



ORACLE

AI Never Sleeps

Scaling **Data Centers** and **Networks**
for the Agentic Era

Alexandru Neagu

Cloud Engineering Director @ Oracle



Nearly 50 years of
helping customers
manage and
protect the
**world's most
important data**

Surfacing sourcing risks and next-best
actions before disruption hits

Training and deploying AI models
that help improve clinical workflows

Turning regulated healthcare and
life sciences data into governed AI insight

Detecting fraud, risk, and compliance
exceptions in real time

Answering complex insurance and risk
questions in seconds instead of days

Giving field teams and service agents real-time
AI answers from operational data

Modernizing critical energy workloads for performance,
resilience, and regulatory agility

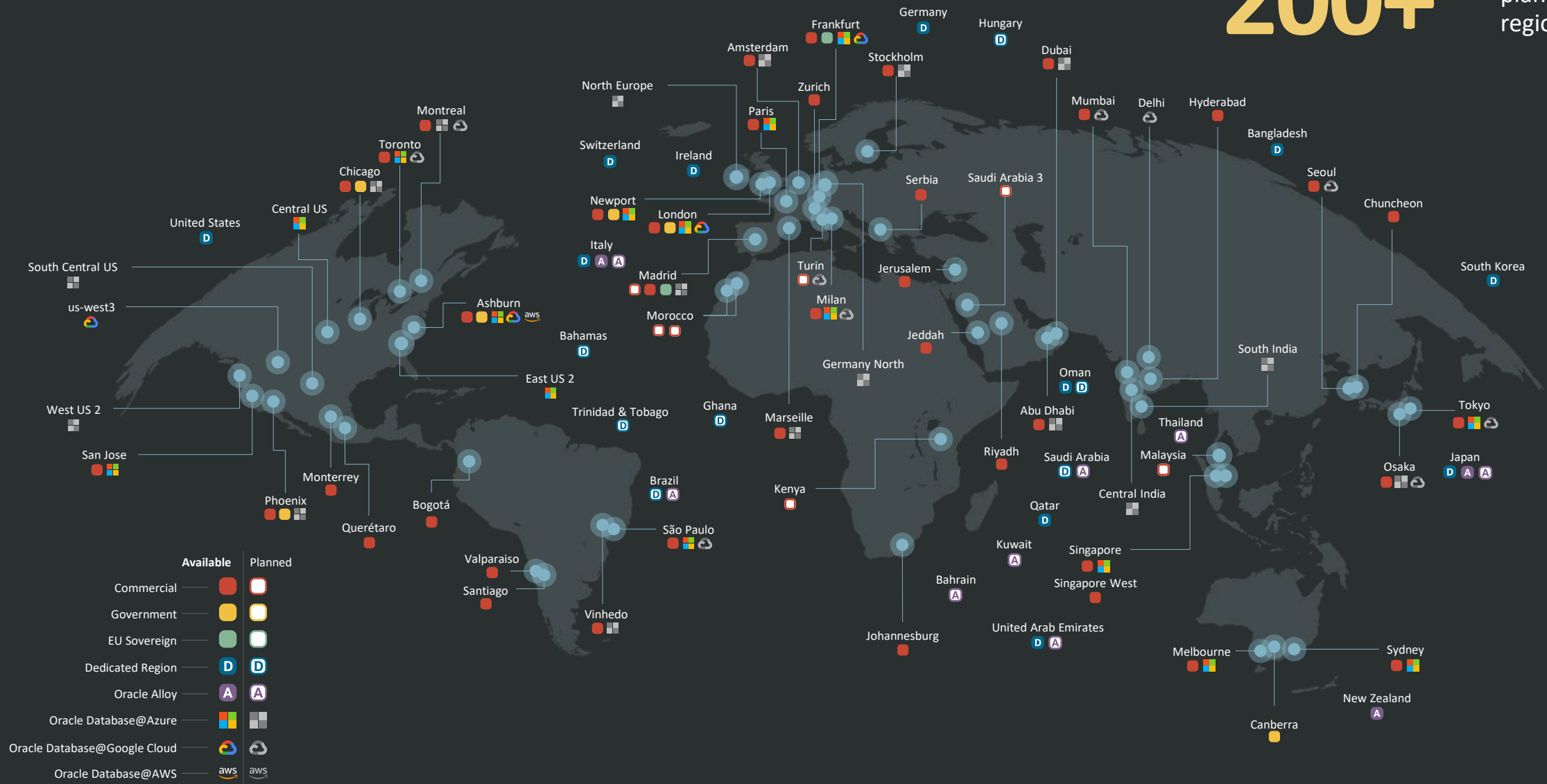
Automating finance, planning, sourcing,
and project workflows with embedded AI

