CLOUDFLARE

Атаки на скорости ИИ

Как защищаться, когда ИИ есть у всех

Alexander Kogan
Senior Territory Account Manager
Cloudflare

GET STARTED

CLOUDFLARE

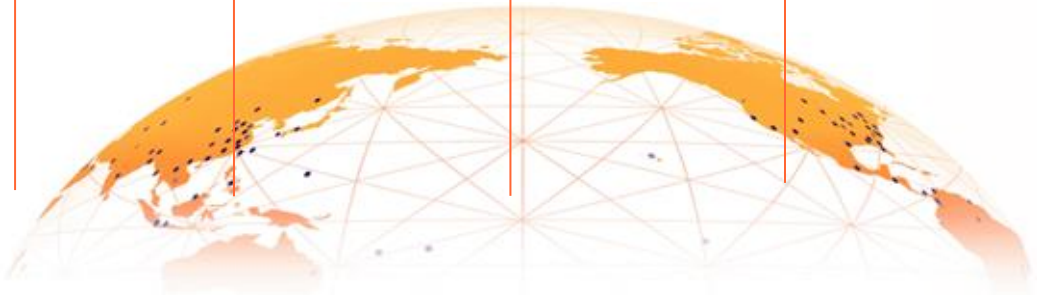
Ready to simplify security and modernize connectivity with Cloudflare One?

300+ cities
delivering all SASE services;
>3x more than other SASE vendors

~13,000 networks
directly connect to Cloudflare including major ISPs, cloud services, and enterprises

Trusted by **80%** of the Top 50 **GenAI** companies

Protects **20%** of the **Web**



One platform

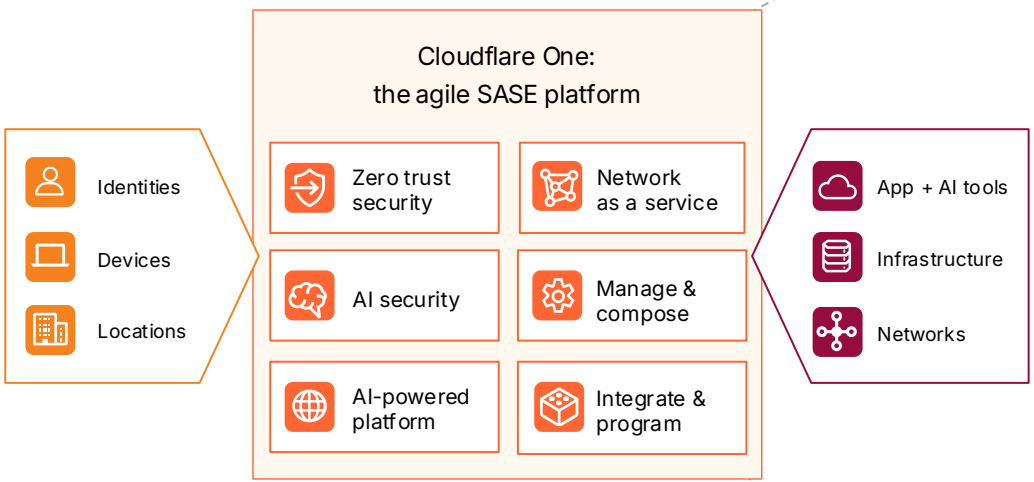
One network

Talk to our team to start a ...

