



Striding Towards the Intelligent World White Paper 2024

Navigating the Journey to All Intelligence



Building a Fully Connected, Intelligent World

Building a Fully Connected,
Intelligent World

▶ Contents

01

| Commercial AI is booming

02

| How enterprises can benefit from AI Intelligence

- Vision: 6A enterprises in the age of AI Intelligence
- Where to start: Focusing on high-value scenarios for greater value
- Infrastructure: Getting the infrastructure right to meet AI application needs in the long term
- Networks: Reshaping network experience and O&M with autonomous driving networks

03

| Implications of AI for networks

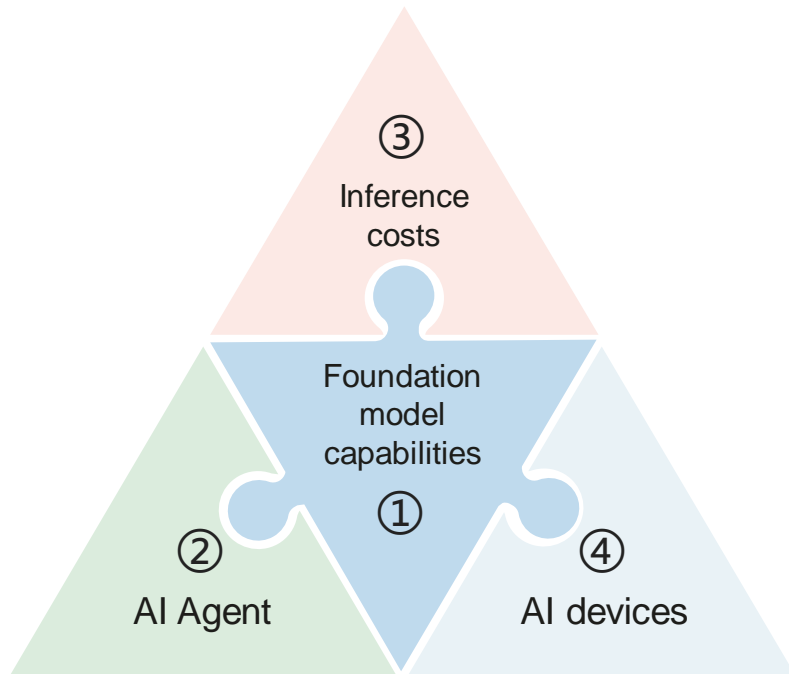
- Network for AI: Evolving networks to advance AI
- AI for Network: Embarking on Autonomous Network Level 4 for greater value

01

| **Commercial AI is booming**



► Four factors driving commercial AI adoption



① Basic capabilities of foundation models continue to evolve.

- **Multi-modal:** Foundation models are going multi-modal, supporting inference tasks based on voice, text, images, and more. Voice-based human-to-machine interaction is expected to reach human-to-human interaction levels (latency down to 300 ms, supporting proper interruption, and being able to recognize and demonstrate emotions).
- **Chain of Thought (CoT):** Technologies such as multi-step generation and policy search can be used to improve advanced inference capabilities of models.
- **Longer memory:** Models' context lengths continue to increase (up to 1 million tokens).
- **Reduced hallucinations:** AI governance rules and technologies, such as retrieval-augmented generation (RAG), can reduce hallucinations of models.

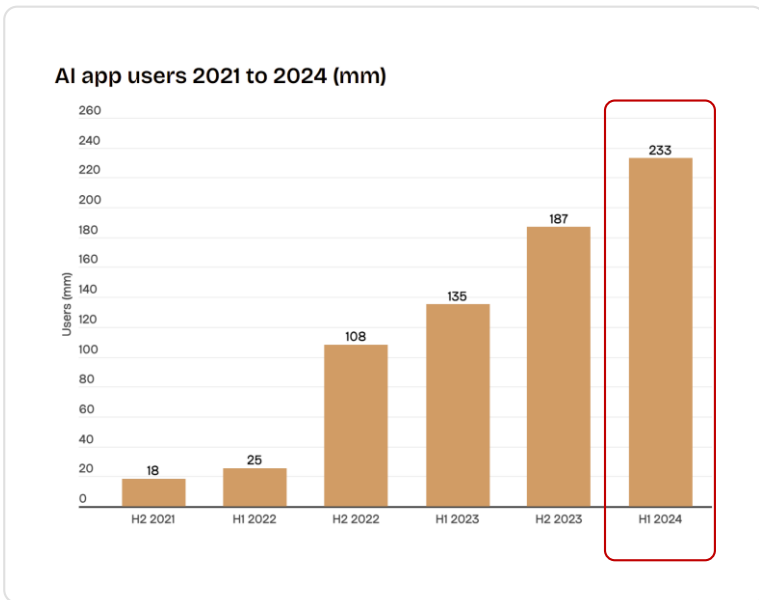
② AI agent frameworks and technologies are gradually maturing, with agents able to sense, plan, decide, and act independently in complex scenarios.