AI = Data + Infrastructure + Expertise

Oracle Cloud Infrastructure

200+

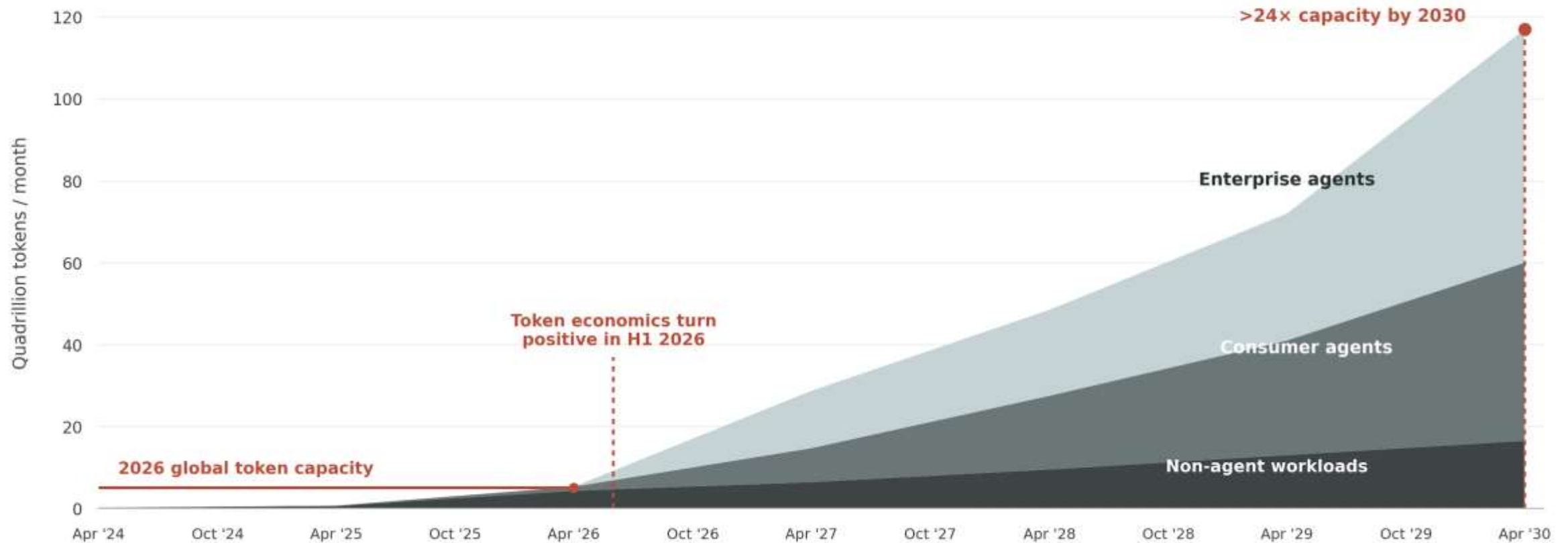
live and
planned regions¹



¹Some regions have not been announced and are therefore not shown on the map



Token use by AI agents is expected to multiply 24 times by 2030



[Source: Goldman Sachs Research](#)

What is an agent?