- Phishing retro scan
- Cloudflare Test Drive
- Architecture workshop
- Business value assessment

SOLUTION

Cloudflare One SASE services



Identities	Devices	Locations
Human users AI agents / non-human services	Managed Unmanaged IoT	Offices Remote WiFi / cellular

App+ AI tools	Infrastructure	Networks
SaaS and GenAI IaaS and PaaS Self-hosted	Servers Databases Storage	Internet WAN/LAN Cloud

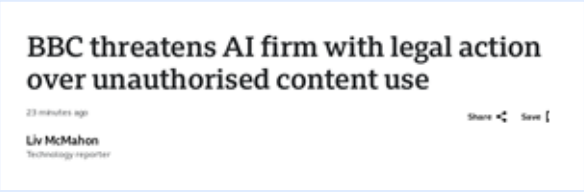
Zero trust security	Network as a service
Zero trust network access Secure web gateway Cloud access security broker Data loss prevention Remote browser isolation Email security	Composable on-ramps / connectivity Tunnel & traffic routing Firewall as a service Load balancing DNS & IP address management Post-quantum cryptography

AI security	Manage & compose
AI usage controls & discovery Data & intent guardrails AI agent governance MCP server access controls	Digital experience monitoring Logging, observability & forensics One UI, API & Terraform provider Data localization suite

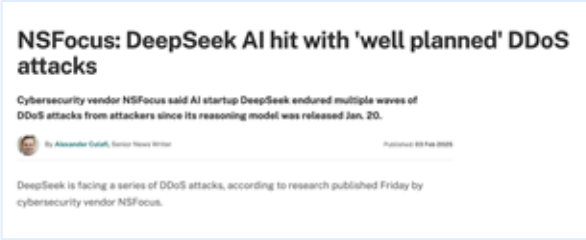
AI-powered platform	Integrate & program
Threat detection models 'Cloudy' AI agent orchestration Data loss prevention models Context awareness	Microsoft CrowdStrike & SentinelOne Okta & Ping Identity Programmable Workers

