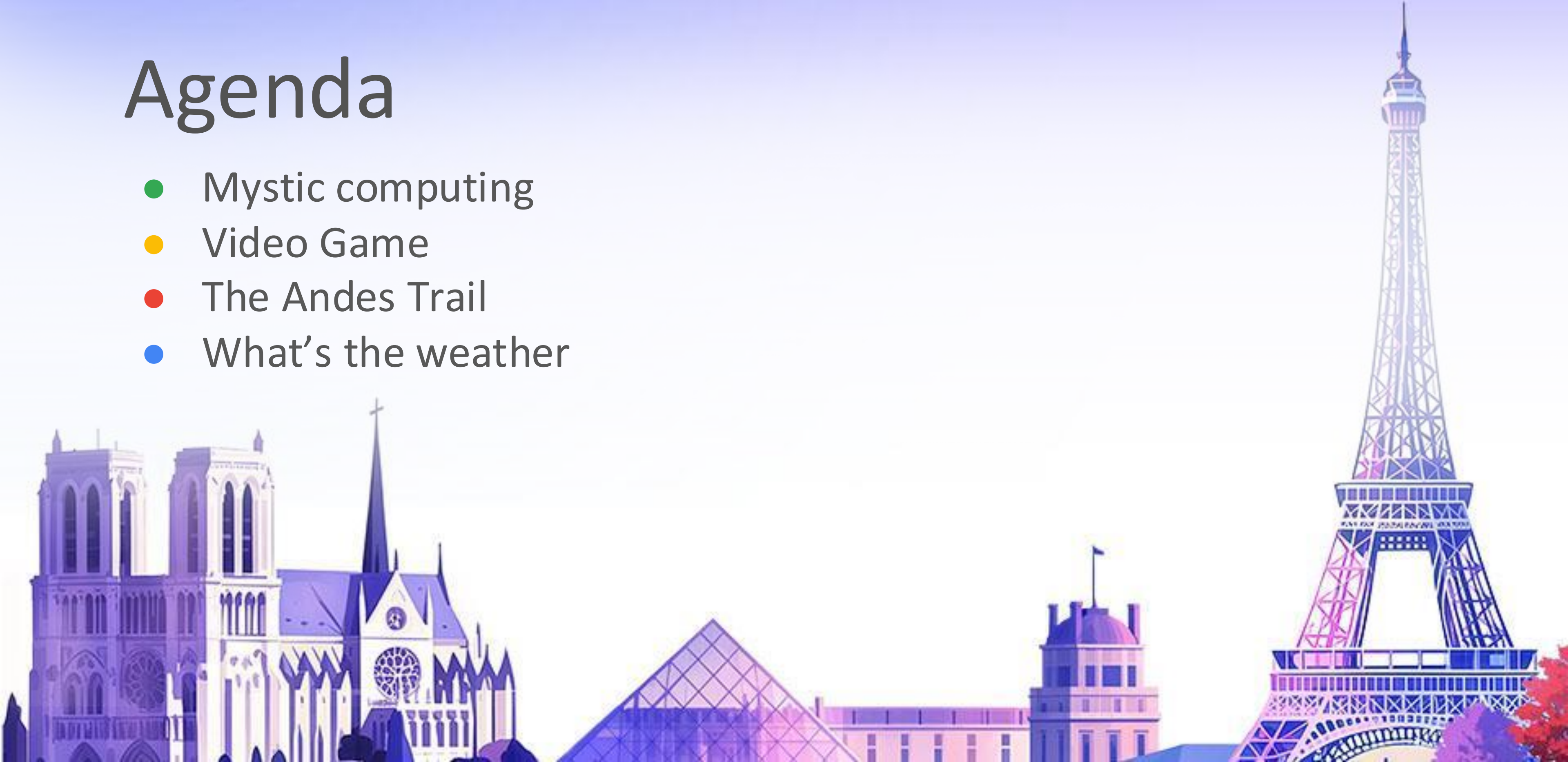
Make your LLM Serverless

Guillaume Blaquiere
@gblaquiere

