU21	223	224	223	222	222	224	223	222	224	221	226	224	224	221	226	224	193	191	196	193	32	X	32	33
CPU22	223	222	223	222	223	222	224	222	223	222	223	222	223	222	223	222	192	196	194	196	33	32	X	33
CPU23	223	225	223	223	223	225	223	223	224	222	227	224	224	222	227	224	194	193	197	195	34	33	33	X

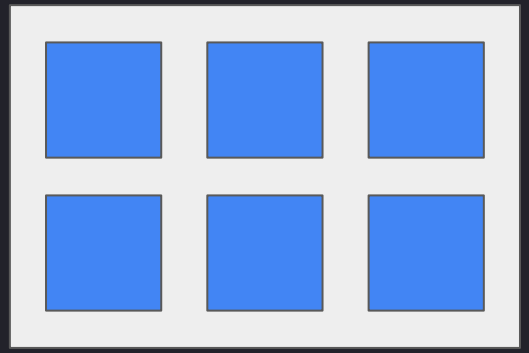
Situation in the incident (probably)



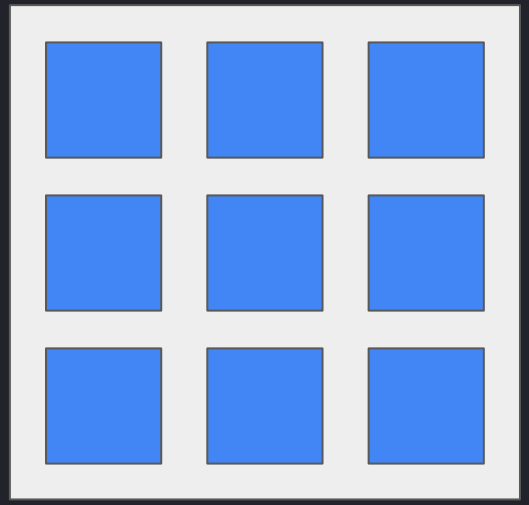
Grafana Labs

How did we resolve the incident?

