The Way to Edge AI

Alisson Sol

March 6th, 2026

As telecommunications evolve toward 6G, Radio Access Networks (RAN) will fundamentally transform application architecture beyond today's approach of endpoints calling LLMs on remote data centers.

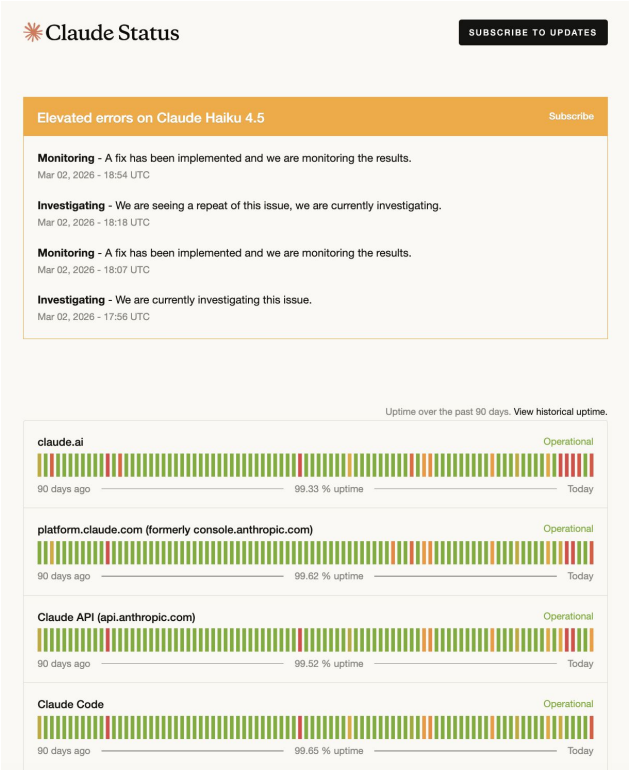
This presentation examines three critical software development challenges for the next decade:

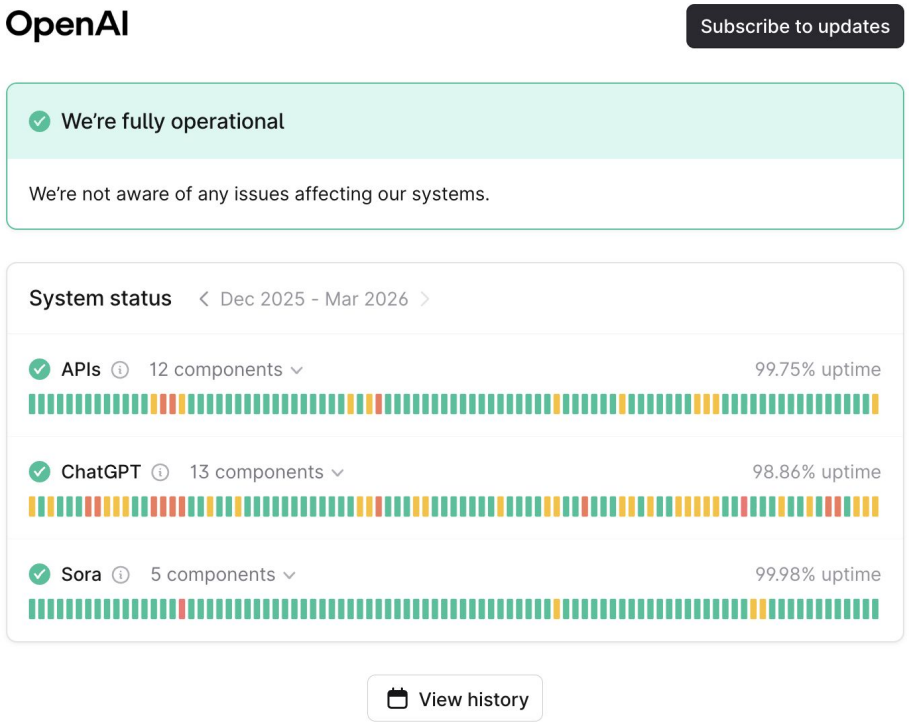
- **Real-time AI at the Edge**: Moving from cloud-dependent inference to distributed processing meeting RAN's strict latency requirements.
- **Multi-tenant Resource Optimization**: Transitioning to shared platforms where AI workloads coexist while maintaining service-level agreements.
- **Federated Learning and Privacy**: Shifting from centralized training to federated approaches preserving privacy across distributed networks.

The goal is sharing a roadmap for AI-native infrastructure where intelligence lives at the network edge.