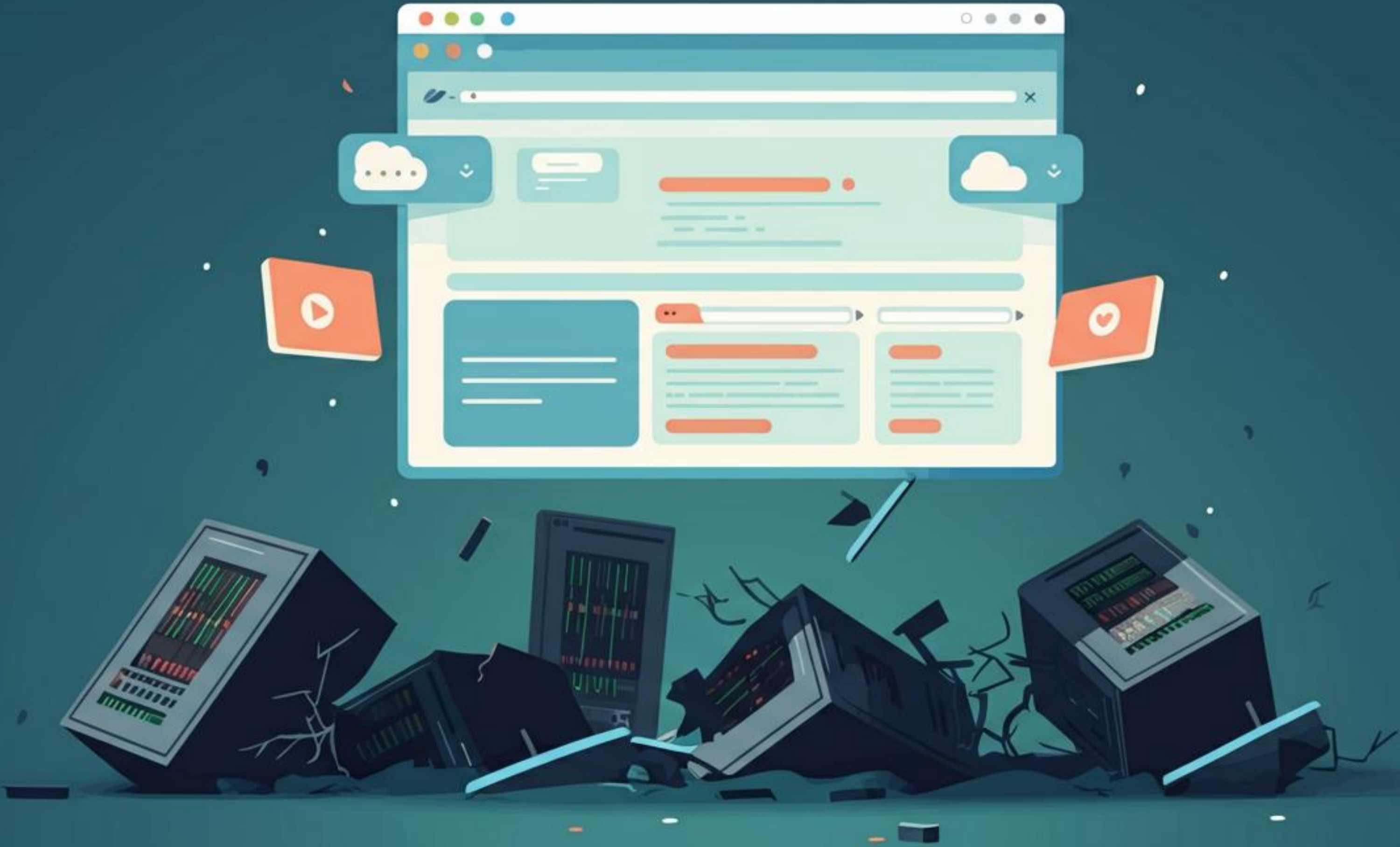
Google Developer Expert - Cloud
Group Data Architect - Carrefour