A model in a loop with tools and state, running until the goal is reached or a budget runs out.

1

Model

proposes: text and tool calls

2

Tools

the agent's only hands, and
your real API surface

3

Harness / loop

orchestration, state,
budgets, tracing

4

Context / memory

what the model sees

the whole idea, in pseudo-code

```
while not done and within_budget:
    response = model(instructions
                     + context + tool_schemas)
    if response.tool_calls:
        # ← your code, your rules
        results = execute(response.tool_calls)
        context += results
    else:
        done = True
```

03:12, Sunday

A cluster of cell sites starts failing. Nobody is awake. The agent is.

Let's follow one incident and count what it costs the infrastructure.

NOC · alarm feed · region North-West

03:12:04	CRIT	S1/NG link down · gNB CJ-0417
03:12:05	CRIT	Cells unavailable · CJ-0417
03:12:09	MAJ	RRC success 78% · CJ-0420
03:12:11	MAJ	MW link BER high · MW-112
03:12:13	CRIT	S1/NG link down · gNB CJ-0423
03:12:20	MIN	Rectifier alarm · CJ-0417
03:12:31	MAJ	Throughput -62% · cluster CJ-N
03:12:40	...	+1,846 correlated alarms in 5 min

Illustrative scenario. Alarm names follow 3GPP / vendor conventions.

Telco is already running agents in the network

This isn't a lab scenario. Operators are moving to autonomous networks now.

**hours → ~1
min**

Deutsche Telekom's RAN Guardian agent: time to handle a major network event; 100+ autonomous fixes in month one

Deutsche Telekom, Feb 2026

81%

of operators target Level 4 autonomous networks by 2030; 20% expect it by 2027

TM Forum survey via Light Reading, Jul 2026

51%

of enterprises already run agentic AI that acts in production; 82% accept some AI network changes without prior approval

Cisco / Omdia, Sep 2026

Oct 2026

first AI-RAN call on a commercial network: the radio network now runs AI inference too

NVIDIA, Nokia, T-Mobile, GTC Berlin

Three ways to handle the same incident

Same alarm storm. Very different load on the infrastructure behind it.

A Standard process

rules + runbook + human

- Rule engine correlates alarms
- Ticket opened, engineer paged
- Human opens 6 dashboards
- Fix in ~2 h (MTTR)

A few DB queries. Mostly idle humans.

B Naive agent

one prompt, one answer

- Dump alarms into one prompt
- Model guesses a root cause
- No verification, no memory
- Pastes summary into ticket

1–3 LLM calls. Fast, cheap, often wrong.