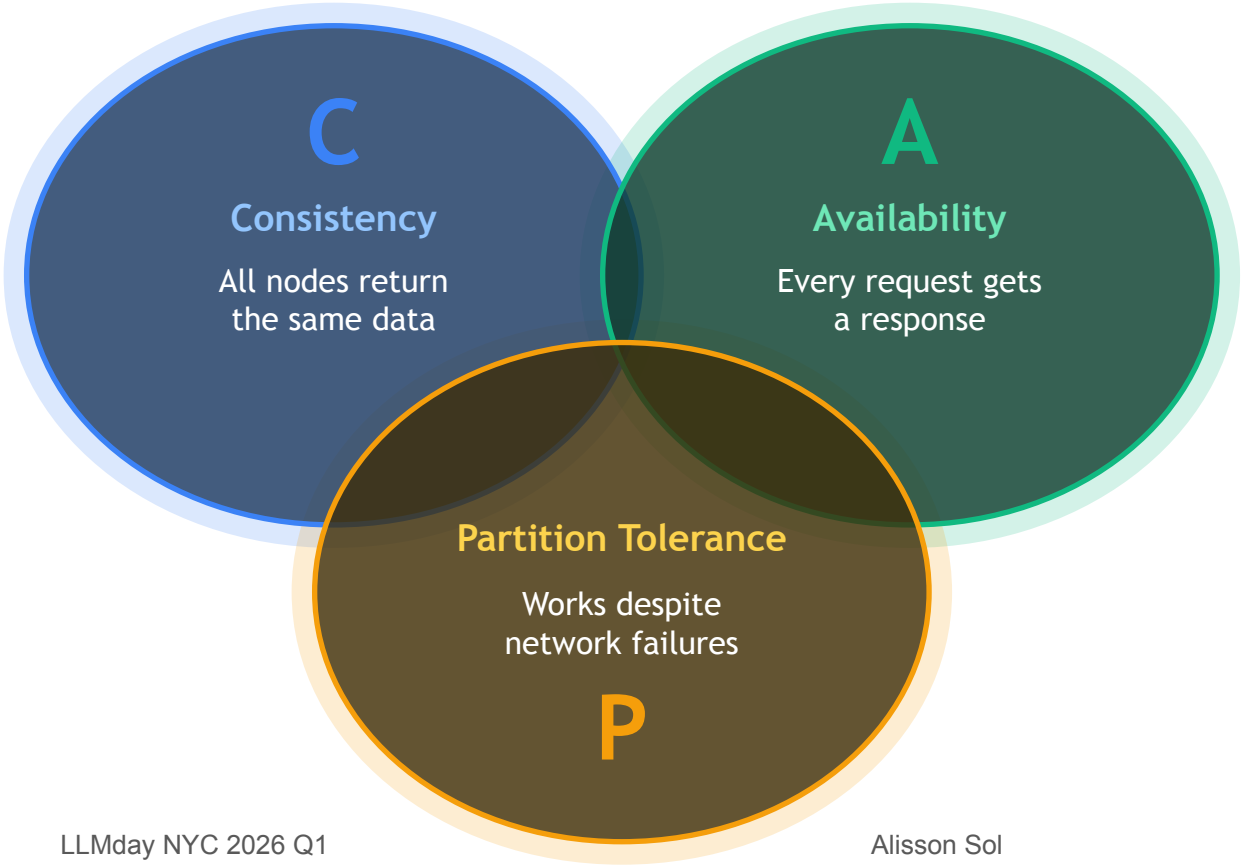
LLMday NYC 2026 Q1

Why Edge AI?





CAP Theorem



PACELC

PAC (If there is a Partition...)
- A (Availability) or
- C (Consistency)

ELC (...Else...)
- L (Low-Latency) or
- C (Consistency)

v2026-03-06c

Why CAP matters?

🚗**Autonomous Vehicle Routing** → Availability + Low-Latency (PACELC: A, L).
Must respond in <1ms even if network is partitioned. Stale map data is OK; a missed turn is not.

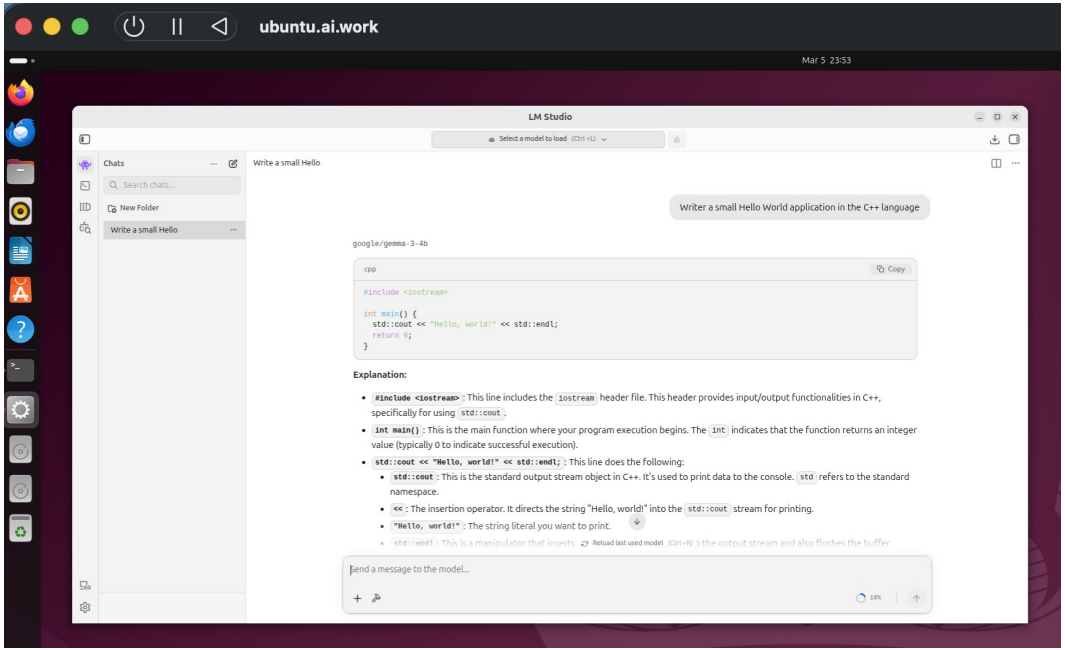
🧠**Federated Model Aggregation** → Consistency + Latency Tolerance (PACELC: C, L-tolerant).
Model sync lag of seconds is acceptable. All nodes must eventually agree on weights.

⚡**Multi-tenant SLA Scheduling** → Availability + Low-Latency (PACELC: A, L).
Better to serve with a stale SLA policy than block all tenants awaiting a consistent state update.

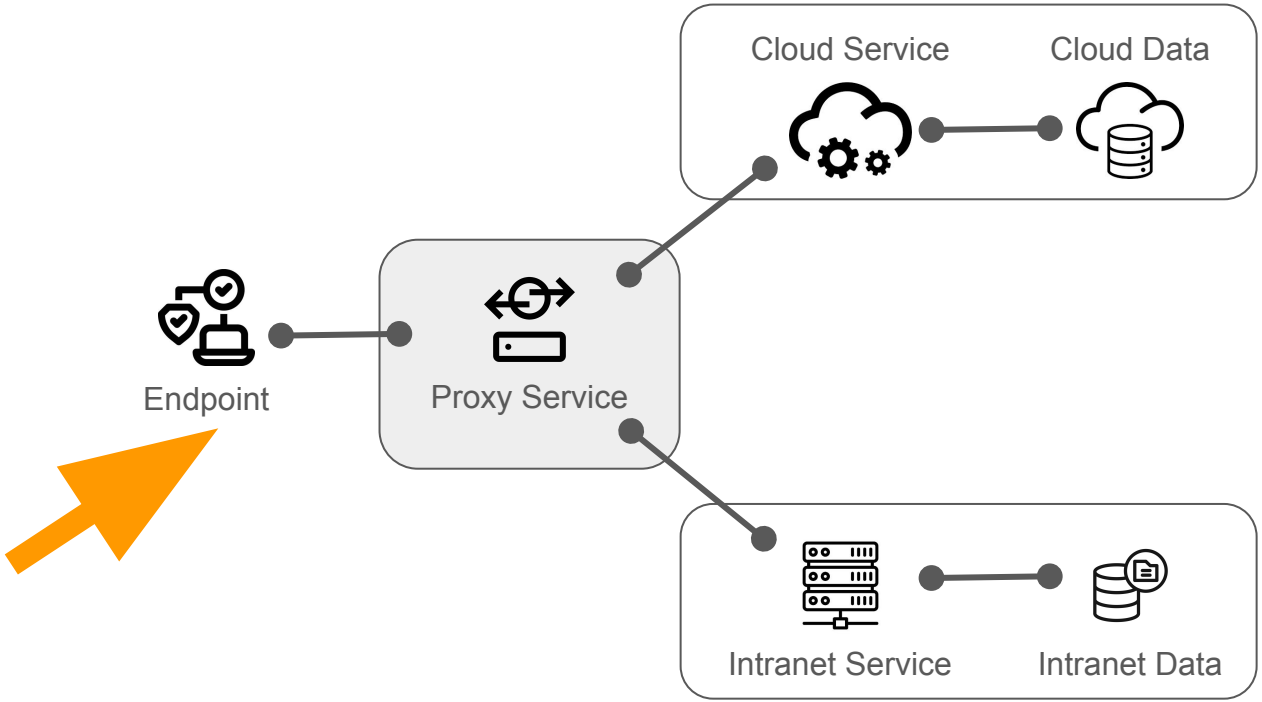
Key insight: 6G's 0.1ms latency target makes the "L" in PACELC the dominant constraint for real-time inference. Architects must design for partition explicitly, not as an afterthought.