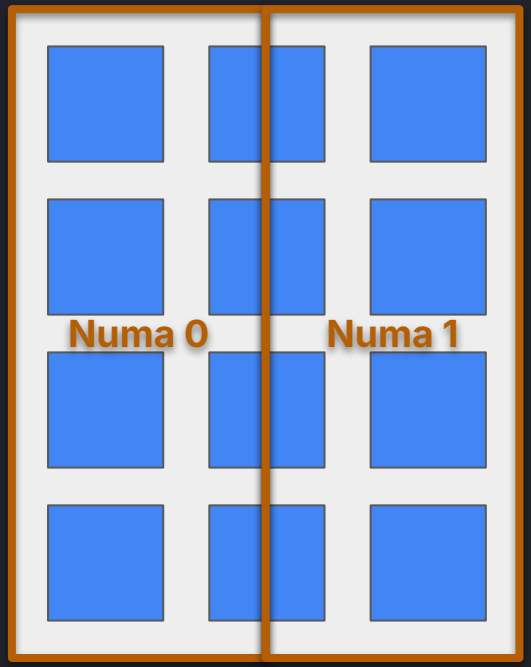
Say you start with small instances



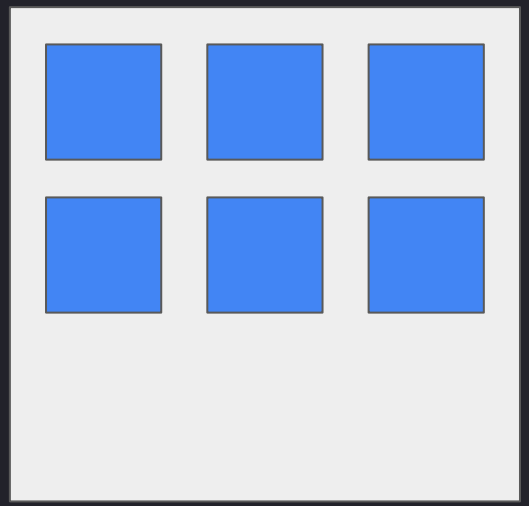
Then you move to bigger instances

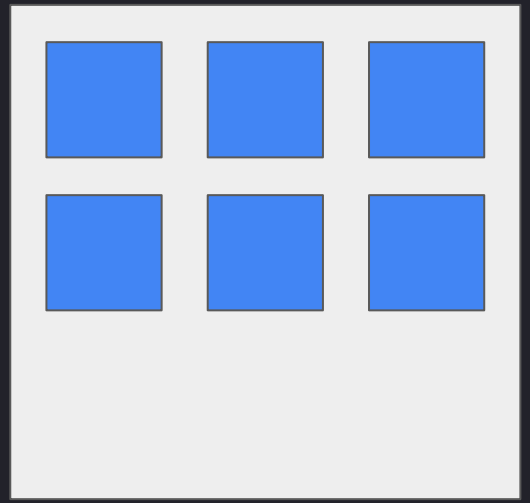


And bigger



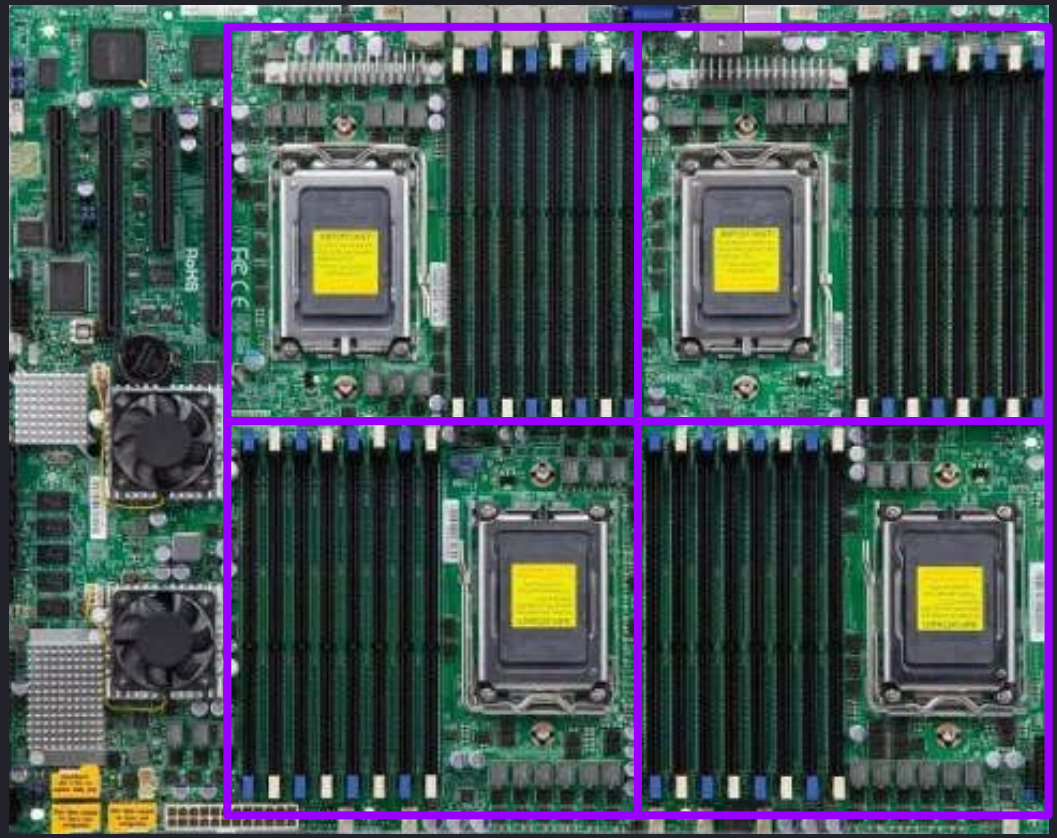
You might be better off with multiple smaller instances





But remember there are per-instance overheads

Physical view



Grafana Labs

Kubernetes

CPU Manager

`--cpu-manager-policy` kubelet flag

- **`none`**: the default policy.
- **`static`**: for containers in Guaranteed pods with integer CPU requests

“CPUManager aligns CPU allocations at the NUMA boundary”

https://kubernetes.io/docs/concepts/policy/node-resource-managers/

Topology Manager

`scope=pod` - to locate a set of containers together on the same NUMA nodes.