Agenda

- Mystic computing
- Video Game
- The Andes Trail
- What’s the weather



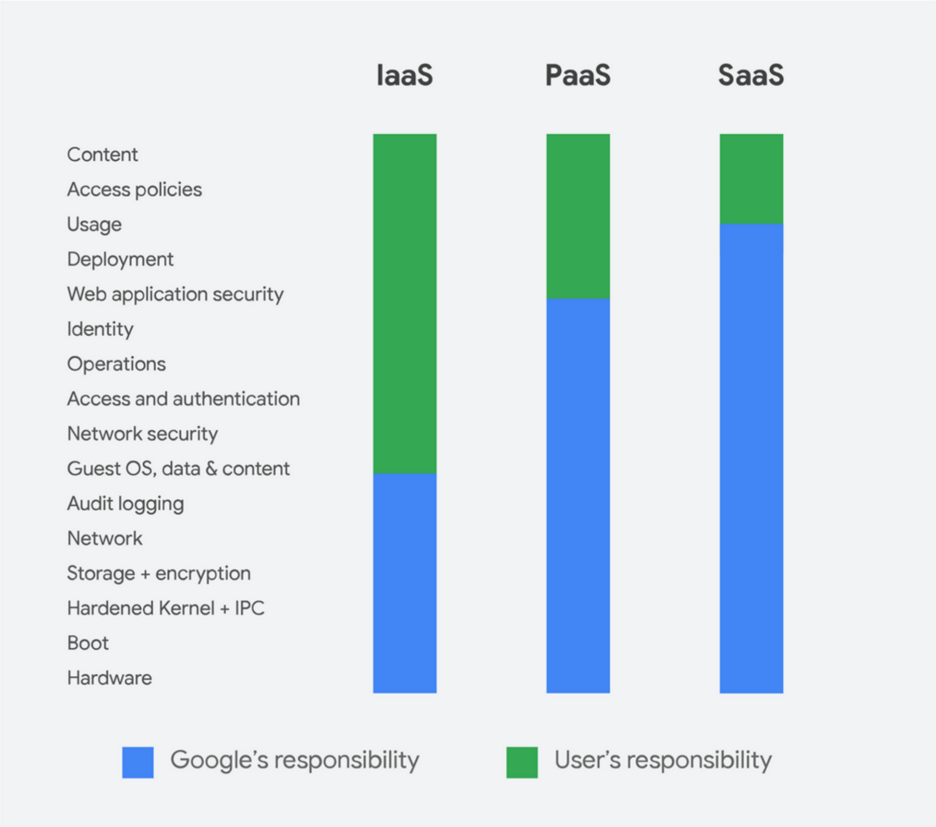
Mystic Computing



Serverless computing

A shared responsibility model

Run your application without **worrying about** the servers



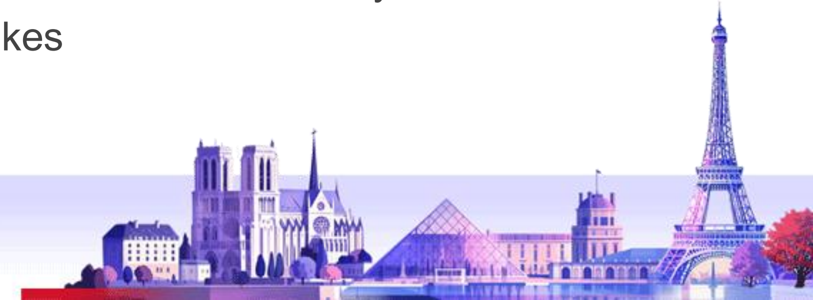
Cloud run service

Serverless container platform

Zero config deployments
gcloud run deploy

Auto-scaling
to support
peak traffic spikes

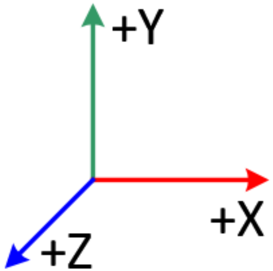
Pay only
while your code runs



Video Game foundations

Mathematics in the core

Vertex and 3D computation



$$\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$$

Vector

translation

$$\begin{pmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} v_x \\ v_y \\ v_z \\ 1 \end{pmatrix} = \begin{pmatrix} v_x + t_x \\ v_y + t_y \\ v_z + t_z \\ 1 \end{pmatrix}$$