AI risk is here, now











Investor Pressure

64%

of CEOs under pressure to accelerate AI adoption

Universal Urgency

97%

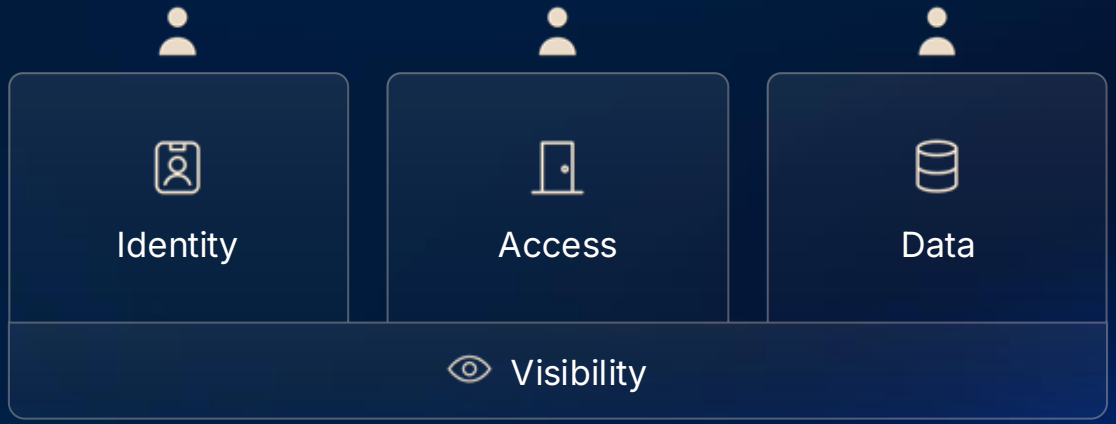
of executives feel urgency to adopt AI

High Cost of Inaction

$4.5M

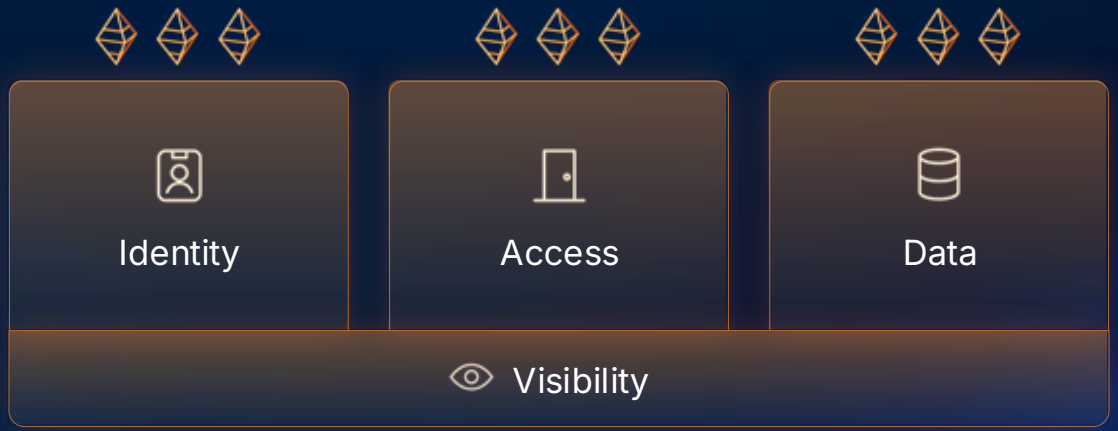
average cost of an AI incident per Gartner

The principles we applied to humans



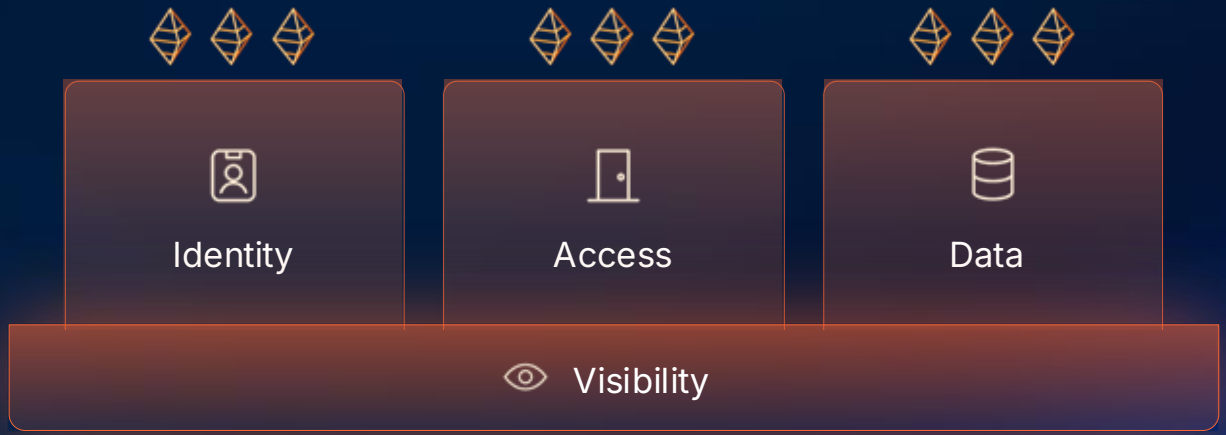
CLOUDFLARE

Same principles. Implemented differently.