Meanwhile, locally…

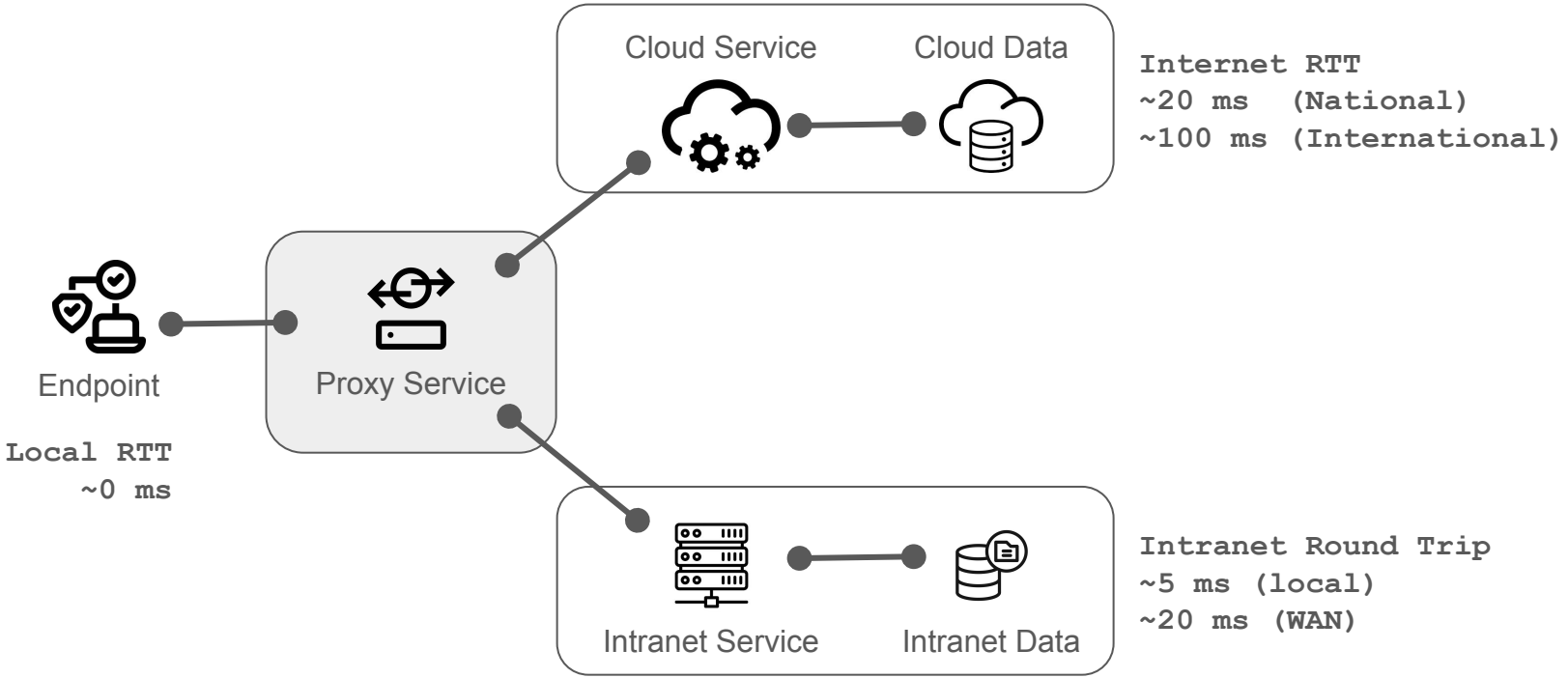
- Try our Virtual Development Environment (VDE) at https://yuruna.com



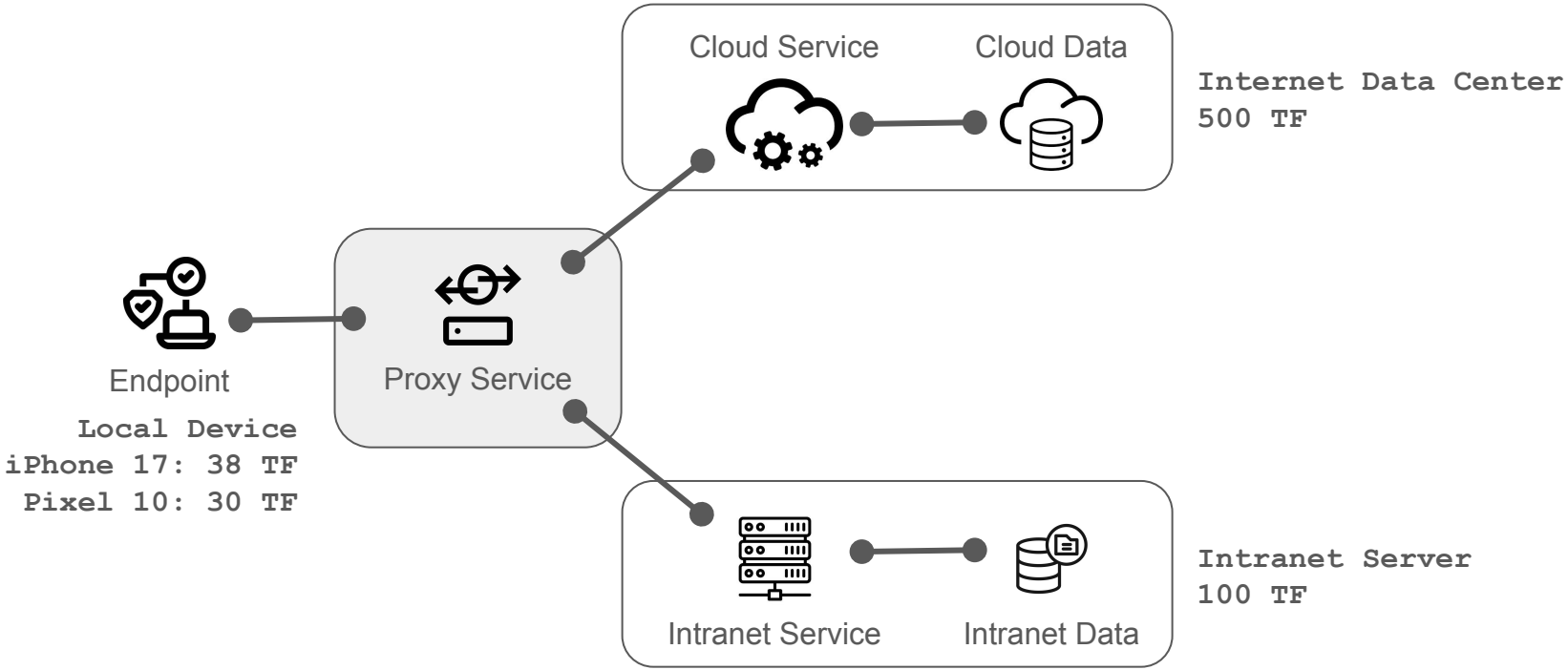
Enterprise endpoints (circa 2026)



