



| **SCALITYDAY26**

How fast can your company think?

Why performance matters.



Charles Prost
VP Engineering
Scality

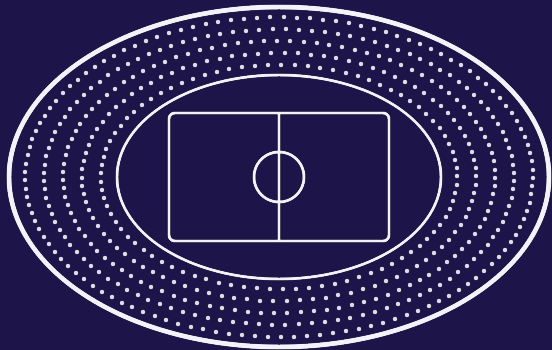
Performance



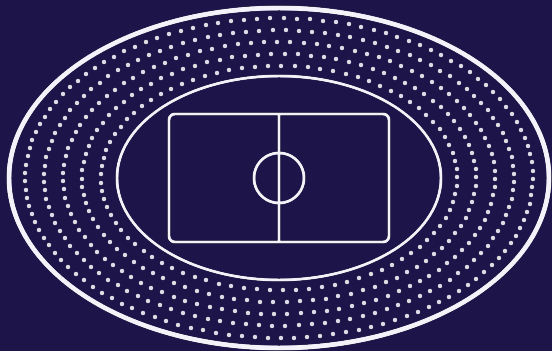
numbers



numbers



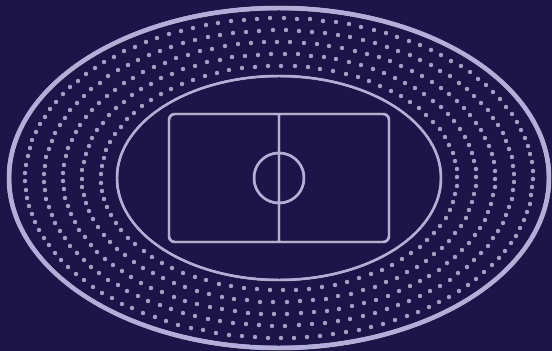
numbers



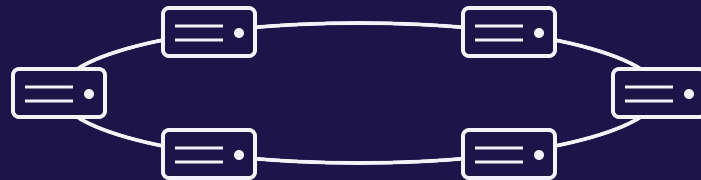
80,000



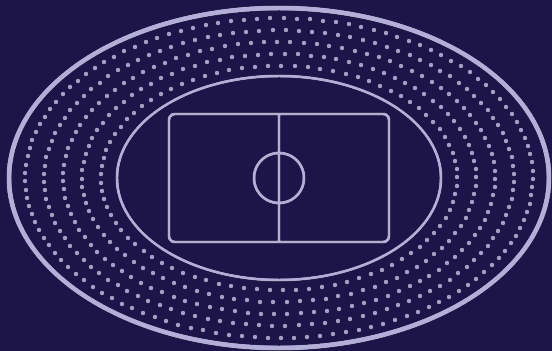
numbers



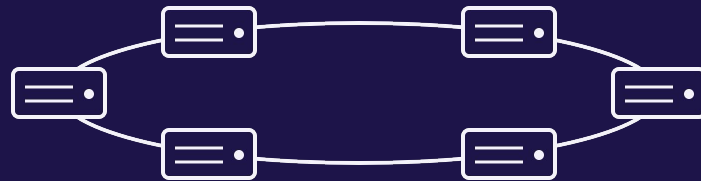
80,000



numbers



80,000

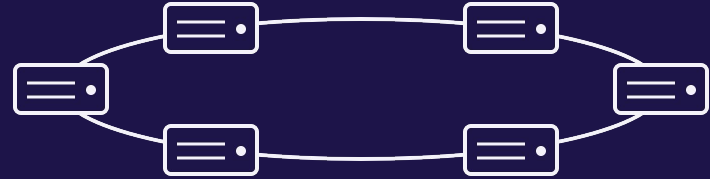
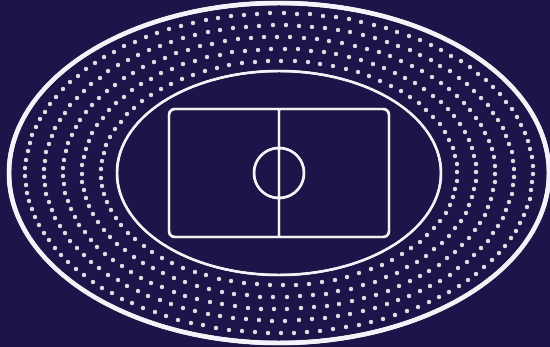


10 PB



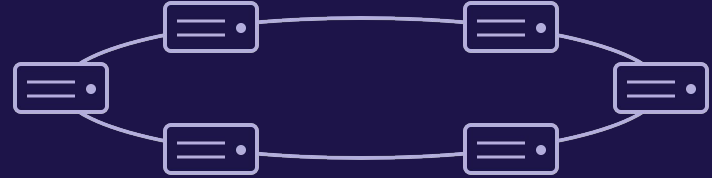
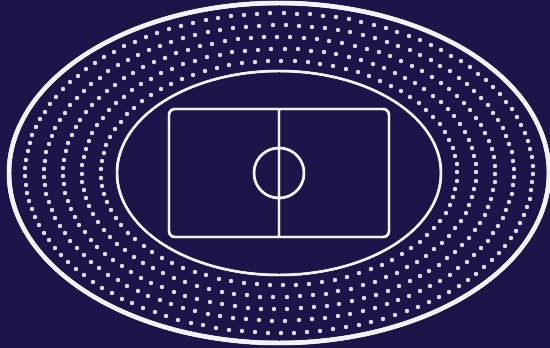
 **numbers**

 **time**



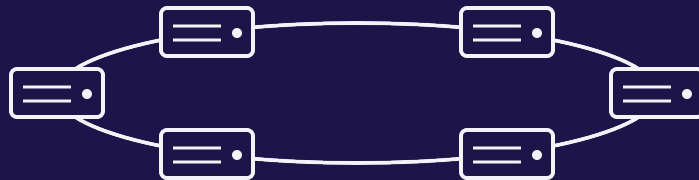
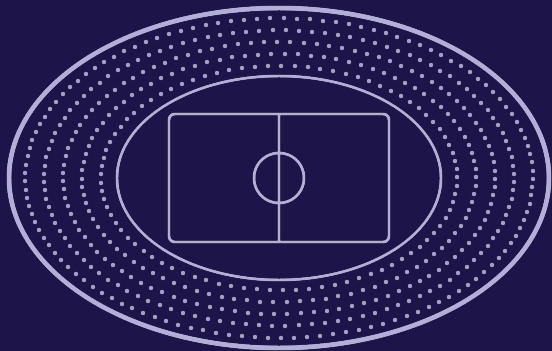
 **numbers**

 **time**



 **numbers**

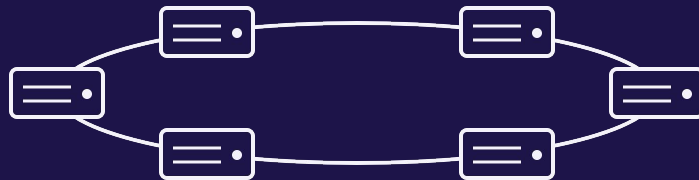
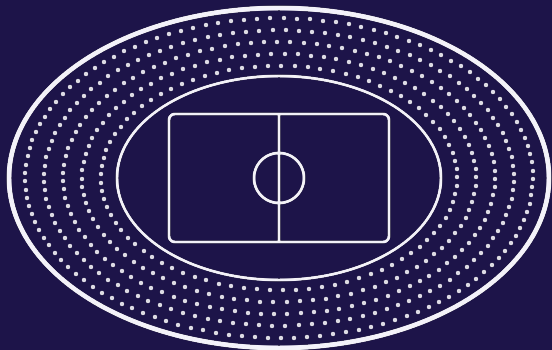
 **time**