CLOUDFLARE

The foundation: visibility



CLOUDFLARE

With an agent, we see every step



CLOUDFLARE

That’s visibility we never had for people

The foundation for governing agents

CLOUDFLARE

Identity

Access

Data

VISIBILITY agent-requests.log — insider-risk-agent · for rohit · 3 subagents · 13 requests · 6 seconds

```
09:14:02 agent=insider-risk-agent for=rohit evt=task.start req="triage IR-2291: mass download"
09:14:02 agent=insider-risk-agent for=rohit evt=llm.invoke model=model-large status=200
09:14:03 agent=insider-risk-agent evt=policy.check action=agent.spawn result=allow
09:14:03 agent=insider-risk-agent evt=agent.spawn children=3 subagents
09:14:03 agent=logs-subagent for=rohit evt=mcp.call tool=logs.query user=user-4471 range=30d status=200
09:14:03 agent=access-subagent for=rohit evt=mcp.call tool=access.list user=user-4471 status=200
09:14:04 agent=escalation-subagent for=rohit evt=mcp.call tool=case.create case=IR-2291 status=201
09:14:04 agent=logs-subagent for=rohit evt=mcp.call tool=mail.search scope=all-users status=200
09:14:04 agent=access-subagent for=rohit evt=mcp.call tool=access.set user=user-4471 level=none status=200
09:14:04 agent=escalation-subagent for=rohit evt=http.post url=files-drop.example/upload bytes=1.9G status=200
09:14:06 agent=insider-risk-agent for=rohit evt=llm.invoke model=model-large status=200
09:14:07 agent=insider-risk-agent evt=agent.exit children=3
09:14:08 agent=insider-risk-agent for=rohit evt=mcp.call tool=case.escalate to=human-review status=200
```

CLOUDFLARE

Identity

Cloudflare Access | MCP Server Portals

- Is it a registered agent with a valid token?
- Which task is the token scoped to?
- Which user is it acting for?

Access

Data

VISIBILITY agent-requests.log

insider-risk-agent · for rohit · 3 subagents · 13 requests · 6 seconds

```
09:14:02 agent=insider-risk-agent for=rohit evt=task.start req="triage IR-2291: mass download"  ✓
09:14:02 agent=insider-risk-agent for=rohit evt=llm.invoke model=model-large status=200  ✓
09:14:03 agent=insider-risk-agent evt=policy.check action=agent.spawn result=allow  ✓
09:14:03 agent=insider-risk-agent evt=agent.spawn children=3 subagents  ✓
09:14:03 agent=logs-subagent for=rohit evt=mcp.call tool=logs.query user=user-4471 range=30d status=200  ✓
09:14:03 agent=access-subagent for=rohit evt=mcp.call tool=access.list user=user-4471 status=200  ✓
09:14:04 agent=escalation-subagent for=rohit evt=mcp.call tool=case.create case=IR-2291 status=201  ✓
09:14:04 agent=logs-subagent for=rohit evt=mcp.call tool=mail.search scope=user-4471 status=200 action=limited  ↓
09:14:04 agent=access-subagent for=rohit evt=mcp.call tool=access.set user=user-4471 level=none status=200  ✓
09:14:04 agent=escalation-subagent for=rohit evt=http.post url=files-drop.example/upload bytes=1.9G status=200  ✓
09:14:06 agent=insider-risk-agent for=rohit evt=llm.invoke model=model-large status=200  ✓
09:14:07 agent=insider-risk-agent evt=agent.exit children=3  ✓
09:14:08 agent=insider-risk-agent for=rohit evt=mcp.call tool=case.escalate to=human-review status=200  ✓
```

CLOUDFLARE

Identity

Cloudflare Access | MCP Server Portals

Access

MCP Server Portals | Secure Web Gateway | Browser Isolation | AI Gateway

- What system does it need access to?
- Is it authorized for this MCP server, read or write?
- Which model can it call, within what budget?

Data

VISIBILITY agent-requests.log — insider-risk-agent · for rohit · 3 subagents · 13 requests · 6 seconds

```
09:14:02 agent=insider-risk-agent for=rohit evt=task.start req="triage IR-2291: mass download"
09:14:02 agent=insider-risk-agent for=rohit evt=llm.invoke model=model-large status=200 ✓
09:14:03 agent=insider-risk-agent evt=policy.check action=agent.spawn result=allow ✓
09:14:03 agent=insider-risk-agent evt=agent.spawn children=3 subagents
09:14:03 agent=logs-subagent for=rohit evt=mcp.call tool=logs.query user=user-4471 range=30d status=200 ✓
09:14:03 agent=access-subagent for=rohit evt=mcp.call tool=access.list user=user-4471 status=200 ✓
09:14:04 agent=escalation-subagent for=rohit evt=mcp.call tool=case.create case=IR-2291 status=201 ✓
09:14:04 agent=logs-subagent for=rohit evt=mcp.call tool=mail.search scope=user-4471 status=200 action=limited ↓
09:14:04 agent=access-subagent for=rohit evt=mcp.call tool=access.set user=user-4471 level=read-only status=200 action=limited ↓
09:14:04 agent=escalation-subagent for=rohit evt=http.post url=files-drop.example/upload bytes=1.9G status=200 ✓
09:14:06 agent=insider-risk-agent for=rohit evt=llm.invoke model=model-large status=200 ✓
09:14:07 agent=insider-risk-agent evt=agent.exit children=3
09:14:08 agent=insider-risk-agent for=rohit evt=mcp.call tool=case.escalate to=human-review status=200 ✓
```