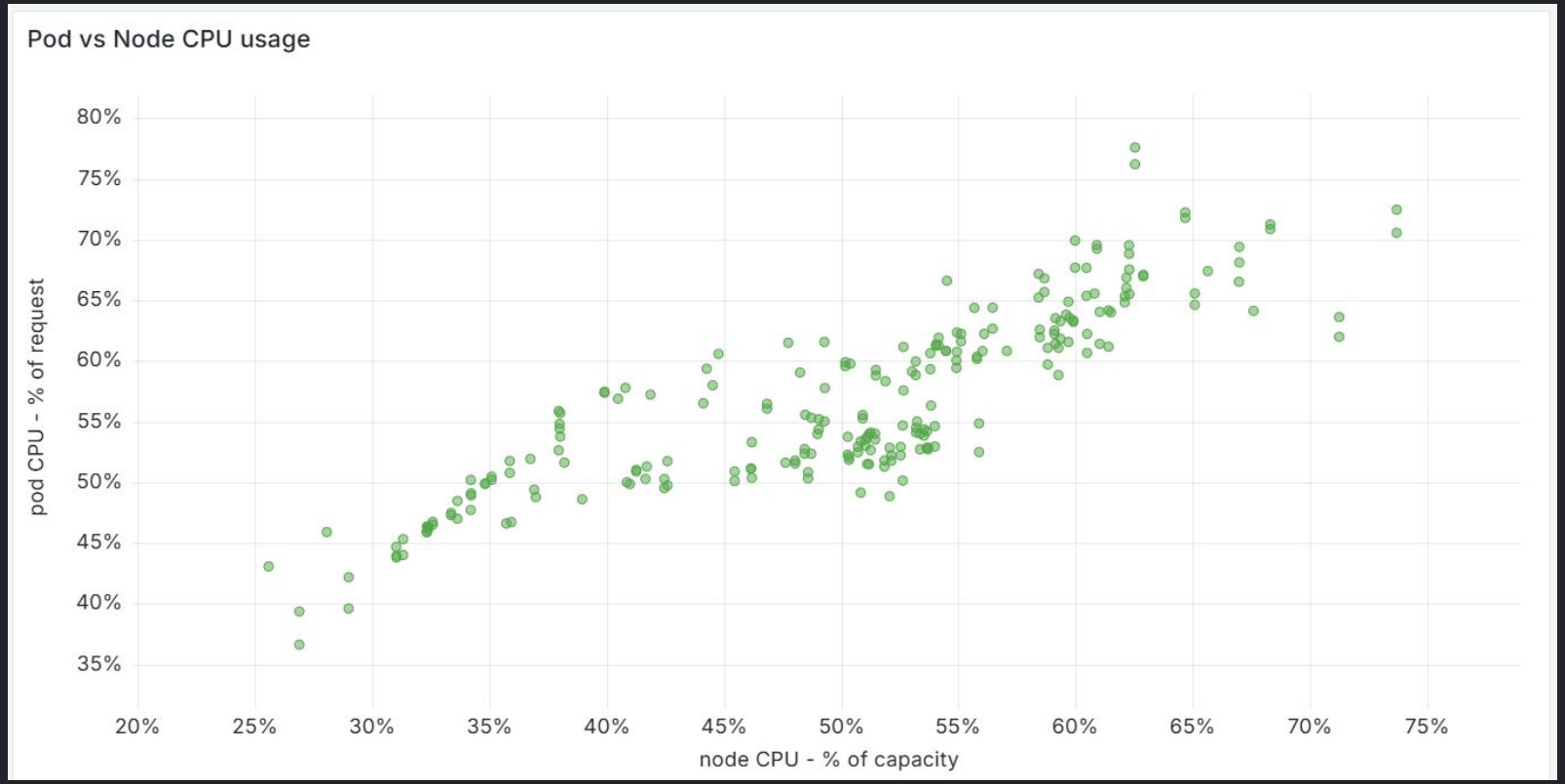
`policy=single-numa-node` - will reject a pod if it can't all run on one node.