C Production agent

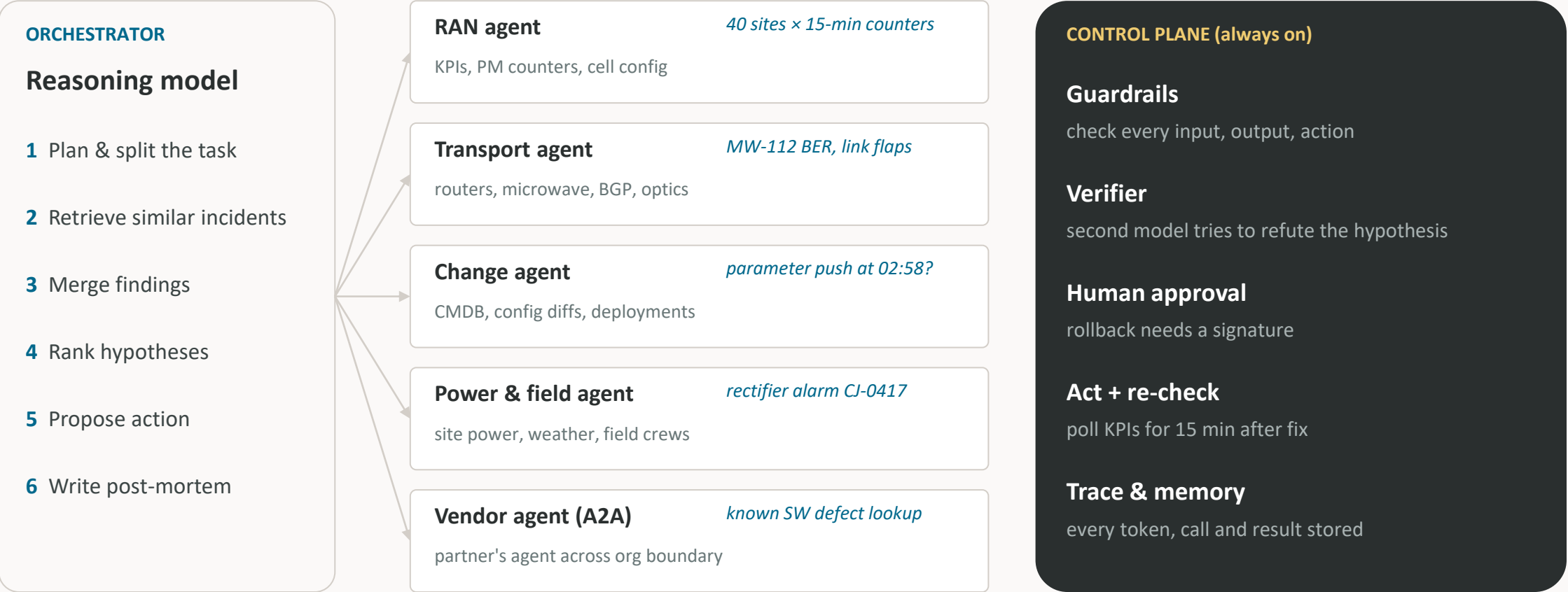
plans, investigates, verifies, acts

- Planner splits the problem
- Specialist agents query RAN, transport, power
- Verifier challenges the hypothesis
- Human approves; agent acts, re-checks KPIs

Hundreds of model and tool calls.

Inside the production agent: one investigation

Every arrow is a network call. Every box is compute. Every step leaves data behind.



Pattern: orchestrator–worker multi-agent system (Anthropic, 2025), A2A protocol (Linux Foundation), MCP for tools.



One incident, three bills

Order-of-magnitude estimate for the incident above. Your numbers will differ; the ratios won't.

	Standard process	Naive agent	Production agent	× vs naive
LLM calls	0	3	~200	~70×
Tokens processed	0	~30 K	~2 M	~70×
Tool / API calls	~10	2	~400	~200×
Telemetry scanned	~50 MB	~1 MB	~5 GB	~5,000×
Data written (traces, memory)	~1 MB	~0	~300 MB	new
Machine-to-machine messages	~20	~10	~3,000	~300×
Time to mitigation	~2 h	~1 min*	~12 min	

** fast, but nobody can trust it, so a human repeats the investigation anyway.*

Speaker's estimate: 5 sub-agents × ~30 steps, ~10 K-token average context, 40 sites of 15-min PM counters. Cf. Anthropic: multi-agent ≈ 15× chat tokens.



