IDC DIRECTIONS

Agents at the Edge

The Role of Devices in an Agentic Future

Tom Mainelli and Bryan Ma

IDC

The first wave of hardware-based device capabilities didn't shift much inference from the cloud

First wave shortcomings

- OS providers overpromised and underdelivered on features
- ISVs failed to embrace NPU for a multitude of reasons
- Cost benefits versus cloud have failed to materialize to date
- Cloud AI remained vastly superior

✨ **Silver Lining:** Despite shortcomings, progress in installed base, model quantization, and OS hardware abstraction will help set the stage for what comes next

What's different now: Agents represent another bite at the apple for the device ecosystem, especially PCs

Agents change the equation

- AI moves from responses to actions
- Workflows become autonomous and multi-step
- User intent replaces prompts as the starting point
- On-device model constraints become a feature rather than a bug

Question: Why won't agents just live in the cloud?
Answer: Many will, but devices will also orchestrate agents locally, on-prem, and in the cloud

Anybody heard of ~~Claudbot~~ ~~MoltBot~~ OpenClaw?

Open Source On-Device Agent

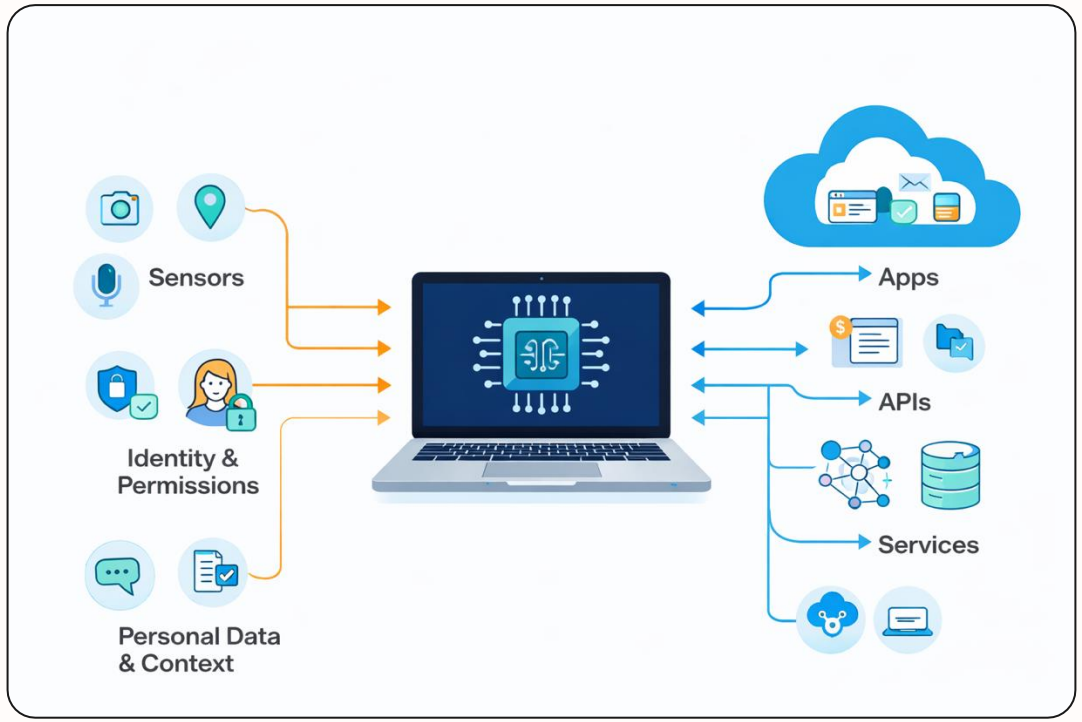
- Shift from chat to autonomous action
- Niche for now, but look at where China is
- Demonstrates risks around security, control, and governance
- Preview of distributed agent architectures

> OpenClaw is “the most important software release probably ever.”
>
> --Nvidia CEO Jensen Huang

Devices matter in Agentic AI because the cloud alone can't deliver autonomous, context-aware experiences