Scaling

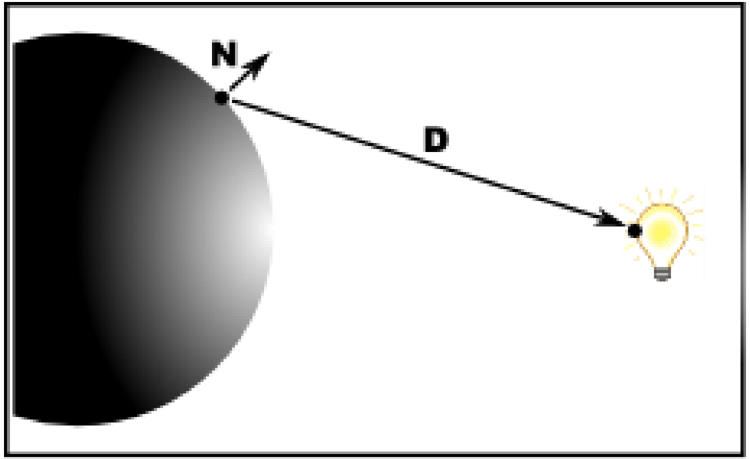
$$\begin{pmatrix} s_x & 0 & 0 & 0 \\ 0 & s_y & 0 & 0 \\ 0 & 0 & s_z & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} v_x \\ v_y \\ v_z \\ 1 \end{pmatrix} = \begin{pmatrix} s_x v_x \\ s_y v_y \\ s_z v_z \\ 1 \end{pmatrix}$$

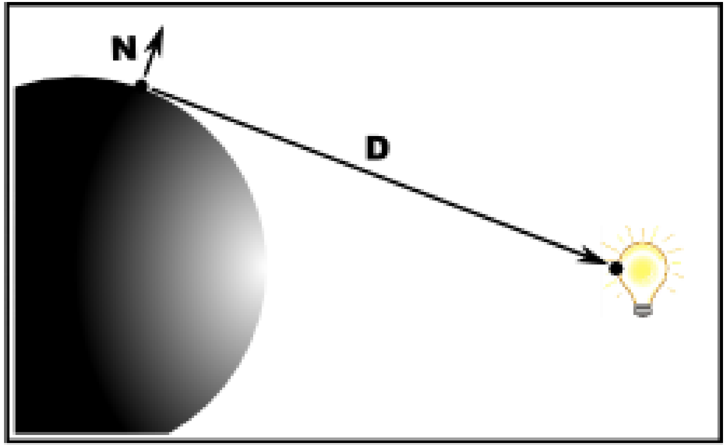
Rotation

$$\mathbf{R}_X(\theta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta) & \sin(\theta) & 0 \\ 0 & -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Mathematics in the core

Light computation





Scalar product to determine the light & reflexion surface alignment

Math recap and secret superpower

School reminder

Addition and multiplication in the core

$$\begin{bmatrix} A & B \\ C & D \\ E & F \end{bmatrix} \times \begin{bmatrix} G \\ H \end{bmatrix} = \begin{bmatrix} A \times G + B \times H \\ C \times G + D \times H \\ E \times G + F \times H \end{bmatrix}$$