https://kubernetes.io/docs/tasks/administer-cluster/topology-manager/

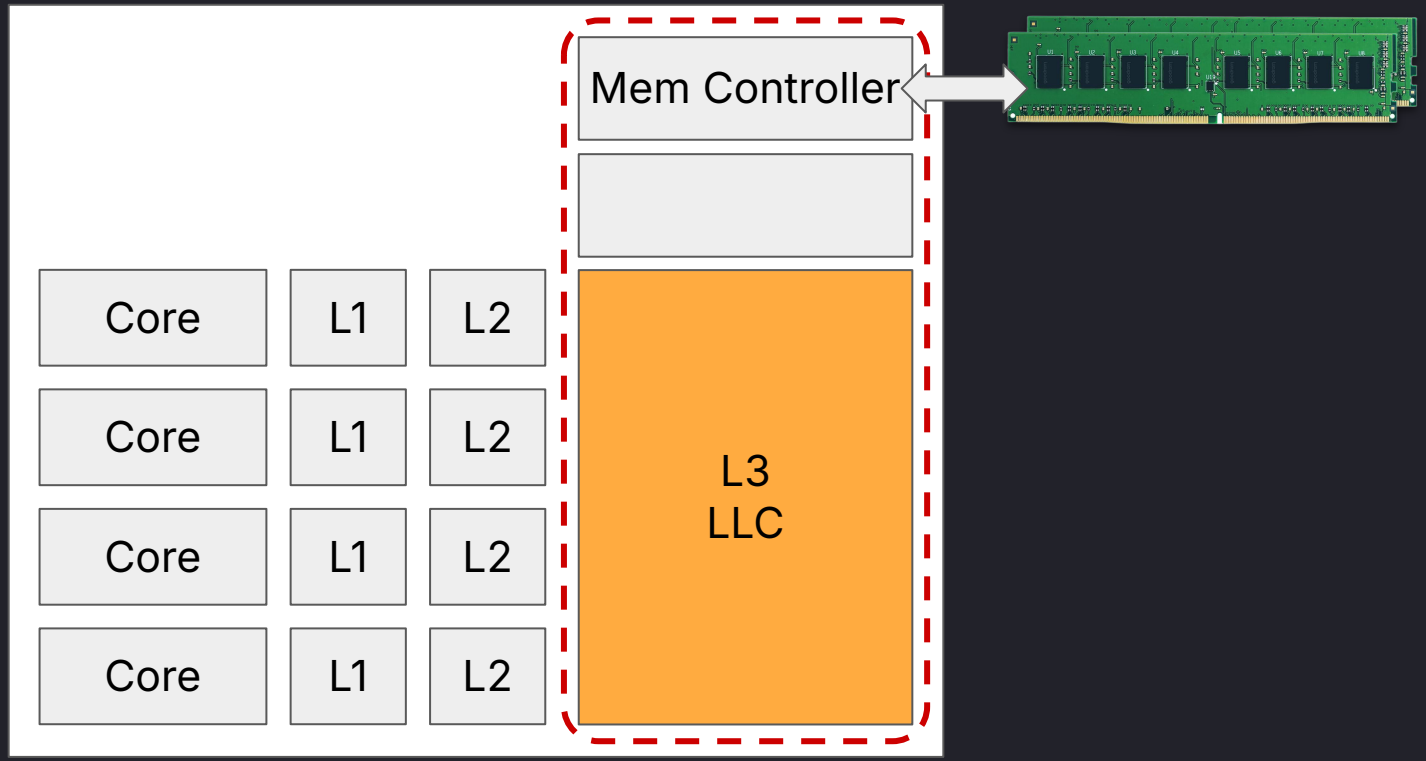
Grafana Labs

“Uncore”

Pods vs instances CPU plot



“Uncore” = the parts outside of cores



CPU Manager

`--cpu-manager-policy=static` kubelet flag

Option: **`prefer-align-cpus-by-uncorecache`**

New in Kubernetes version 1.32 (alpha, hidden by default).

https://kubernetes.io/docs/tasks/administer-cluster/cpu-management-policies/

Wrapping up

Benefit from a little understanding of underlying hardware.

Stay within NUMA zones if you possibly can.

Everything is non-uniform.

Grafana Labs

Thank you

Special thanks to:

JuanJo Ciarlante, Grafana Labs

Anton Kolesnikov, Grafana Labs

Jonathan Perry, Unvariance

Milind Chabbi, Uber

Links

https://github.com/SoilRos/cpu-latency

https://www.intel.com/.../articles/technical/intel-vtune-amplifier-functionality-on-aws-instances

https://jprahman.substack.com/p/amd-cpu-topology-rome-milan

https://www.nextplatform.com/2019/08/15/a-deep-dive-into-amds-rome-epyc-architecture/

https://anarazel.de/talks/2024-10-23-pgconf-eu-numa-vs-postgresql/numa-vs-postgresql.pdf

https://frankdenneman.nl/2019/10/14/amd-epyc-naples-vs-rome-and-vsphere-cpu-scheduler-updates/


Grafana Labs