Network round trip time (RTT)



Computational power (Teraflops)



Processing queries

Compute time only (ms) - 100 tokens query

Device	Mistral Small 3.1 (12GB)	Gemma 3 (13.5GB)	DeepSeek-R1-32B (16GB)
Internet Data Center (500 TF)	9.6	10.8	12.8
Intranet Server (100 TF)	48.0	54.0	64.0
iPhone 17 Pro (38 TF)	126.3	142.1	168.4
Google Pixel 10 (30 TF)	160.0	180.0	213.3

Total request time (ms) - 100 tokens query

Device	Mistral Small 3.1 (12GB)	Gemma 3 (13.5GB)	DeepSeek-R1-32B (16GB)
Internet Data Center (500 TF, 100ms)	109.6	110.8	112.8
Intranet Server (100 TF, 20 ms)	68.0	74.0	84.0
iPhone 17 Pro (38 TF, 0 ms)	126.3	142.1	168.4
Google Pixel 10 (30 TF, 0 ms)	160.0	180.0	213.3

Query context changes (100, 1,000, and 10,000 tokens)

Total request time (ms) - 100 tokens query

Device	Mistral Small 3.1 (12GB)	Gemma 3 (13.5GB)	DeepSeek-R1-32B (16GB)
Internet Data Center (500 TF, 100ms)	109.6	110.8	112.8
Intranet Server (100 TF, 20 ms)	68.0	74.0	84.0
iPhone 17 Pro (38 TF, 0 ms)	126.3	142.1	168.4
Google Pixel 10 (30 TF, 0ms)	160.0	180.0	213.3

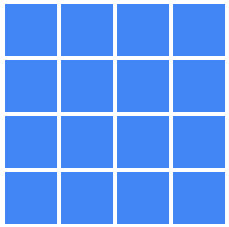
Total request time (ms) - 1,000 : 10,000 tokens queries

Device	Mistral Small 3.1 (12GB)	Gemma 3 (13.5GB)	DeepSeek-R1-32B (16GB)
Internet Data Center (500 TF, 100ms)	115.4 : 158.0	117.3 : 165.2	120.5 : 177.2
Intranet Server (100 TF, 20 ms)	96.0 : 316.0	105.0 : 354.0	121.0 : 416.0
iPhone 17 Pro (38 TF, 0 ms)	217.1 : 928.4	244.2 : 1,052.1	289.5 : 1,246.8
Google Pixel 10 (30 TF, 0 ms)	275.2 : 1,176.0	309.6 : 1,323.0	366.9 : 1,568.0

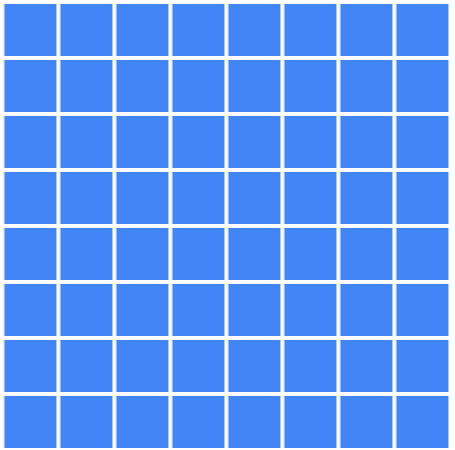
Quadratic growth of LLM attention matrix

2 Tokens (4 ops)

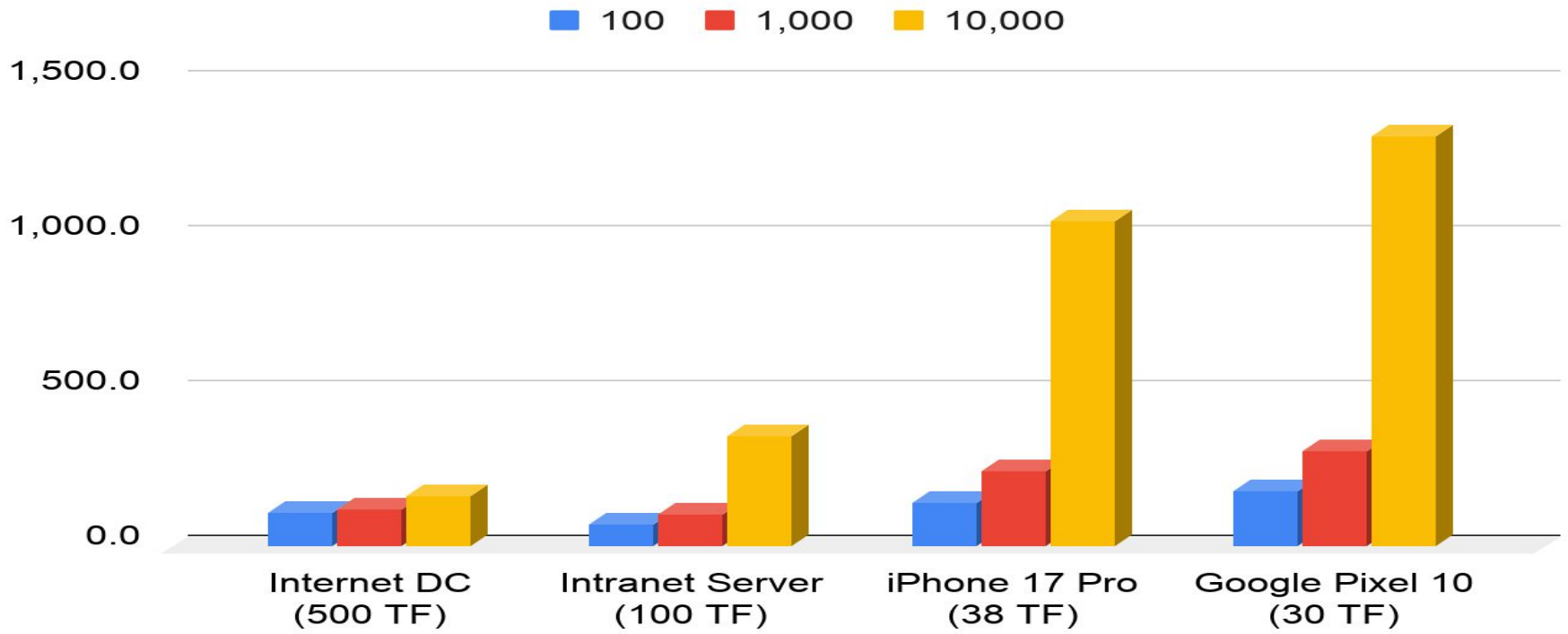
4 Tokens (16 ops)



8 Tokens (64 ops)



Gemma 3 processing time and context size



A case of use cases

When is my meeting with John Doe this week?