The industry data points the same way

COMPUTE

15×

more tokens for a multi-agent system than for a chat (single agent: ~4×

Anthropic Engineering, Jun 2025

1,000×

more tokens for agentic coding than code chat; same task varies up to 30×

Bai et al., arXiv 2604.22750, 2026

7× / year

growth in Google's token volume, to over 3.2 quadrillion tokens a month

Google I/O keynote, May 2026

NETWORK

53%

of all web traffic was automated in 2025, not human

Imperva Bad Bot Report, 2026

25%

of all WAN traffic will be AI inference by 2035, from ~0 today

Cisco, AI impact on WAN traffic, 2026

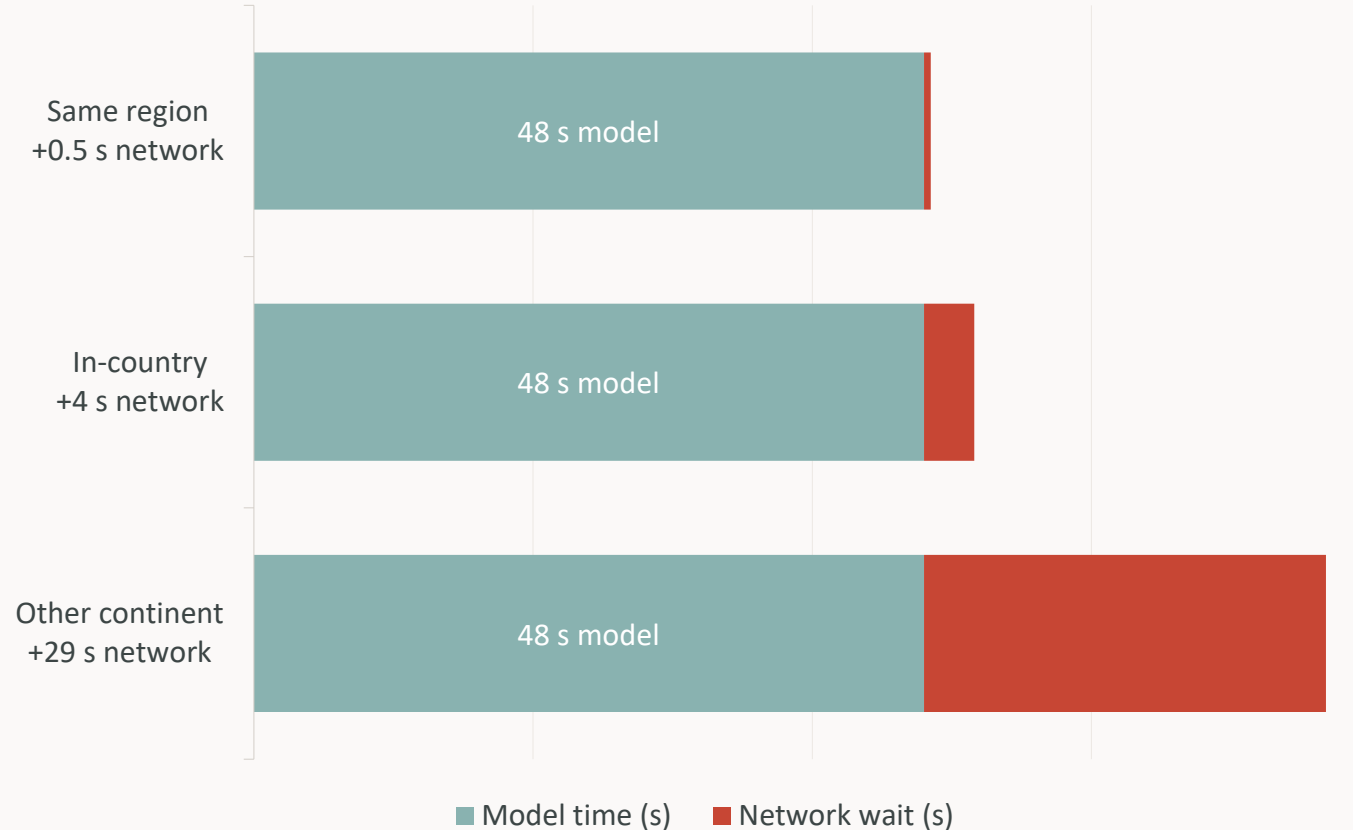
43 of 55

operators saw uplink grow faster than downlink (Q1 2026)

Ericsson Mobility Report, Jun 2026

Agents turn milliseconds into minutes

End-to-end time of one 40-step investigation vs. where the model and tools run.