CLOUDFLARE

Identity

Cloudflare Access | MCP Server Portals

Access

MCP Server Portals | Secure Web Gateway | Browser Isolation | AI Gateway

Data

Data Loss Prevention | CASB | AI Gateway | AI Security for Apps | MCP Server Portals

- Is there a jailbreak or injection in the prompt?
- Is sensitive data going to a model or out?
- Is the destination approved?

VISIBILITY agent-requests.log

insider-risk-agent · for rohit · 3 subagents · 13 requests · 6 seconds

```
09:14:02 agent=insider-risk-agent for=rohit evt=task.start req="triage IR-2291: mass download"  ✓
09:14:02 agent=insider-risk-agent for=rohit evt=llm.invoke model=model-large status=200  ✓
09:14:03 agent=insider-risk-agent evt=policy.check action=agent.spawn result=allow
09:14:03 agent=insider-risk-agent evt=agent.spawn children=3 subagents
09:14:03 agent=logs-subagent for=rohit evt=mcp.call tool=logs.query user=user-4471 range=30d status=200
09:14:03 agent=access-subagent for=rohit evt=mcp.call tool=access.list user=user-4471 status=200
09:14:04 agent=escalation-subagent for=rohit evt=mcp.call tool=case.create case=IR-2291 status=201
09:14:04 agent=logs-subagent for=rohit evt=mcp.call tool=mail.search scope=user-4471 status=200 action=limited  ↓
09:14:04 agent=access-subagent for=rohit evt=mcp.call tool=access.set user=user-4471 level=read-only status=200 action=limited  ↓
09:14:04 agent=escalation-subagent for=rohit evt=http.post url=files-drop.example/upload bytes=1.9G status=403 action=blocked  ✗
09:14:06 agent=insider-risk-agent for=rohit evt=llm.invoke model=model-large status=200  ✓
09:14:07 agent=insider-risk-agent evt=agent.exit children=3
09:14:08 agent=insider-risk-agent for=rohit evt=mcp.call tool=case.escalate to=human-review status=200
```