 **numbers**

 **time**

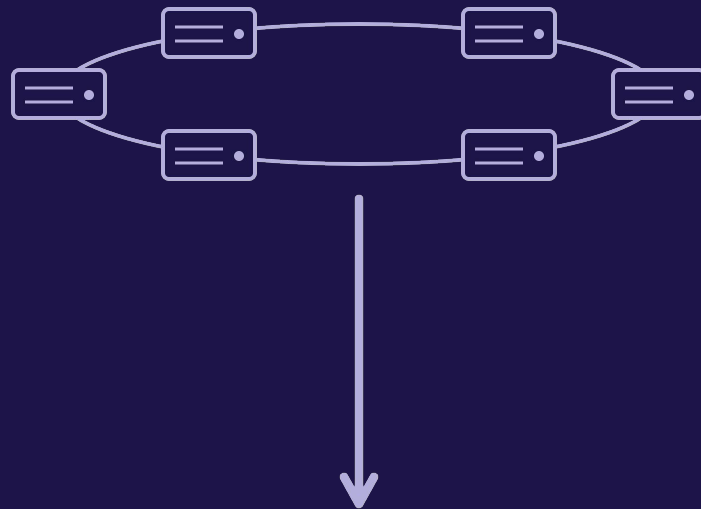
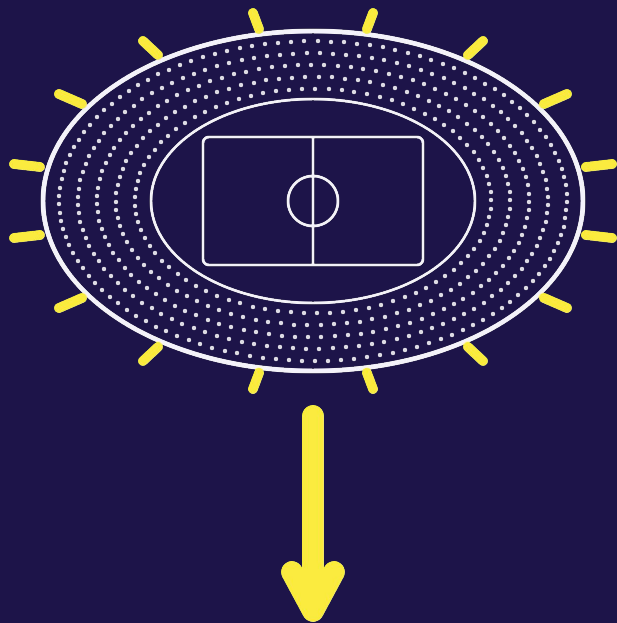
 **speed**



 **numbers**

 **time**

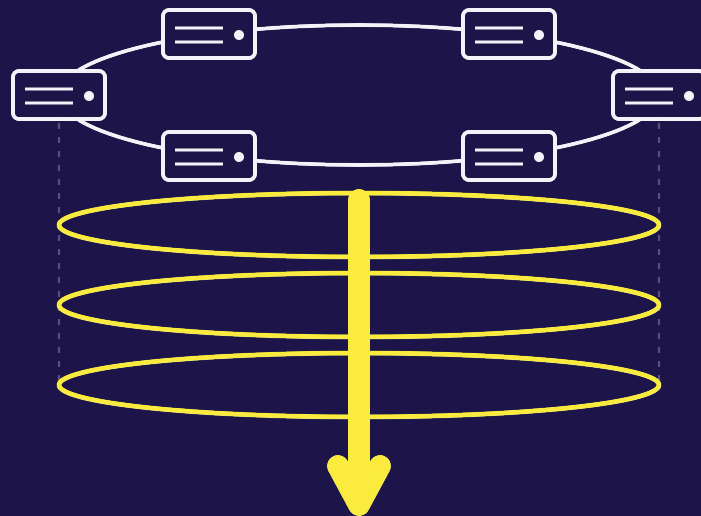
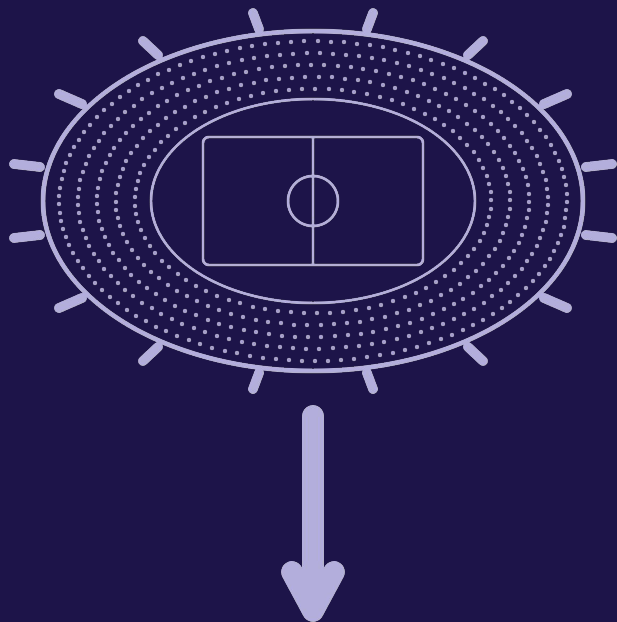
 **speed**



 **numbers**

 **time**

 **speed**

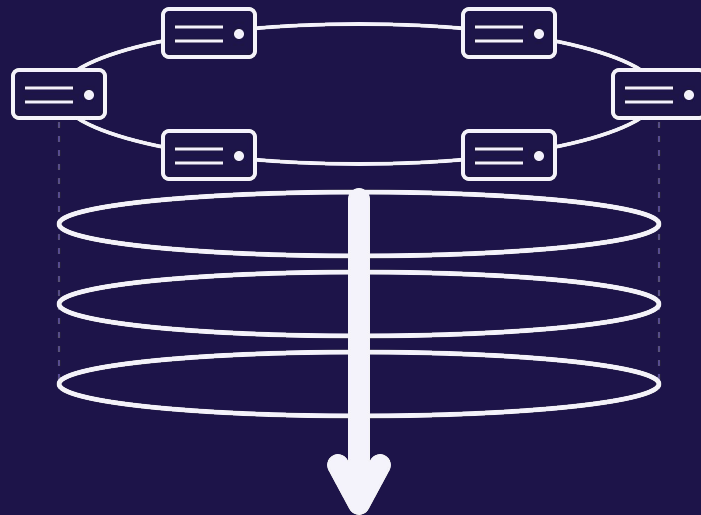
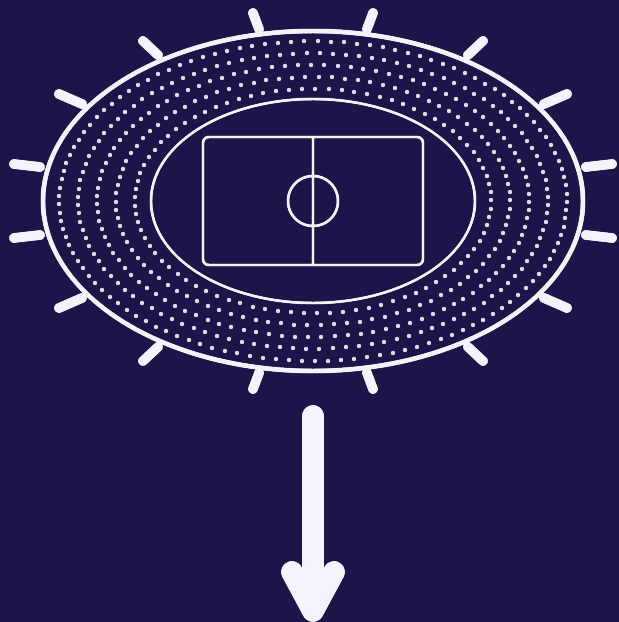