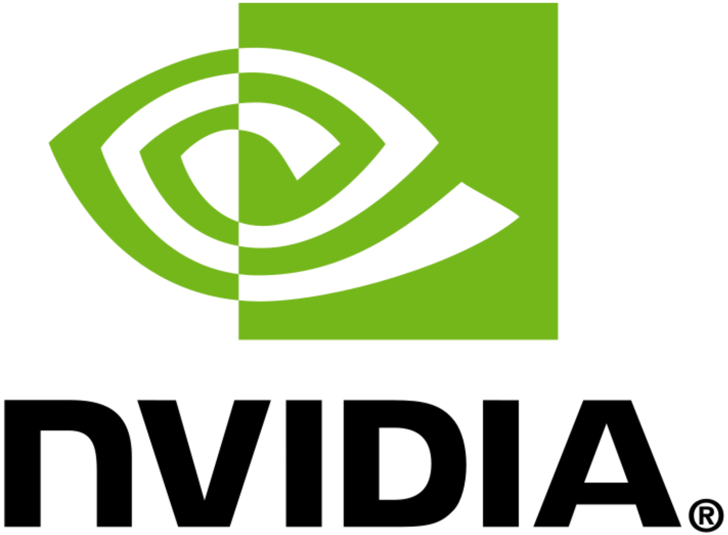
← **Matrix - vector multiplication**

Scalar product →

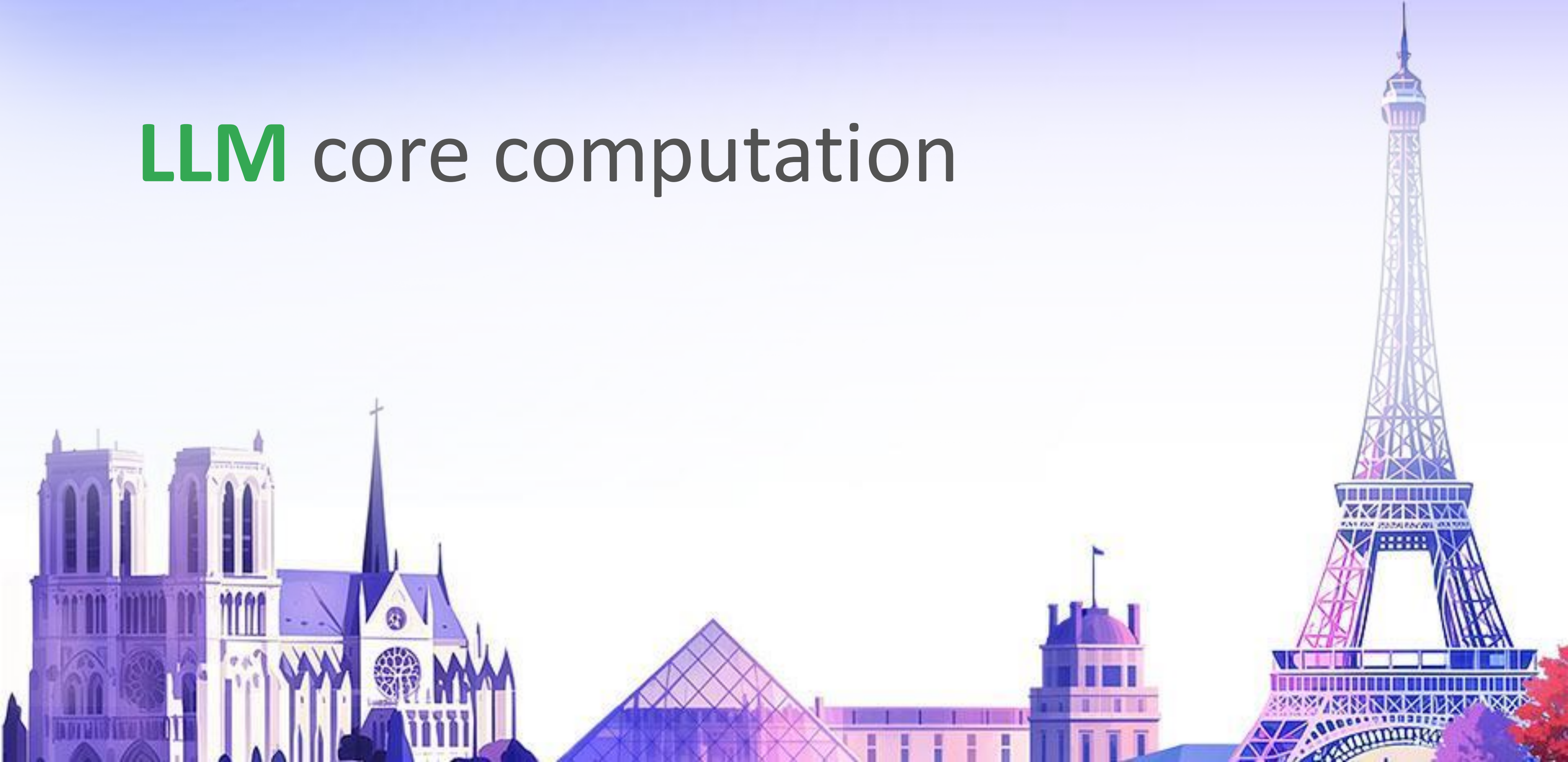
$$\begin{bmatrix} A_1 & A_2 & A_3 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} = A_1B_1 + A_2B_2 + A_3B_3 = \vec{A}.\vec{B}$$