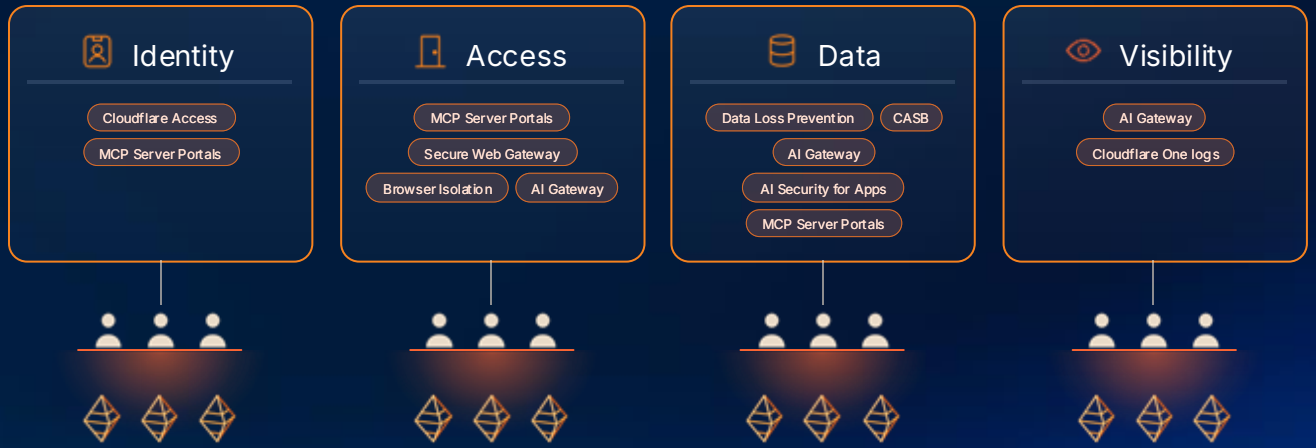
CLOUDFLARE

Applied for people



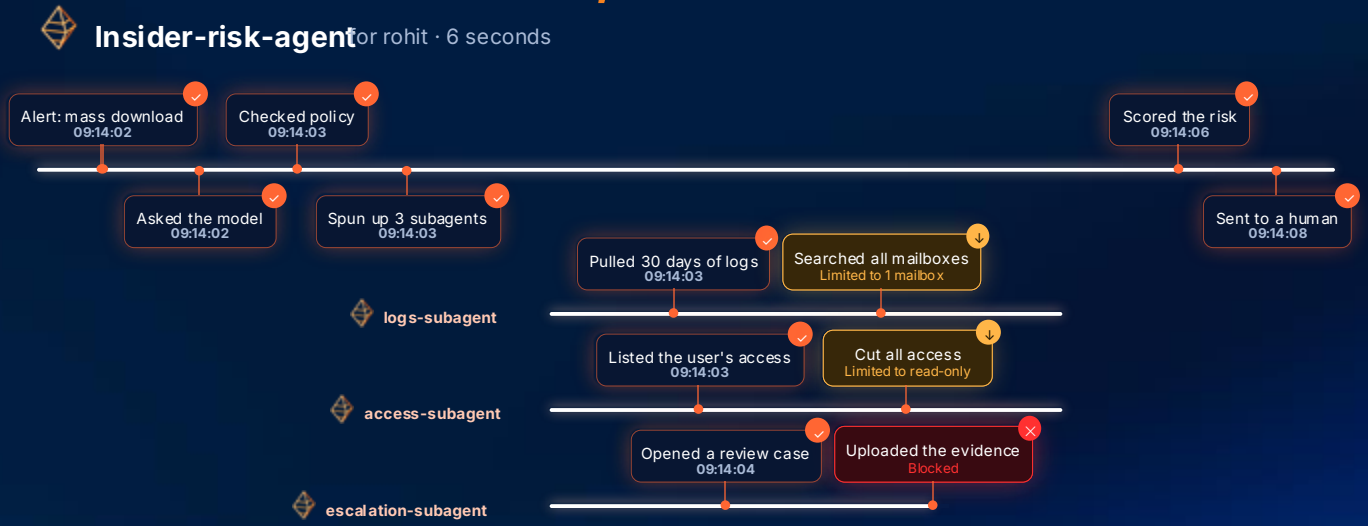
CLOUDFLARE

Extended to agents



CLOUDFLARE

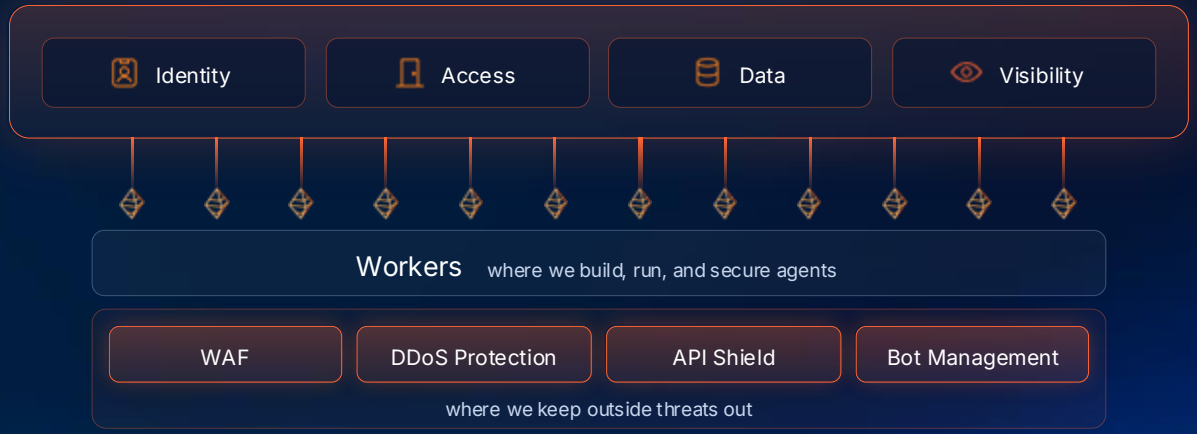
Controls on every step
Inline, real time



9 allowed 2 limited 1 blocked

CLOUDFLARE

Built by Cloudflare
Built on Cloudflare