 **numbers**

 **time**

 **speed**

 **flow**

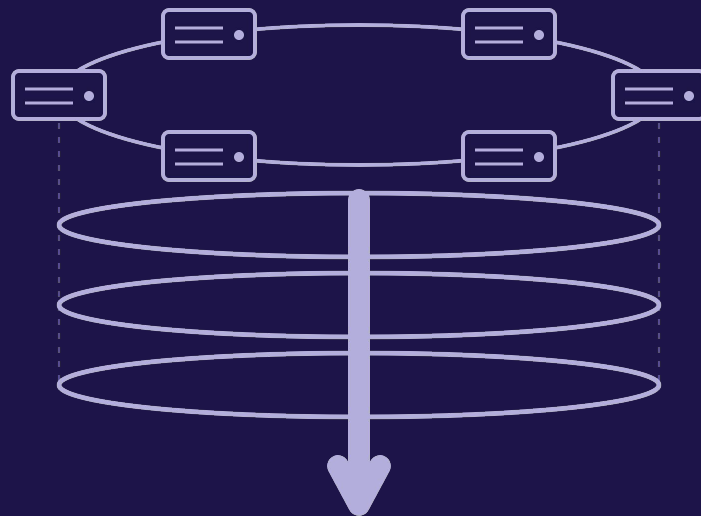
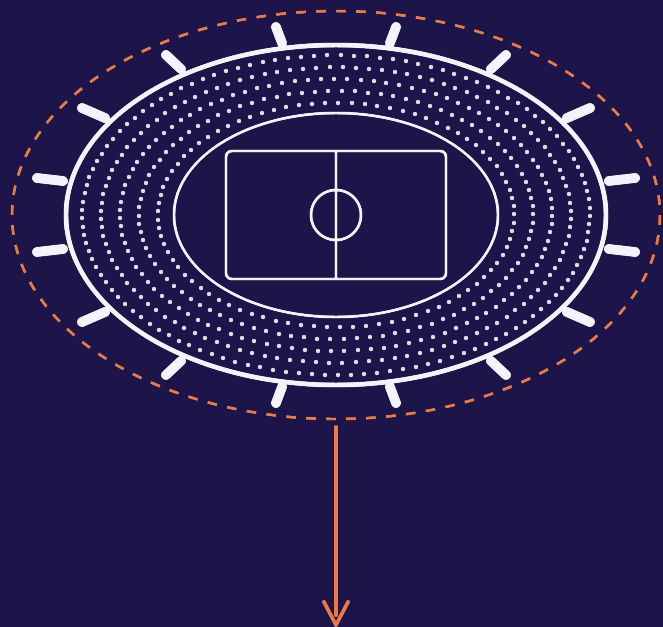


 **numbers**

 **time**

 **speed**

 **flow**

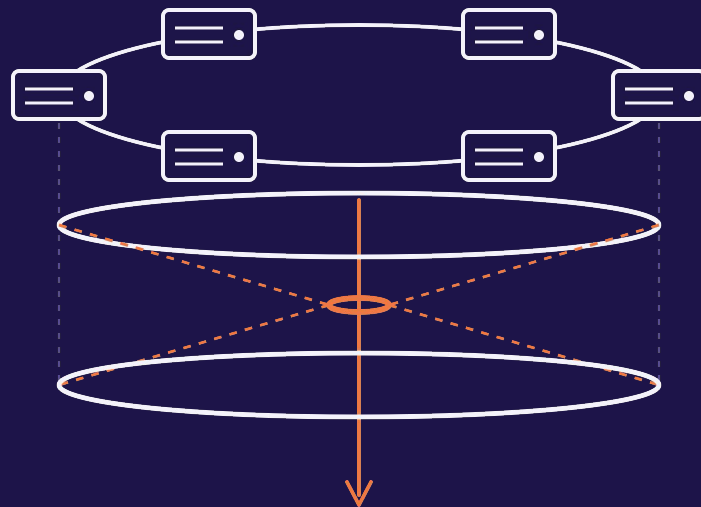
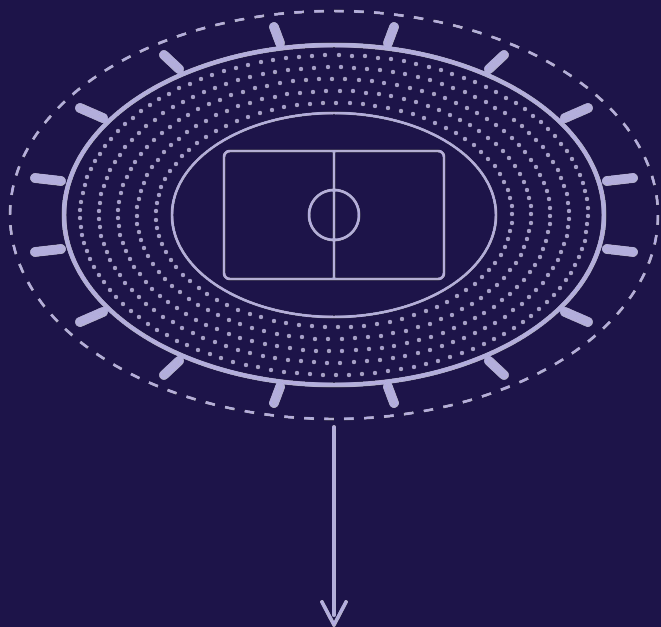


 **numbers**

 **time**

 **speed**

 **flow**

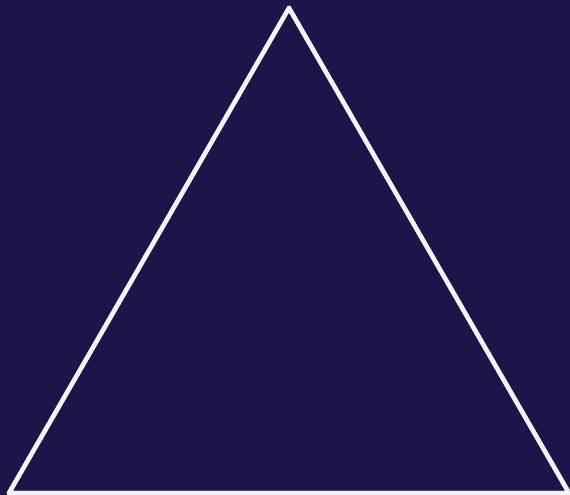


Performance

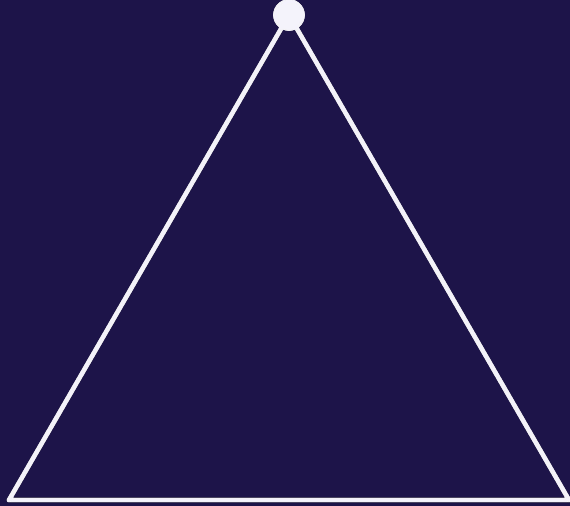


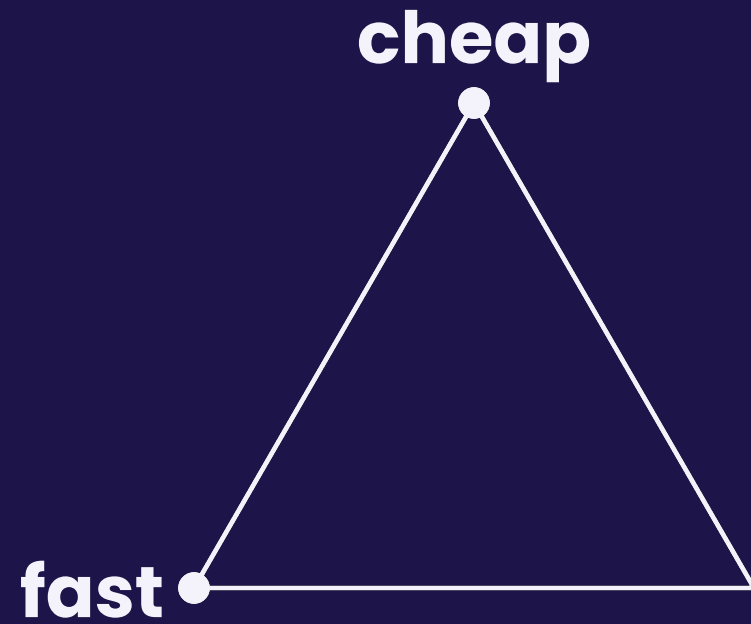
Performance
= the flow of the full system