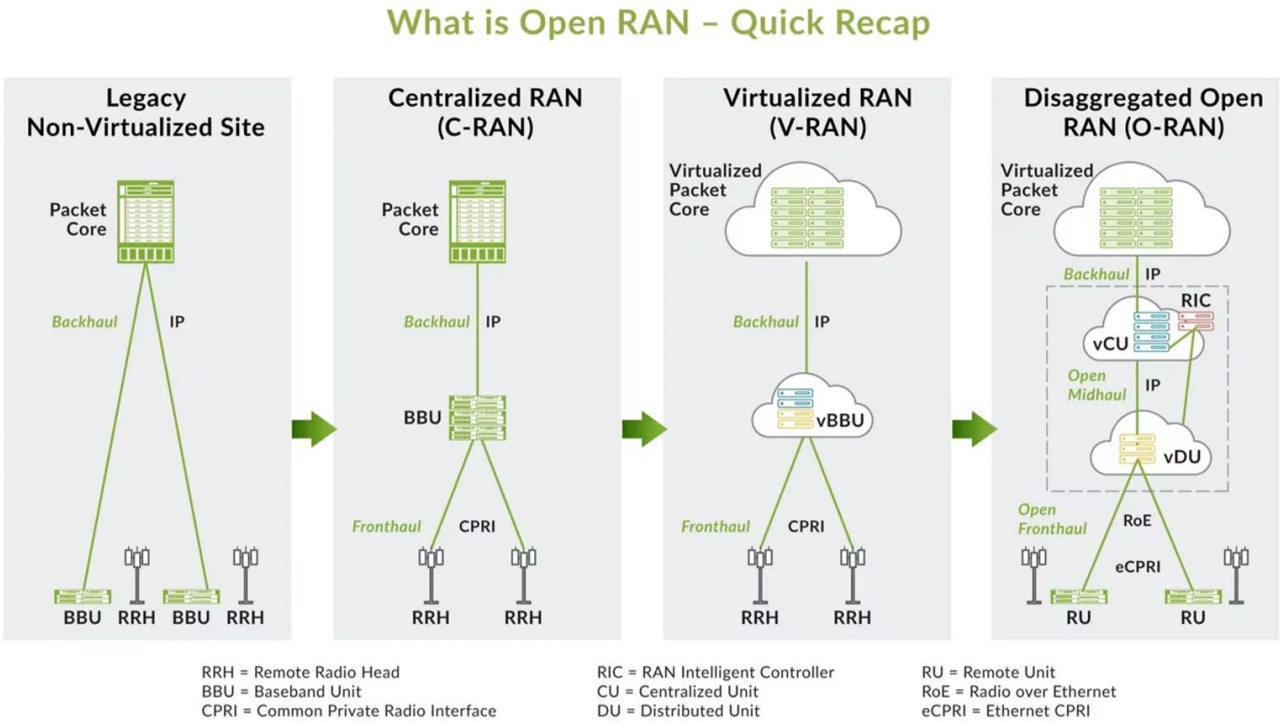
What is the average value of insurance contracts in ZIP code XYZ?

Resource optimization

- Reduce NPU clock speed
- Optimize battery usage
- Drop activity to reduce temperature

- Tenant A in rack with dataset A
- Scale if latency above 50ms for 10s
- Pause usage if cost above Z$

The RAN “compute power”