CLOUDFLARE

Built for us

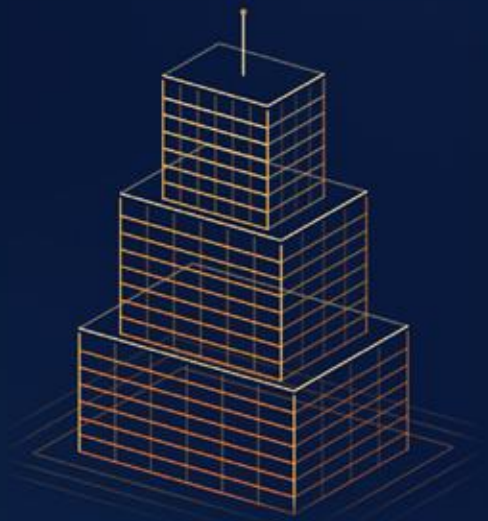


CLOUDFLARE

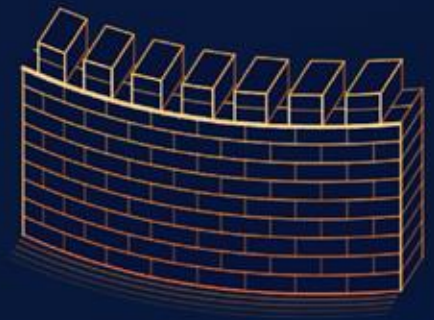
Ready for you



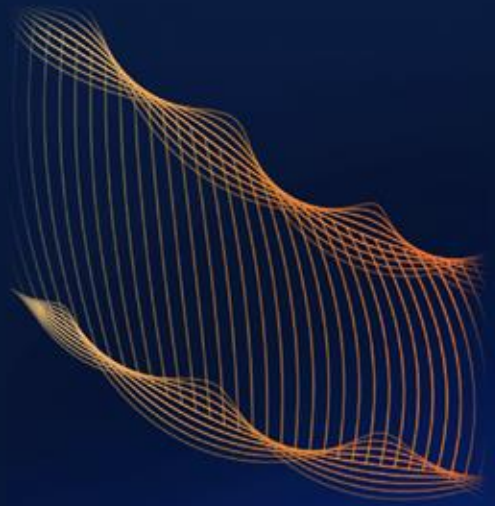
CLOUDFLARE



01 Build



02 Secure



03 Scale

CLOUDFLARE

CLOUDFLARE

Thanks!

akogan@cloudflare.com
Alexander Kogan
Senior Territory AE
Cloudflare