GPUs revolution

Video game super power



LLM core computation



LLMs, matrices and vectors

Feel the similarities

“This is a prompt”

Prompt

Array of bytes

This
is
a
pro
mpt

Vector of tokens

0.5	1.2	0.35	8.9	0.54	1
0.4	0.01	1.6	5.1	1.3	0.57
1.8	0.4	5.6	10	4.4	0.24
5.1	0.25	0.7	0.92	0.3	0.02
0.25	0.68	0.08	1.25	5.91	0.99

Vector of token array values

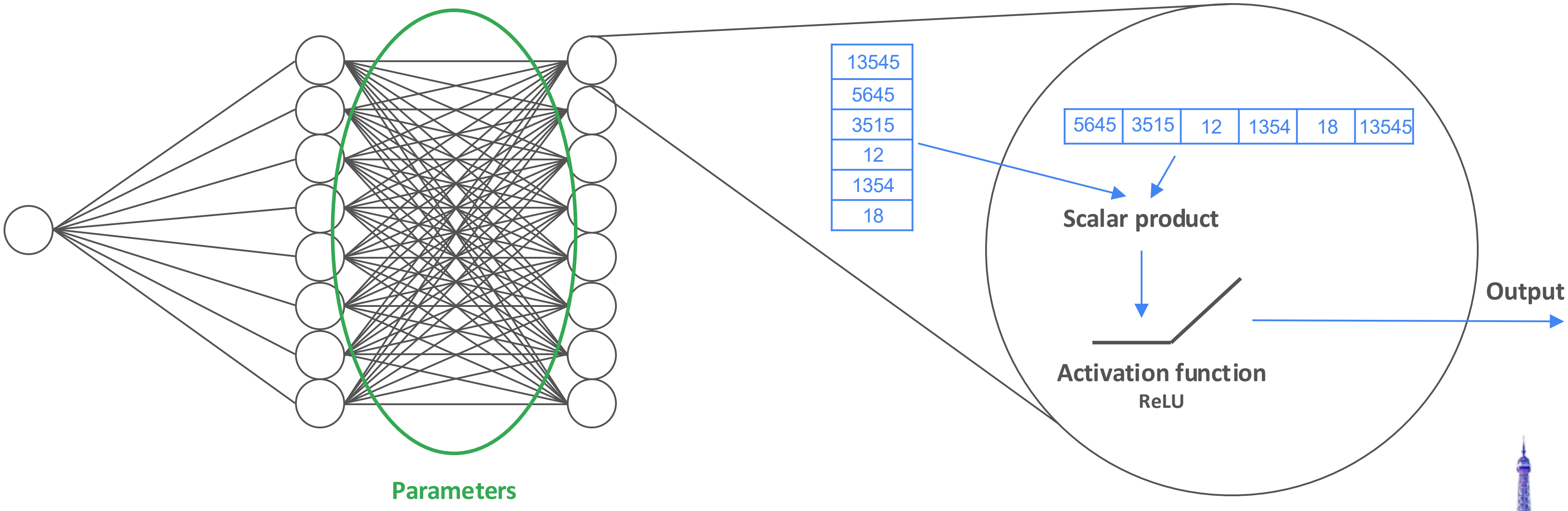
Matrix

13545
5645
3515
12
1354
18
1561531
1812
15644

Embeddings

Neuron activation

Scalar product and activation function



Not so different

GPUs in the core

		Video Games	LLMs
Matrix - vector multiplication	→	Vertices transformation	Tokenization and embeddings
Scalar product	→	Light effect	Neuron alignement & activation

Cloud run & GPUs

Since September 2024

The Andes trail

LLAMA
GEMMA
MISTRAL
OH!