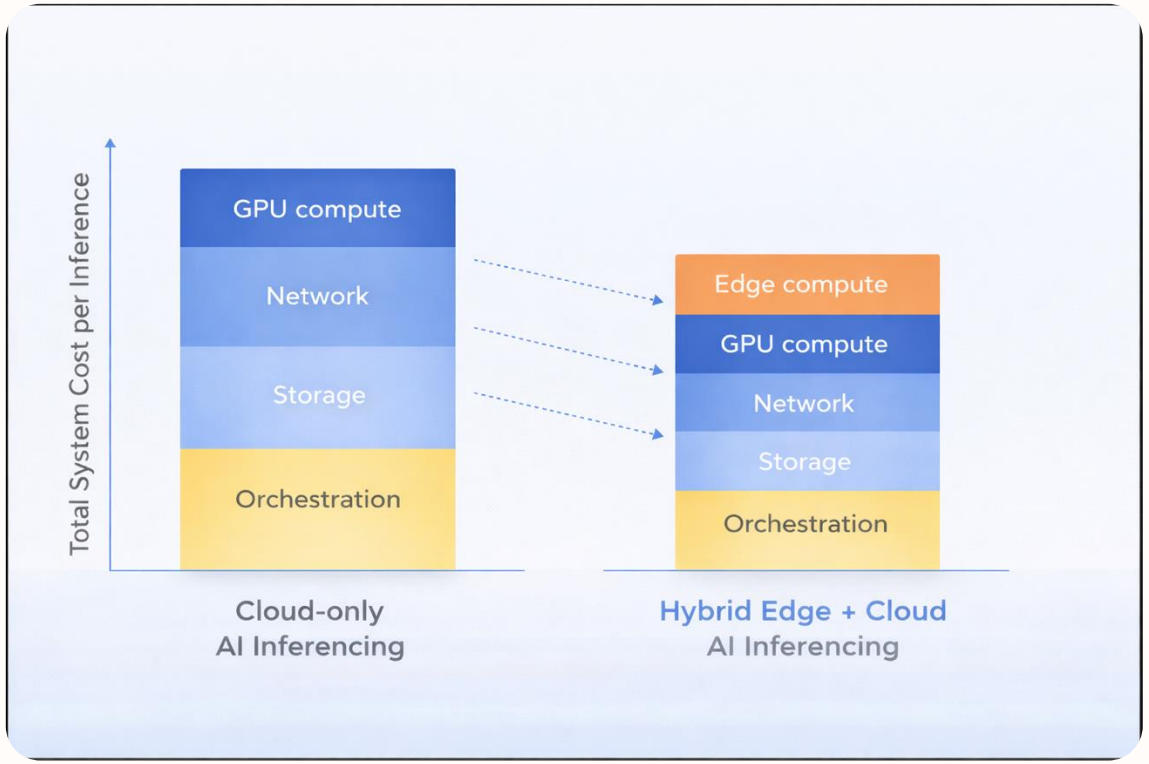
Structural device advantages

- Sensors and real-world inputs originate on-device
- Identity and permissions are often best governed locally. Plus personalization often requires local memory and context
- Note that this brings emphasis back toward GPU and CPU, not just NPU



Question: Aren't cloud models still better?
Answer: For some tasks, yes. Over time we expect devices to handle continuous perception and orchestration; cloud handles escalation and heavy reasoning

Edge inference reshapes AI economics and enables agent-scale deployment



Hybrid economics

- Structural reduction in marginal inference cost
- Lower network and centralized compute burden
- Dynamic cloud to edge workload distribution
- Enables cost-efficient scaling of millions of AI agents

Question: If cloud AI keeps getting cheaper, won't the edge advantage disappear?
Answer: Cheaper cloud AI lowers compute costs, but will it drop fast enough? Plus, latency, bandwidth, and data gravity still favor hybrid edge-cloud architectures at scale

Trust Will Be the Critical Gatekeeper for On-Device Agentic AI and the Battle for Control Has Begun

Who delivers transparent autonomy?

- Device vendors define the physical trust boundary
- OS owners govern permissions and orchestration
- Silicon anchors secure, verifiable autonomy
- Network providers control the real-time fabric agents depend upon

Key questions: Where does the critical data live, and who do device owners trust?

Agents at the edge herald the age of physical AI

Edge sensors and agents on devices → real-world action

- Wearables and smart glasses: AI augmenting human capability and interpreting the physical world
- Robots, drones, and autonomous vehicles: AI executing physical tasks based on sensors
- Survey: consumers like vacuums but hesitant in healthcare settings, concerned about reliability and price

Key point: When AI agents move to devices, embodied intelligence moves into the physical world

What to watch: The enablers that will determine how fast agentic on the edge begins to scale

Critical building blocks

- Multimodal models built for action, not just conversation
- OS-level agent frameworks with native app integration
- Trust, safety, and governance standards appear that users and regulators accept
- ISVs fully embrace and build for local agent capabilities across XPUs

Key insight: These enablers don't evolve independently, progress in one accelerates the others, creating compounding momentum.

Essential guidance

- Agents will redefine the device, so architect **devices and apps** for this future
- Hybrid AI will win in the end, so build for **and accelerate the pace** of cloud/edge cooperation
- Trust will decide winners, so ecosystem players must aborate to guard identity, permissions, and data
- The embodiment of intelligence via physical AI is next, so plan for a new class of intelligent devices