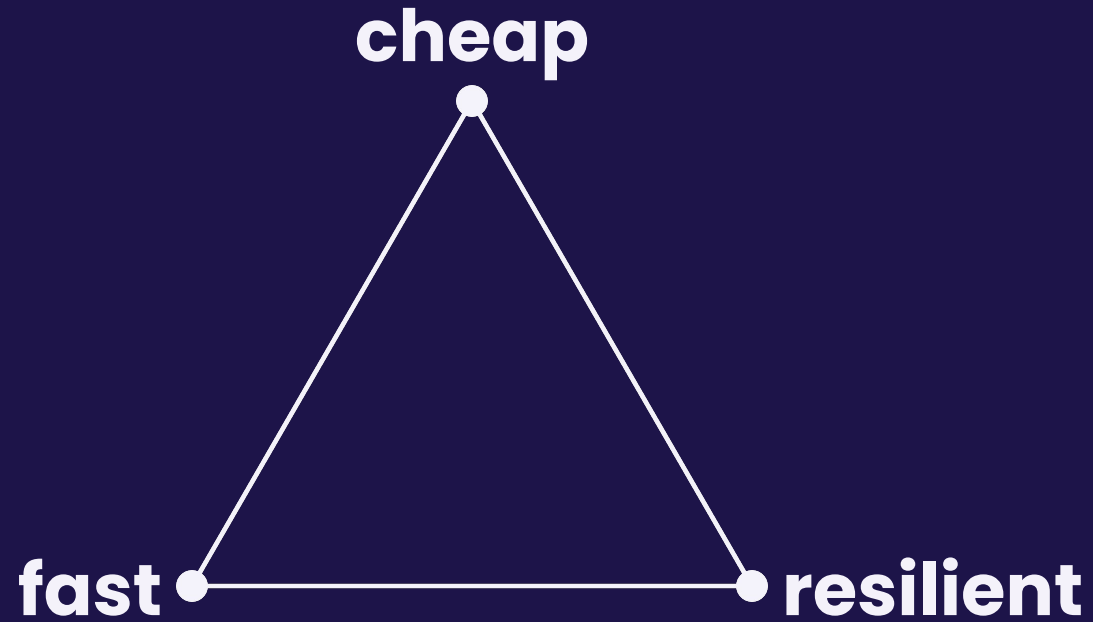


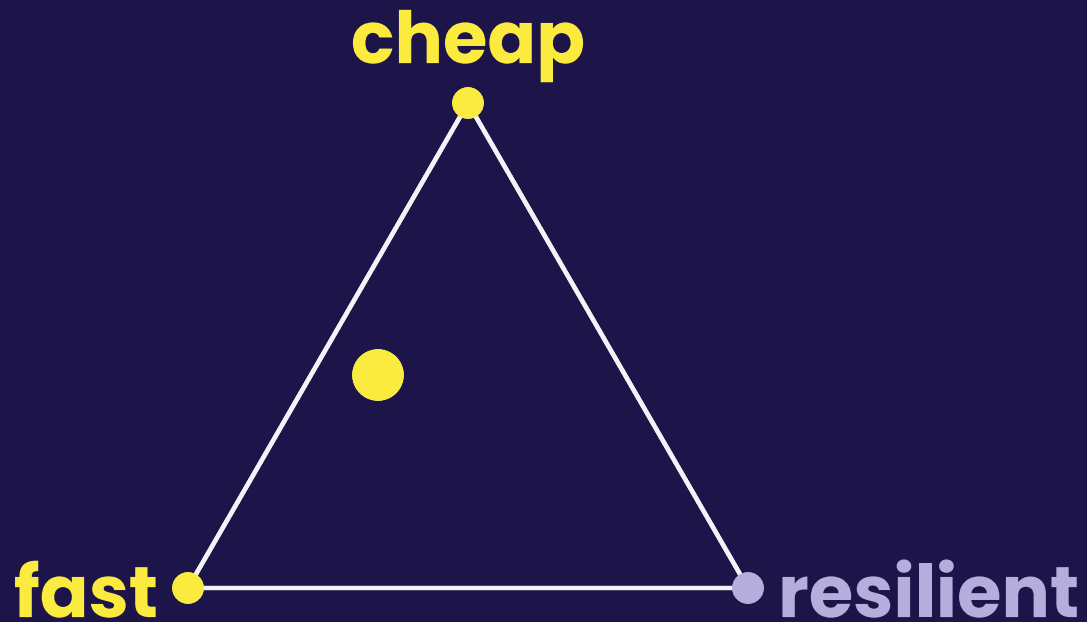


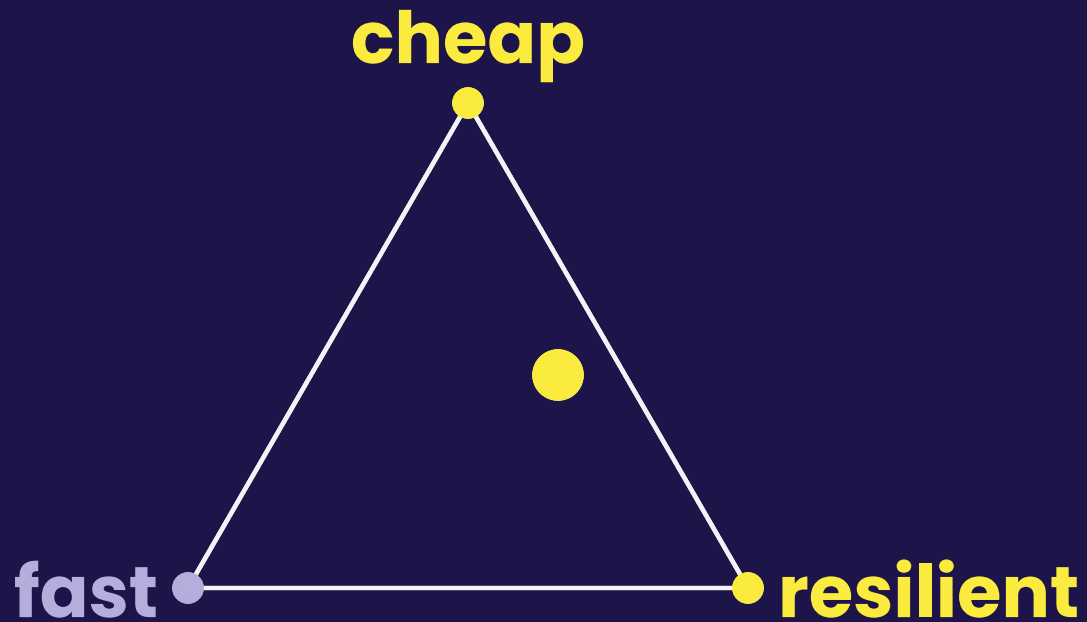
cheap

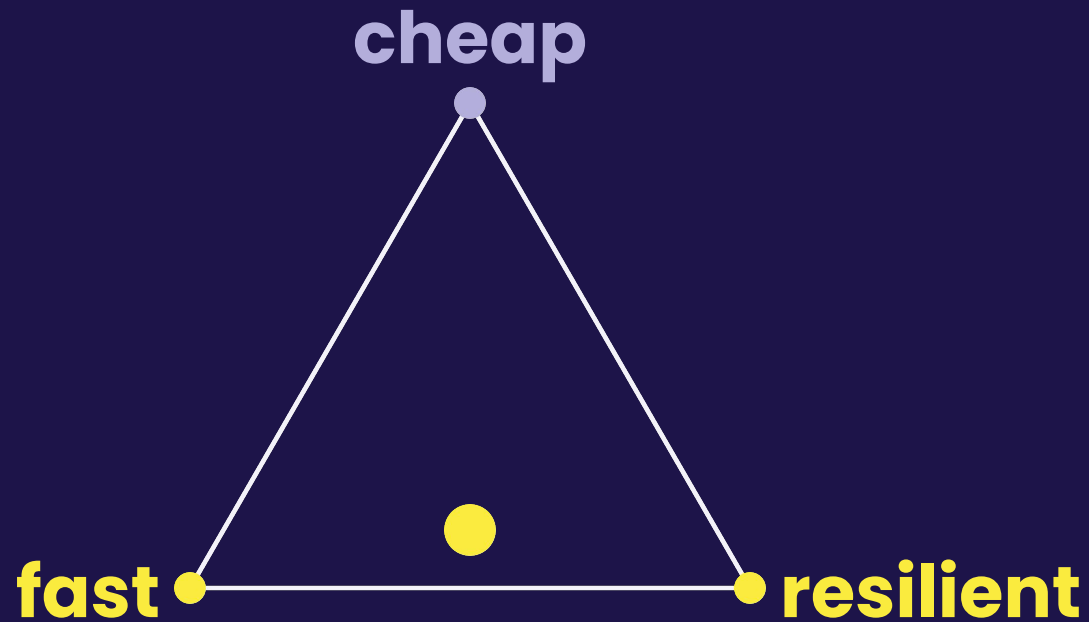


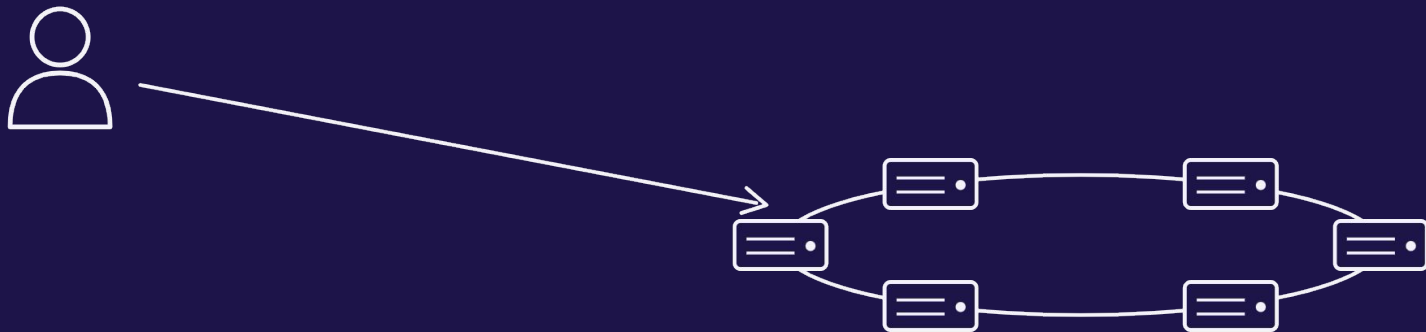


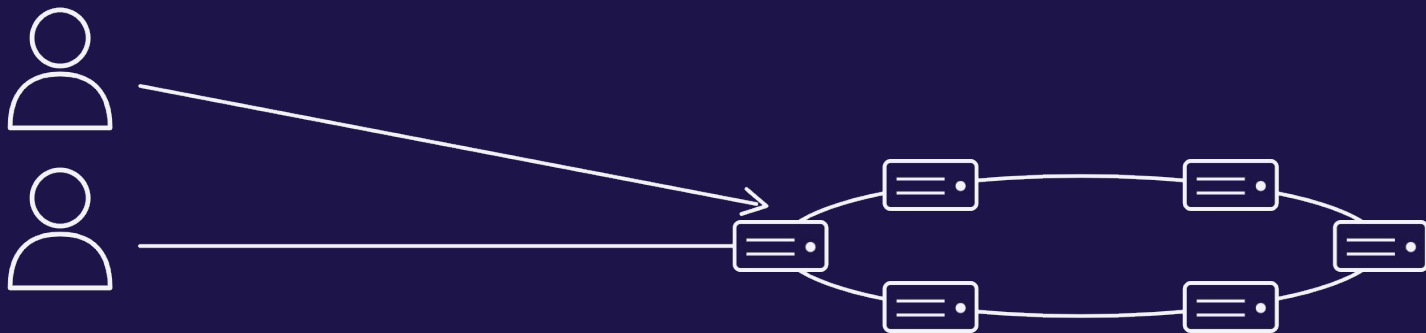


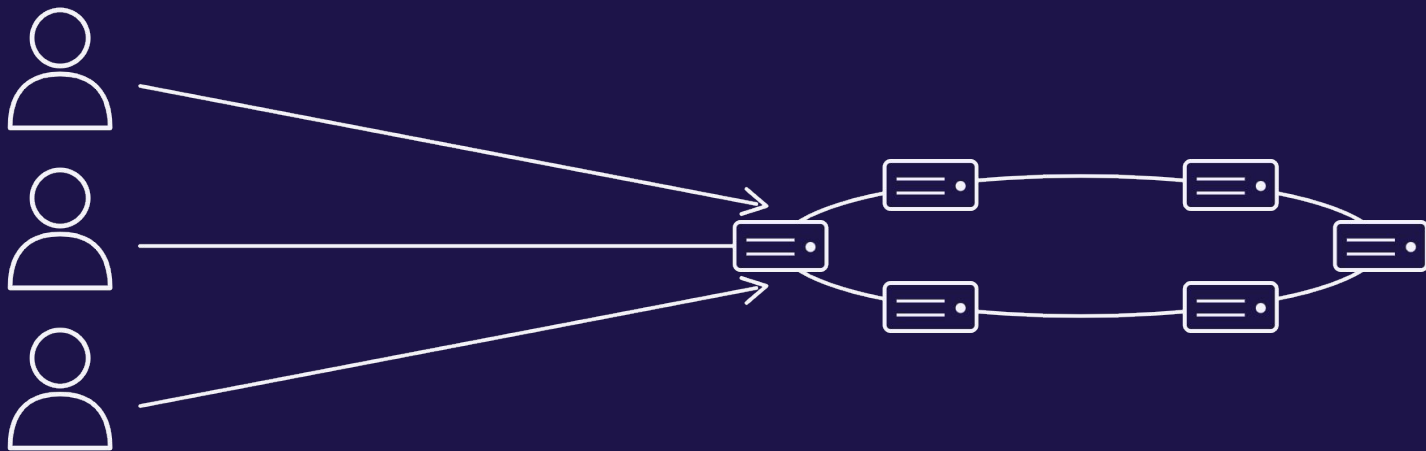


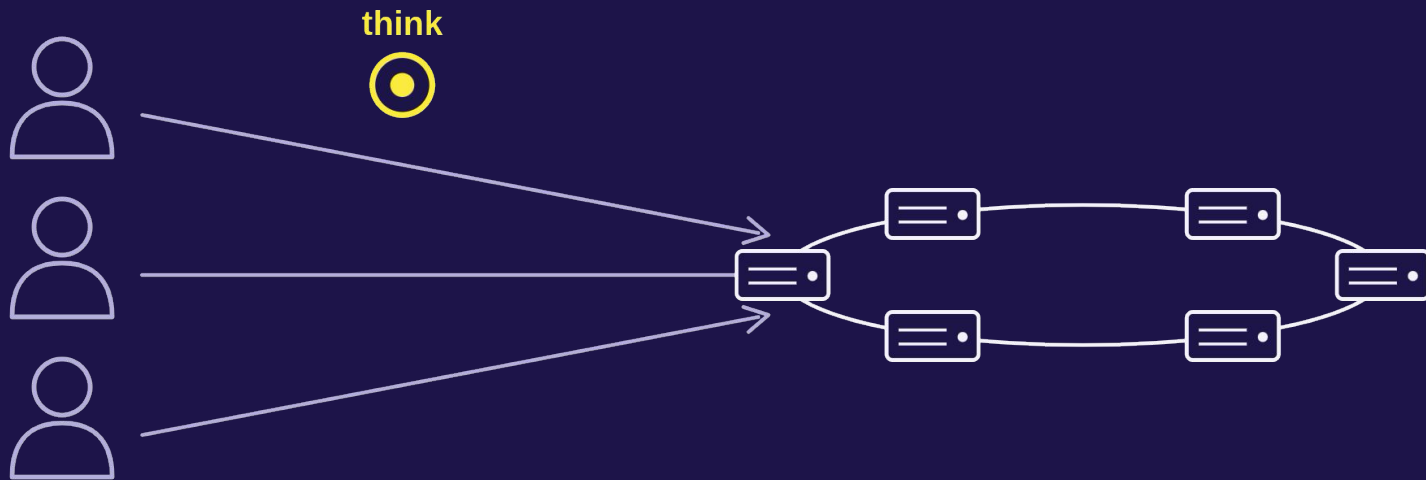


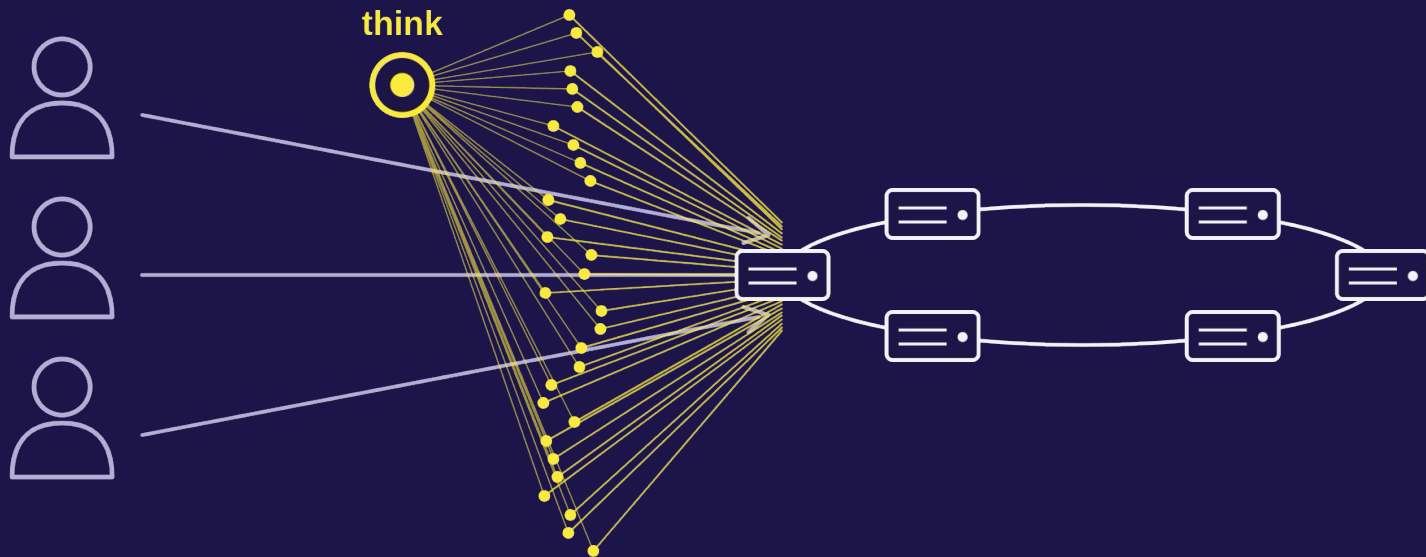


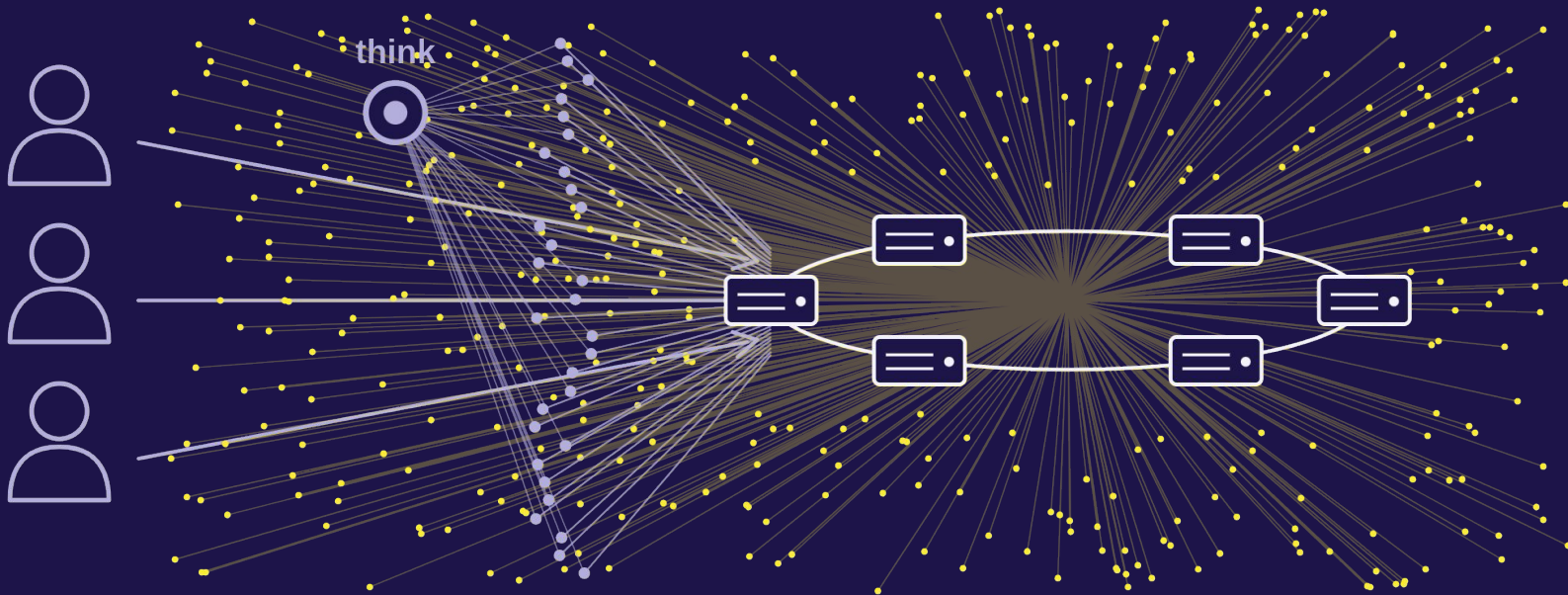


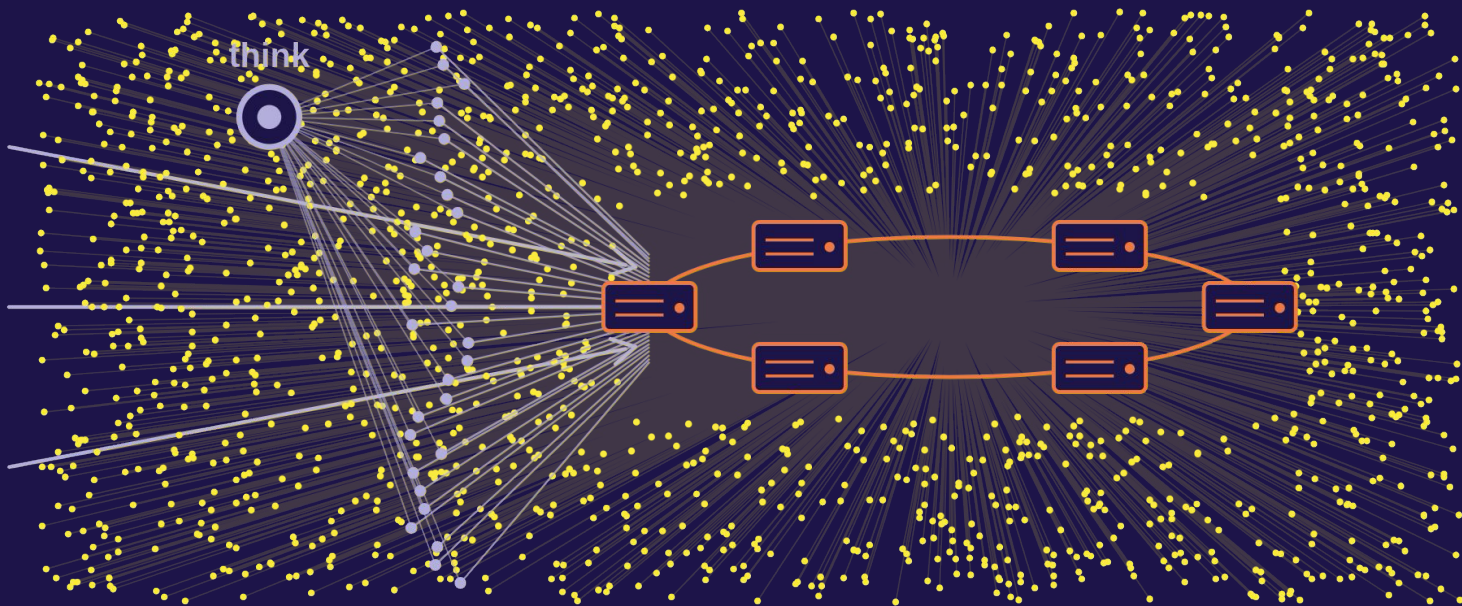












more traffic



more traffic

more complex traffic



more traffic

more complex traffic

more unpredictable traffic



capacity tier



capacity tier **→** **flow tier**



more flow, on paper



more flow, on paper

better network



more flow, on paper

better network

flash instead of HDD



more flow, on paper

better network

flash instead of HDD

chase the shiny number



more flow, in real life



more flow, **in real life**