Ollama swiss knife

Serving LLM, easily

Multi LLM support

Gemma

LLaMA by Meta

MISTRAL AI_

Adaptive runtime

NVIDIA

~~NVIDIA~~

Open Source

Secure

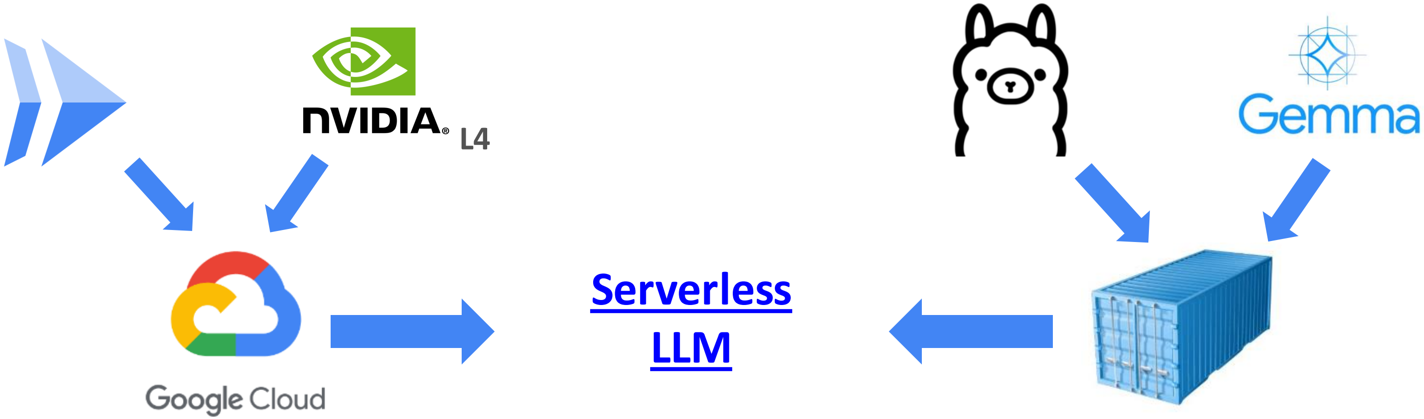
Weather **mix**

Running LLM, serverless

Mix all the ingredients

Cloud Run X GPU

Ollama X Gemma



Serverless LLM

No solution's perfect

Pros and Cons

No overprovisioning

scale to 0

Scale with the traffic

pay as you use

Easy to use