Assumptions

40 sequential steps

6 network round trips per step (tool, auth, data, guardrail)

2 / 15 / 120 ms round-trip time

1.2 s model time per step

A chatbot pays network latency once. An agent pays it on every step, and every sub-agent multiplies it.

Scale inside: the data center becomes the computer

When one agent run touches thousands of GPUs and services, the fabric is the bottleneck.

SCALE-UP

Superclusters

- OCI Zettascale10: up to 800,000 NVIDIA GPUs, 16 zettaFLOPS peak
- Multi-gigawatt clusters inside a 2 km radius
- OCI + AMD: 50,000 MI450 GPUs, 432 GB HBM4 each
- Liquid-cooled, 72-GPU racks

SCALE-OUT

Oracle Acceleron networking

- Dedicated network fabrics and more direct data paths
- Converged NIC: up to 2× throughput
- Multi-planar network: traffic shifts instantly if a plane fails
- Zero-trust packet routing at the host

SERVE

Inference as a factory

- Inference $\approx$ 2/3 of AI compute in 2026
- Split prefill and decode; route requests to cached context
- One 128K-token user on a 70B model: ~40 GB of KV cache
- Context memory becomes its own storage tier

Oracle press releases, 14 Oct 2025 · Deloitte TMT Predictions 2026 · NVIDIA Technical Blog, Sep 2025 · NVIDIA BlueField-4, CES 2026. Vendor figures are company claims.

Scale out: mega campuses and small regions

Train where power is cheap. Reason where the data lives. Act where the network is.

multi-GW · train

Mega AI campuses

Frontier training, large-model serving, batch agent work.

OCI Zettascale10 superclusters

10–40 ms · reason

Public & sovereign regions

Agent orchestration, RAG, in-country data and models.

OCI public & EU Sovereign regions:
200+ live and planned

1–10 ms · data

A region in your building

Inference next to OSS/BSS and subscriber data.

Dedicated Region25 from 3 racks ·
Alloy · Cloud@Customer

1–5 ms · act

Network edge

Small models in the RAN / MEC;
closed-loop control.

AI-RAN on GPUs at the cell site

Latency bands are indicative. Sources: Oracle press releases, 14 Oct 2025; NVIDIA / Nokia AI-RAN announcement, 21 Oct 2026.

Networks get sized for machines, not people

Inside the data center, between data centers, and out to the edge.

INSIDE THE DATA CENTER

The fabric

800G → 1.6T

AI back-end switch sales overtook front-end networks for the first time (2Q26)

- Three networks: scale-up, scale-out, front-end
- Ethernet now leads AI back-end networks
- Lossless, multi-planar, zero-trust

BETWEEN DATA CENTERS

Scale-across

45%

year-on-year growth in data center interconnect optics (2Q26)

- One AI factory spread over many sites
- Lumen: 17M → ~58M fiber miles by 2031
- Over \$16B of new subsea cables due 2026–29

OUT TO THE EDGE

WAN & access

3–5×

mobile uplink traffic by 2031 vs 2025

- Software agents are “almost always on”
- Symmetric, machine-to-machine traffic
- Edge inference: AI-RAN, MEC

The network stops being plumbing. It decides GPU utilisation, agent speed, and where AI is allowed to run.

Why scaling data centers and networks matters now

The agentic era asks infrastructure for four things at once.

01

Scalable

Plan for 10–100× the demo.

Agent workloads compound:
more users, more steps, always
on.

02

Connected

**The network is part of the
model.**

Fabric, scale-across and WAN
decide speed and GPU utilisation.

03

Resilient

**Agents never sleep; neither
can we.**

Failure is normal. Multi-plane,
multi-region, resumable.

04

Trusted

**Sovereign and secure by
design.**

Agent identity, zero-trust routing,
budgets, full traces.

AI never sleeps.

The data centers and networks underneath it can't afford to blink.

Thank you. Questions?

Alexandru Neagu · Oracle

Slides & sources: <https://neagu.cloud>