applications



more flow, **in real life**

applications

network



more flow, **in real life**

applications

network

firewall



more flow, **in real life**

applications

network

firewall

load balancer



more flow, **in real life**

applications

network

firewall

load balancer

site topology



more flow, **in real life**

applications

network

firewall

load balancer

site topology

storage



more flow, in real life

applications

network

firewall

load balancer

site topology

storage

NVIDIA



more flow, in real life

applications

network

firewall

load balancer

site topology

storage

NVIDIA

+ Mellanox

a network company, 2020



more flow, **in real life**

applications

network

firewall

load balancer

site topology

storage



flow



flow =



flow = concurrency



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



$$\text{flow} = \frac{\text{concurrency} \times 10}{\text{latency}}$$



$$\text{flow} = \frac{\text{concurrency} \times 10}{\text{latency} \times 50}$$



$$\text{flow} = \frac{\text{concurrency} \times 10}{\text{latency} \times 50} = \text{less flow}$$



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

Scality RING



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

Scality RING

sites 3



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

Scality RING

sites 3

servers 150



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

Scality RING

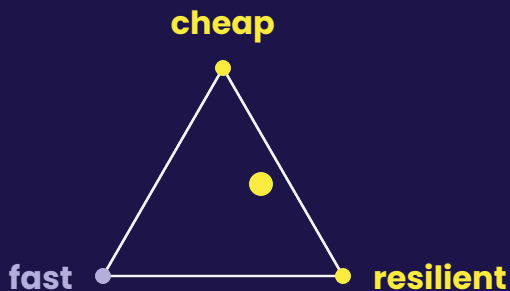
sites 3

servers 150

drives 10,000 HDD



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



Scality RING

sites 3

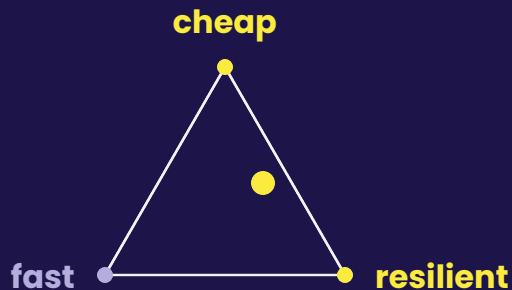
servers 150

drives 10,000 HDD



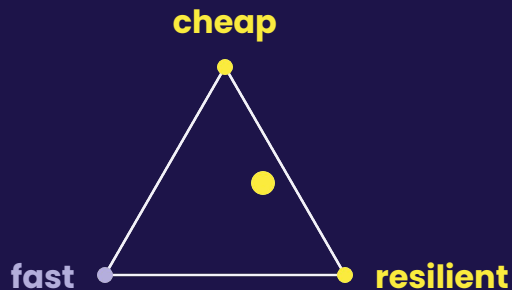
$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

each layer adds latency
+ random reads



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