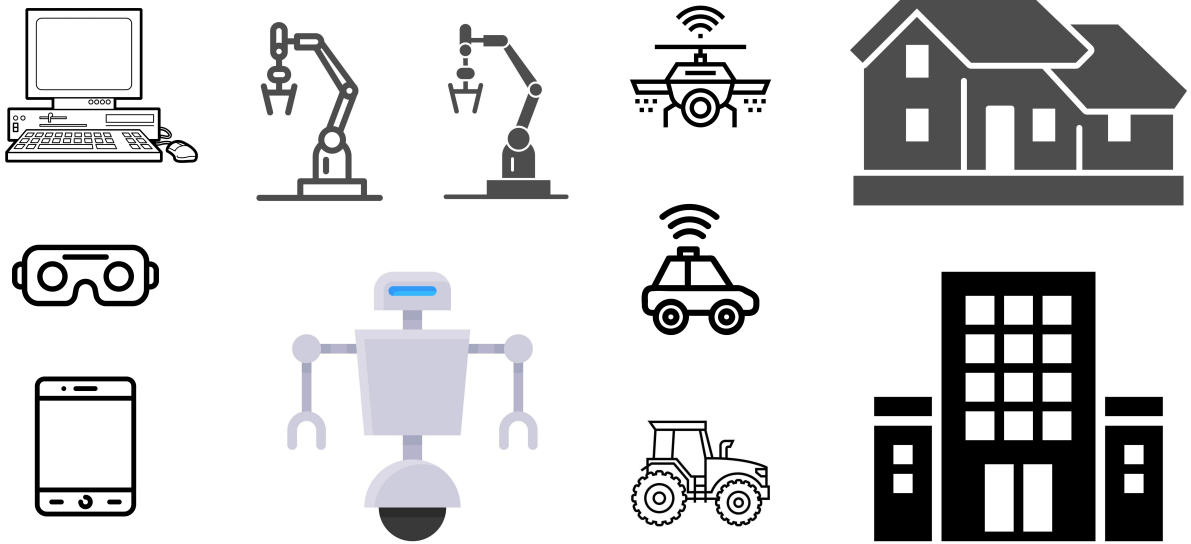


https://www.hpe.com/us/en/what-is/open-ran.html

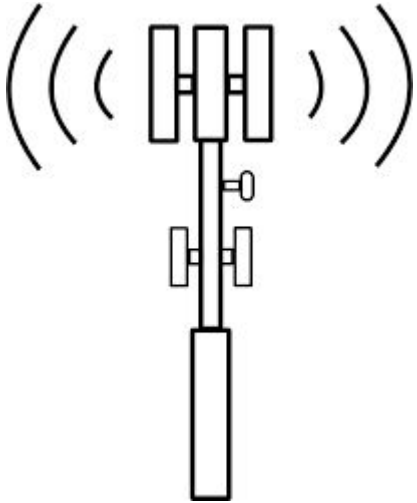
v2026-03-06c

Working backwards from the 6G-connected world

1Gbps everywhere, 0.1 milliseconds latency



Terahertz bands
100 GHz to 10 THz



Comparison: 5G vs. 6G targets

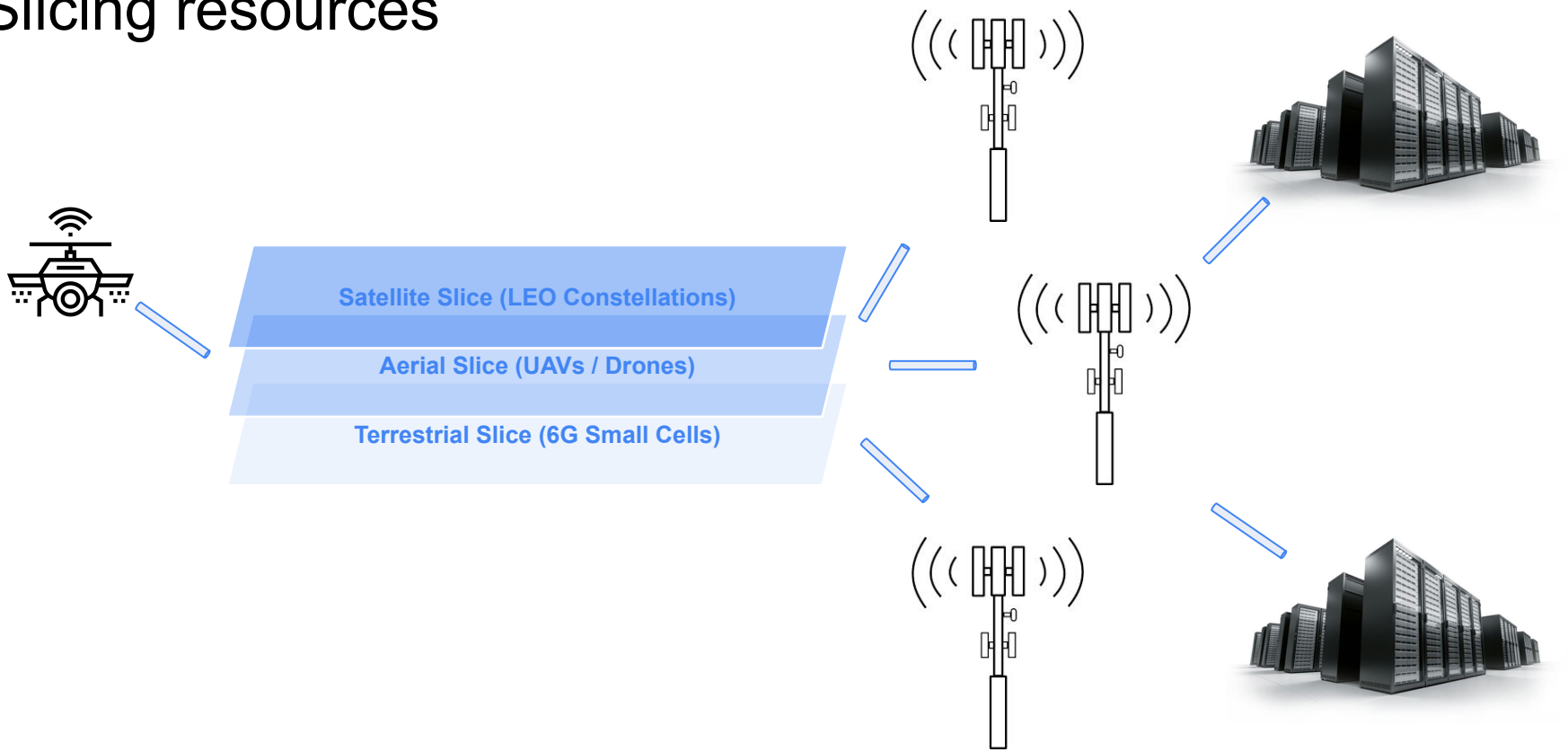
Metric	5G (IMT-2020)	6G (IMT-2030)	Improvement
Peak Data Rate	20 Gbps	1 Tbps	50x
User Experienced Rate	100 Mbps	1 Gbps	10x
Air Interface Latency	1 ms	0.1 ms (100 µs)	10x
Connection Density	10^6 devices/km²	10^7 devices/km²	10x
Mobility	500 km/h	1,000 km/h	2x

Manhattan density: 30K/km²
Peak: 300K/km² during workday

International Mobile Telecommunications (IMT) for 2030 and beyond (IMT-2030)
https://www.itu.int/rec/R-REC-M.2160/en

v2026-03-06c

Slicing resources



What is a Network Slice?

- A network slice is a logically isolated, end-to-end virtual network built on top of a shared physical infrastructure. Each slice has its own guaranteed compute, memory, bandwidth, and latency budget — invisible to other slices running on the same hardware.

eMBB Slice
Enhanced Mobile Broadband
Use: 4K video, AR/VR
Priority: high throughput
Latency: ~10ms OK

URLLC Slice
Ultra-Reliable Low Latency
Use: autonomous vehicles, surgery
Priority: deterministic delivery
Latency: <1ms required

mMTC Slice
Massive Machine-Type Comms
Use: IoT sensors, smart cities
Priority: massive device count
Latency: seconds OK