Auto driver installation

Cold start

First request latency

Limitation

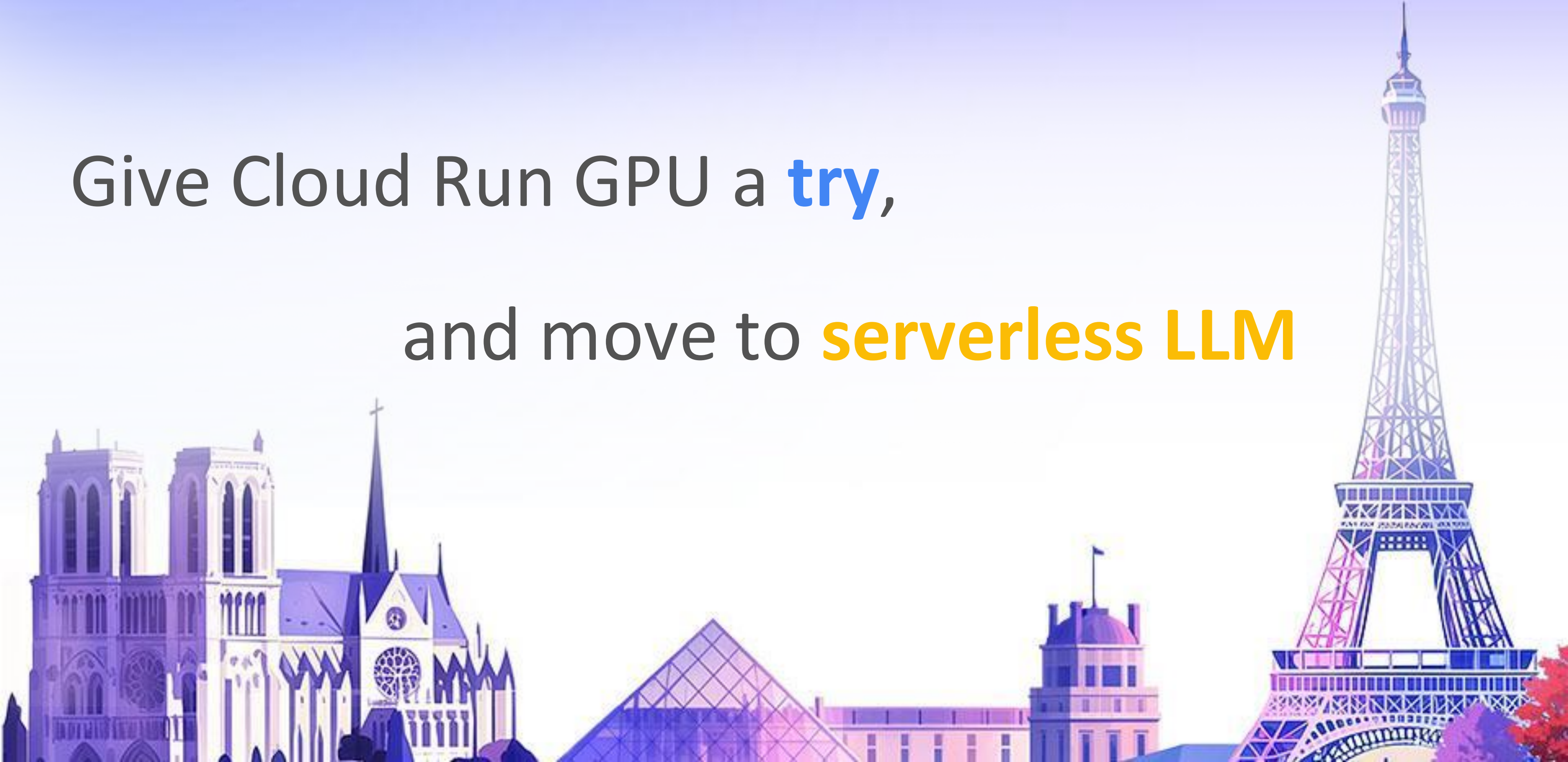
Regions & max instances

GPUs available

only NVidia L4

Give Cloud Run GPU a **try**,

and move to **serverless LLM**



THANK YOU

Article https://medium.com/google-cloud/cloud-run-gpu-make-your-llms-serverless-5188caacc667

Find me on :

Twitter	@gblaquiere
Medium	@guillaume-blaquiere
GitHub	guillaumeblaquiere
LinkedIn	guillaume blaquiere