each layer adds latency
+ random reads
+ TLS



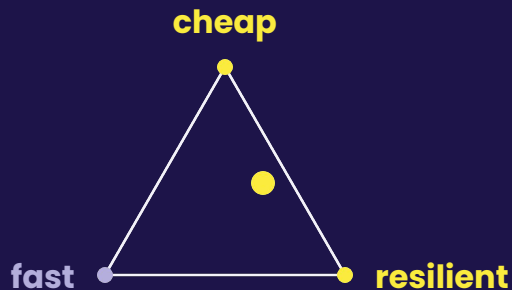
$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

each layer adds latency

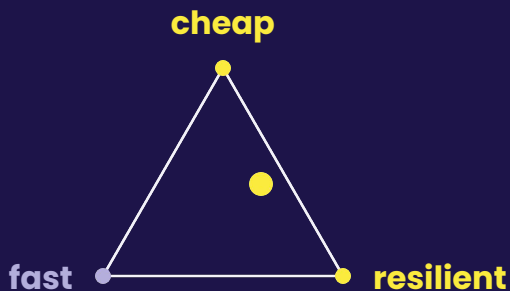
+ random reads

+ TLS

+ firewall



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



each layer adds latency

+ random reads

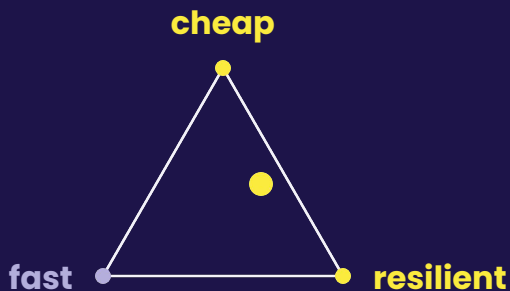
+ TLS

+ firewall

+ load balancer



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



each layer adds latency

+ random reads

+ TLS

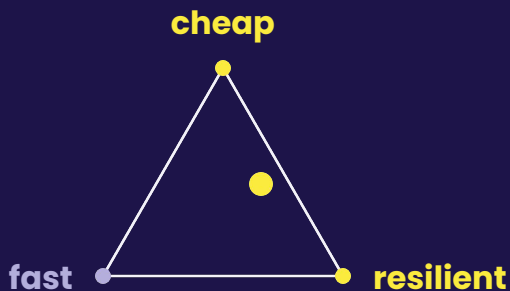
+ firewall

+ load balancer

+ three sites



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



each layer adds latency

+ random reads

+ TLS

+ firewall

+ load balancer

+ three sites

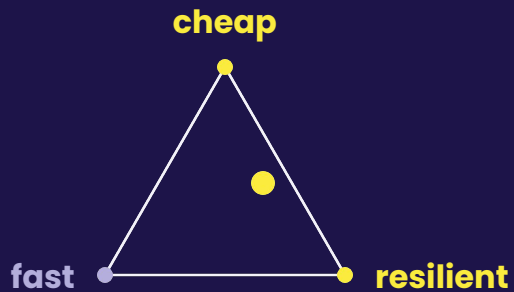
→ more concurrency



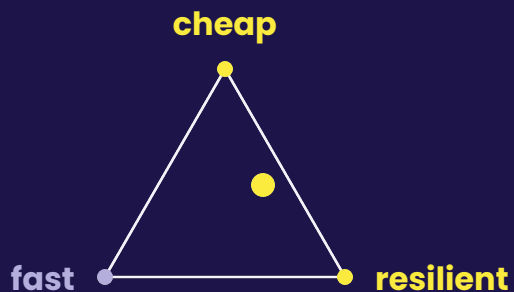
$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