Architectural shift: Cloud-Centric → Edge-Native

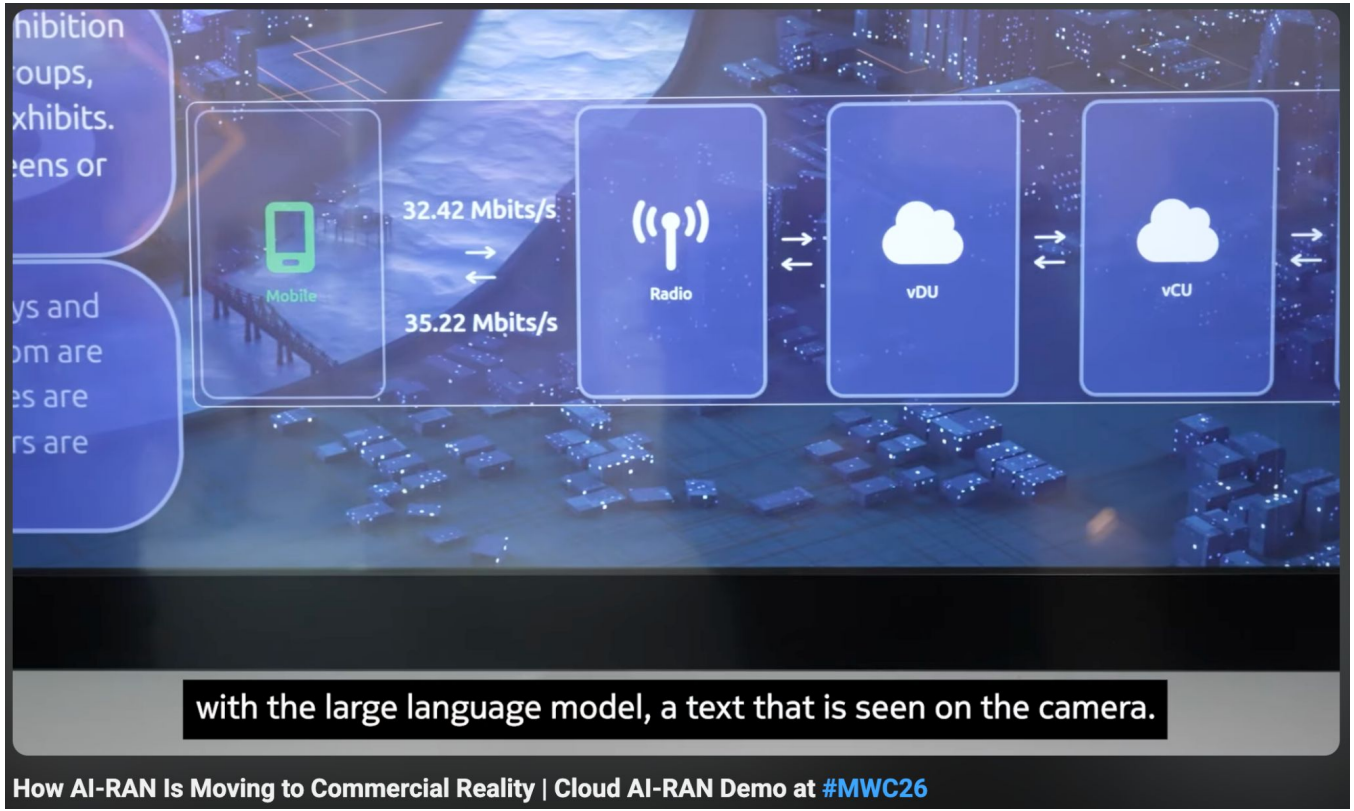
TODAY: Cloud-Centric Architecture

- Endpoint calls remote LLM in cloud data center over 100ms+ round-trip
- All context, private data, and inference logic travels over public internet
- Single point of failure: cloud outage = no AI

6G TARGET: Edge-Native Architecture

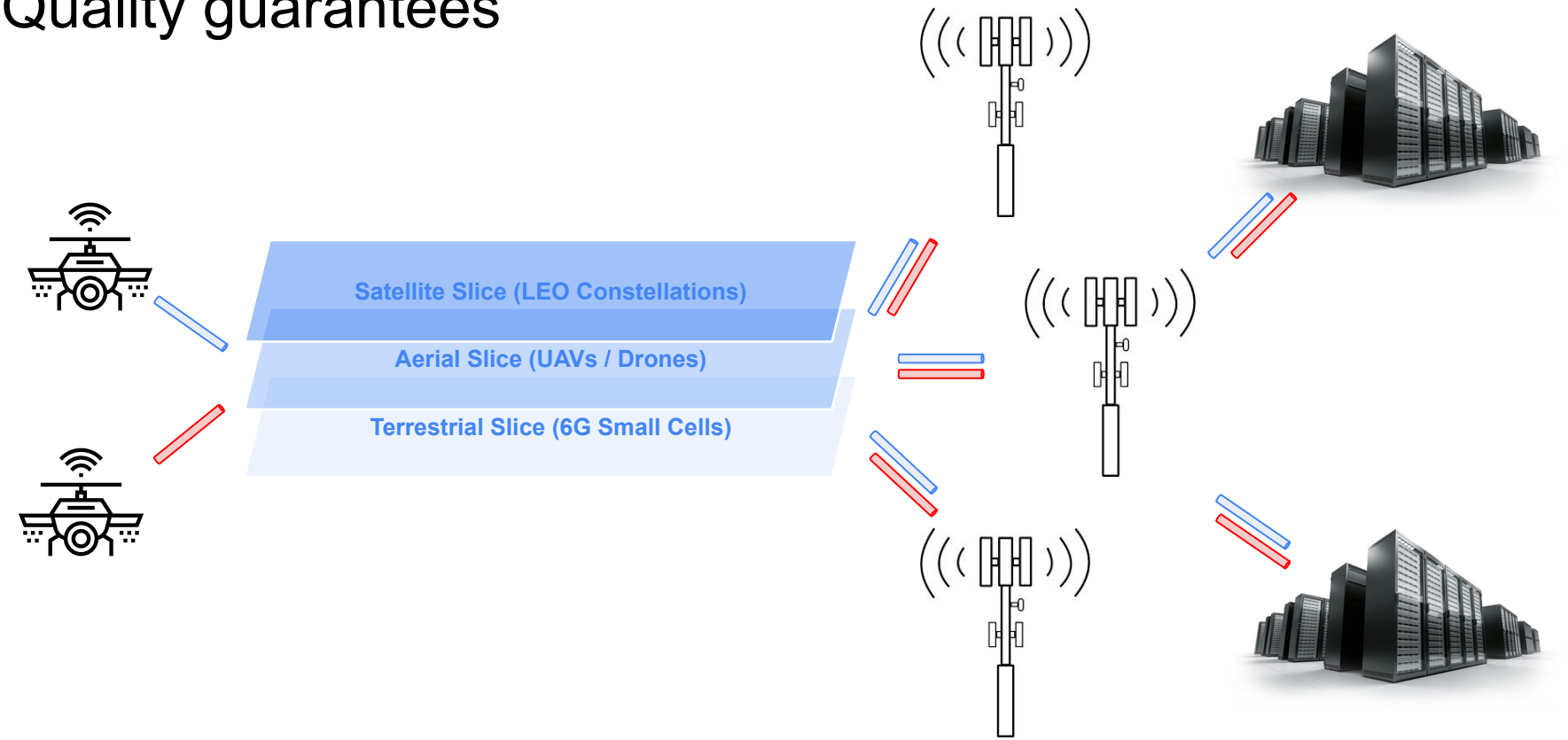
- Inference runs at RAN edge node: 0.1ms latency, no internet round-trip required
- Context and model weights cached at edge; private data never leaves the network slice
- Cloud used only for non-real-time tasks: model updates, aggregation, long-context batch jobs

Moves to the edge: Model inference, context caching, routing logic, SLA enforcement, privacy filtering.
Stays in the cloud: Model training, global aggregation, long-context tasks exceeding edge memory.

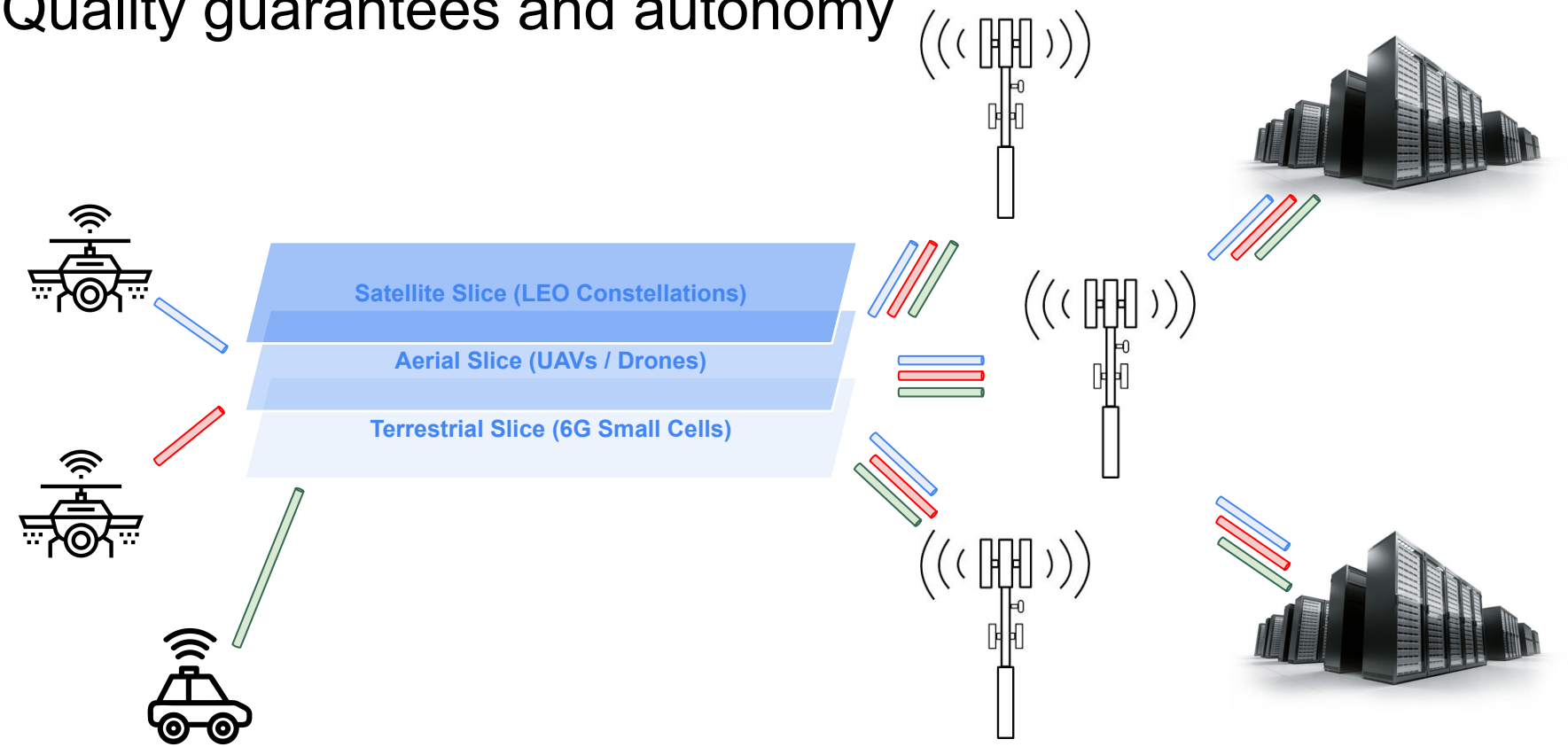


https://www.youtube.com/watch?v=7QNdKqGmPbw
https://ai-ran.org/

Quality guarantees



Quality guarantees and autonomy



Multi-tenant “AI Grid” requirements

- **Deterministic performance**: prioritize workloads with zero-latency jitter. Some cores pinned to network tasks. Others “rented out”.
- **Dynamic “slack” utilization**: reclaim resources “instantly” (including eviction of low-priority tasks).
- **Full-stack isolation (slicing)**: SLA for compute, storage, and networking.

From ai-ran.org: Unified AI-RAN Fabric

Federated scenarios

- **Car**: need a new route to [destination] due to accident
  - Or new route to avoid passing by location, person, …
- **Drone**: birds ahead, need to find alternative path
  - No dependency on “traffic control” past initial instructions
- **Monitoring camera**: detect person
  - Outside the “recognized house set” ⇒ Obfuscate identity
- **Federated learning or fine-tuning**: commute optimization
  - Without disclosing individual routes, tune models to optimize load on main roads

Federation and privacy

- Redact photo background
- Biometric data in enclave

- Data encrypted at rest
- No PII in training sets

Federated learning at the RAN edge

Train locally, share only gradients, aggregate globally — raw data never leaves the device.

- **Step 1 — Local Training:** Each edge device (phone, vehicle, sensor) trains the model on its own private data. The global model weights are downloaded first; only gradient updates are produced.
- **Step 2 — Encrypted Gradient Sharing:** Gradient updates are encrypted (differential privacy noise added) and transmitted to the RAN edge aggregator. Raw training data stays on-device.
- **Step 3 — Aggregation at Edge Server:** The RAN edge node uses FedAvg (or a variant) to merge gradient updates from all participating devices into an improved global model.
- **Step 4 — Model Distribution:** The updated global model is pushed back to all edge devices, completing the cycle. With 6G bandwidth, full model refresh can happen in <100ms.

Key challenges: Non-IID (Independent and Identically Distributed) data across devices, communication overhead on constrained links, Byzantine fault tolerance (malicious gradient poisoning), and balancing global vs. personalized models.

Ref: McMahan et al. (2017) FedAvg
https://arxiv.org/abs/1602.05629

Roadmap: Edge AI maturity timeline (Pragmatic opinion)

Phase 1 (2026–2030): Foundation — On-Device Intelligence (mostly still over 5G)

- Quantized, compressed models run locally on NPUs (today's 30–38 TF devices)
- Privacy sandboxing and TEE enclaves adopted broadly (EU AI Act compliance)

Phase 2 (2030–2036): Transition — RAN-Integrated Inference (early 6G)

- Multi-tenant AI Grids deployed in urban 6G small cells; SLA-aware scheduling standard
- Federated learning pilots at telecom scale; FedAvg replaced by communication-efficient variants

Phase 3 (2036–2042): Maturity — AI-Native RAN

- Intelligence fully embedded in network fabric; autonomous edge orchestration, no cloud dependency for real-time tasks
- Global federated models spanning satellite, aerial, and terrestrial slices (IMT-2030 vision)
- Privacy by design: homomorphic encryption feasible at RAN speeds; developers write to privacy APIs, not raw data

Call to Action: Preparing for the Edge AI future + Q&A

- Architect for distributed inference
  - Shift from cloud-dependent calls to tiered local processing to leverage 6G's 0.1ms latency.
  - Prioritize running real-time tasks at the network edge to eliminate Internet RTT bottlenecks.
- Master resource-aware development
  - Design for multi-tenant "AI Grids" where compute and memory are dynamically sliced.
  - Implement deterministic logic to maintain high performance while sharing hardware resources.
- Adopt privacy-first federated workflows
  - Replace centralized data harvesting with federated learning and secure on-device enclaves.
  - Process sensitive data locally and only transmit encrypted, anonymized insights to the network. See EU AI Act (Reg. 2024/1689).