asked

2 Tb/s



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



asked

2 Tb/s

delivered in real life

3.3 Tb/s



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

Scality ADI



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

Scality ADI
for KV cache



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

Scality ADI
for KV cache

sites 1



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

Scality ADI for KV cache

sites	1
servers	6



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

Scality ADI for KV cache

sites	1
servers	6
drives	60 NVMe



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

Scality ADI for KV cache

sites	1
servers	6
drives	60 NVMe
network	400G



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

S3 over RDMA



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

S3 over RDMA

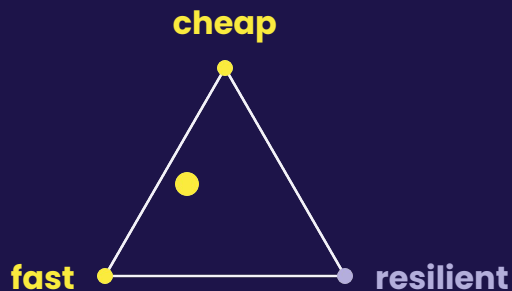
Scality AI Connector



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

S3 over RDMA

Scality AI Connector

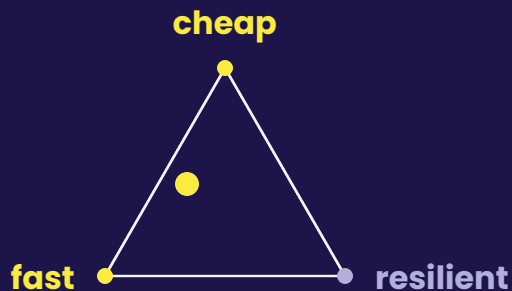


$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$

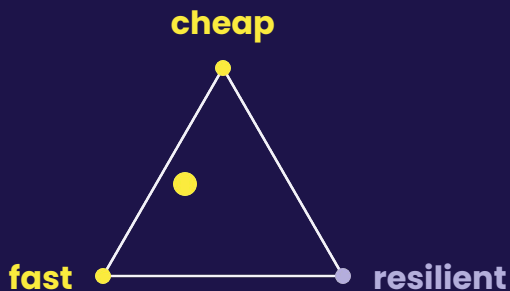
time to first token

S3 over RDMA

Scality AI Connector



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



time to first token

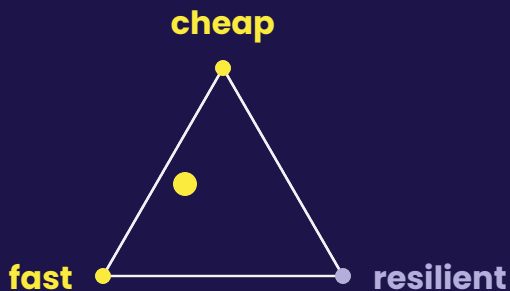
S3 over RDMA

250 ms

Scality AI Connector



$$\text{flow} = \frac{\text{concurrency}}{\text{latency}}$$



time to first token

S3 over RDMA

250 ms

Scality AI Connector

166 ms

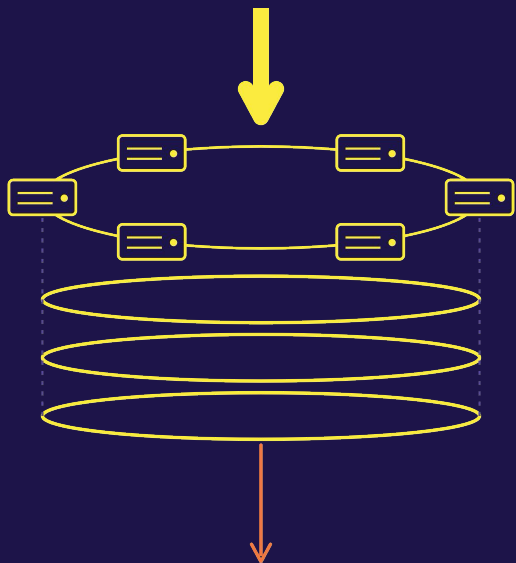


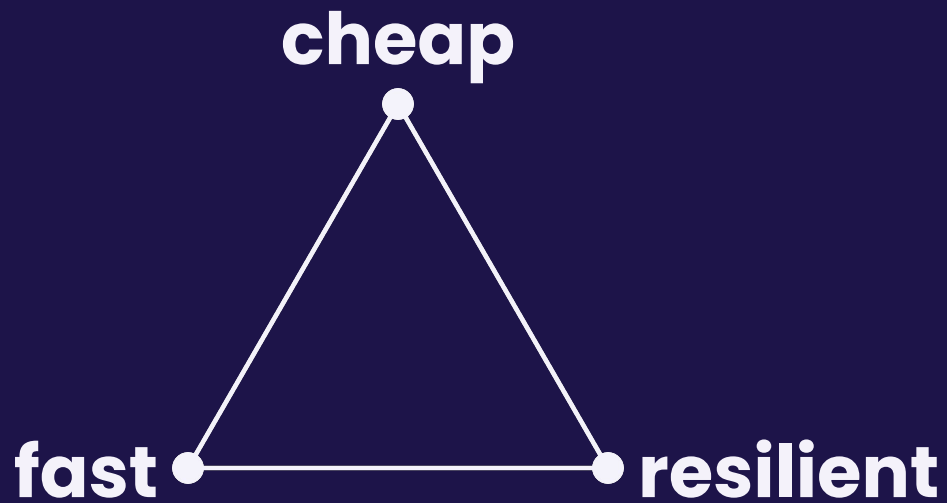
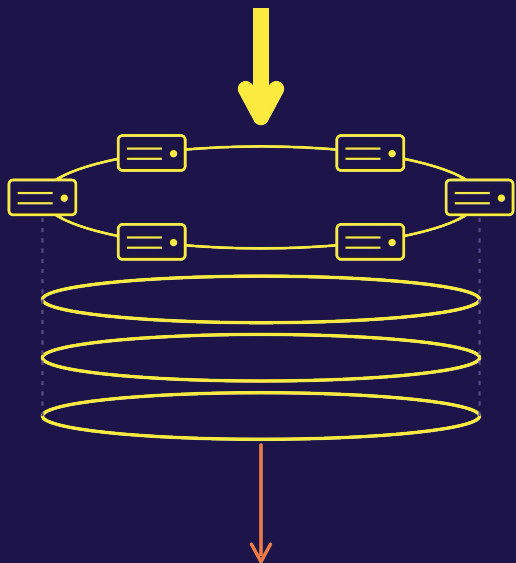
97%

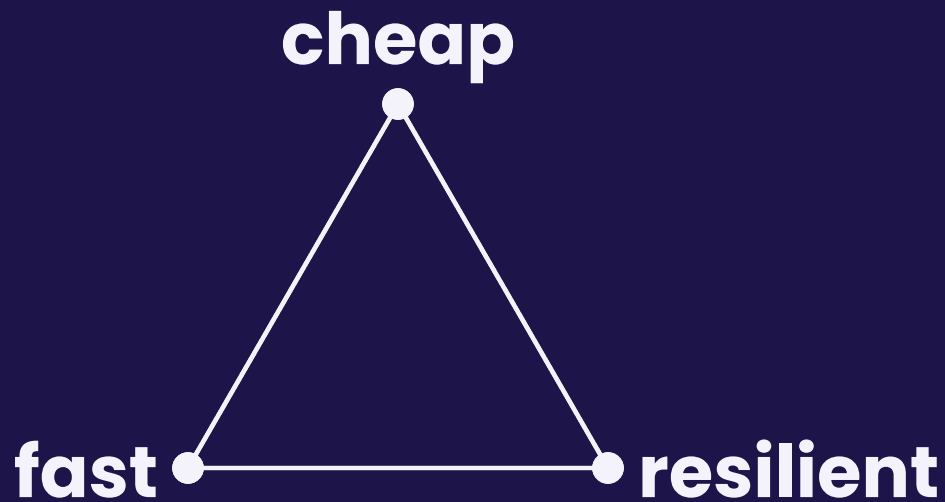
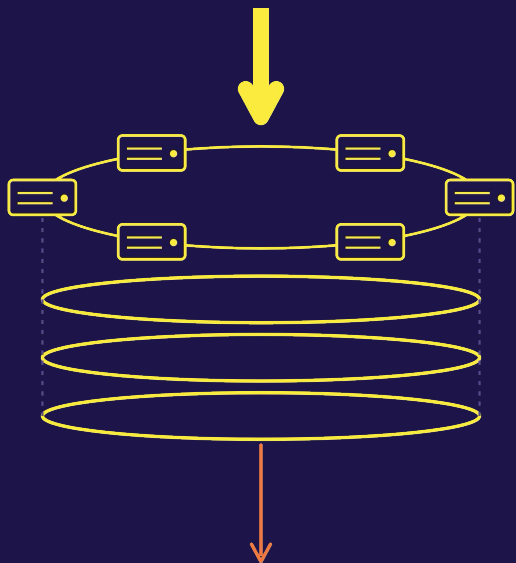


97% × 400G









real life



How long to read our largest full data set?



How long to read our largest full data set?

in real life, not on paper



1 hour?



1 hour?

1 month?



Capacity



Capacity

what you can keep



Capacity
what you can keep

Flow



Capacity
what you can keep

Flow
what you can ask





SCALITY

RELIABLE. SECURE. SUSTAINABLE.

Scality Day 2026

How **fast** can your company think?

Charles Prost, VP Engineering

Scality

8 October 2026





SCALITY

RELIABLE. SECURE. SUSTAINABLE.

Scality Day 2026

How **fast** can your company **think**?

Charles Prost, VP Engineering

Scality

8 October 2026





THANK YOU

SCALITY
DAY26