③ Inference costs are dropping steeply. The average inference costs decreased by more than 10 times during the last year.

④ Various AI devices are emerging. With AI devices like AI phones, AI PCs, AI glasses, and AI tutoring tablets hitting the market, more AI applications will be launched faster.

► B2C: Rapid growth of AI app users

- From June 2023 to June 2024, the number of AI app users increased from 135 million to **233 million** globally, and jumped from 8.2 million to **61.7 million** in China alone (a year-on-year increase of **653.3%**).
- A variety of B2C AI apps are emerging, including **AI assistants** and apps for content generation, **content editing**, **companionship**, **search**, and **education**.

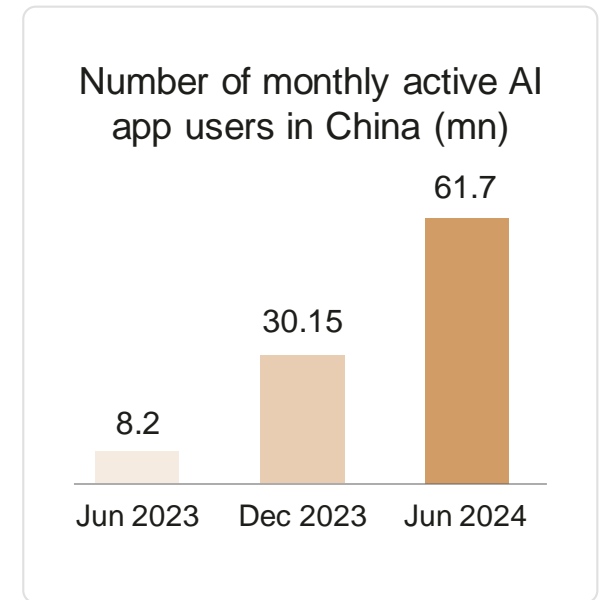


Sources: Business of Apps analysis and AppMagic

The Top 50 Gen AI Mobile Apps, by Monthly Active Users

1. ChatGPT	11. Facemoji	21. Chatbot AI & Smart Assistant	31. DAVINCI	41. Microsoft SwiftKey
2. Microsoft Edge	12. Remove It	22. Talkie	32. ChatBox	42. Prequel
3. photomath	13. ChatOn	23. Photo AI	33. Question AI	43. LooksMax AI
4. NOVA	14. EPIK	24. Face Dance	34. Cici	44. Umax
5. Bing	15. Translate	25. Luzia	35. Adobe Express	45. Bobbie AI
6. Remini	16. AI Mirror	26. Doubao	36. Copilot	46. ChatPod
7. Chat & Ask AI	17. Photoroom	27. Beat.ly	37. ImagineArt	47. Photoleap
8. BRAINLY	18. ChatBot	28. OANDA	38. PhotoApp	48. Chat AI
9. meitu	19. Hypic	29. SnapEdit	39. AI Chat	49. RIZZ
10. character.ai	20. AI Chatbot: AI Chat Smith 4	30. SNOW	40. Poly.AI	50. perplexity

Source: a16z



Source: QuestMobile

▶ B2C: AI devices expediting AI applications

Vendors are vying to launch AI devices

- In October 2023, Ray-Ban and Meta introduced their next-generation smart glasses.
- In May 2024, Microsoft released its AI PC Surface Pro.
- In June 2024, Apple released Apple Intelligence.
- In June 2024, Huawei unveiled Harmony Intelligence.
- In July 2024, iFLYTEK launched a next-generation AI tutoring tablet.

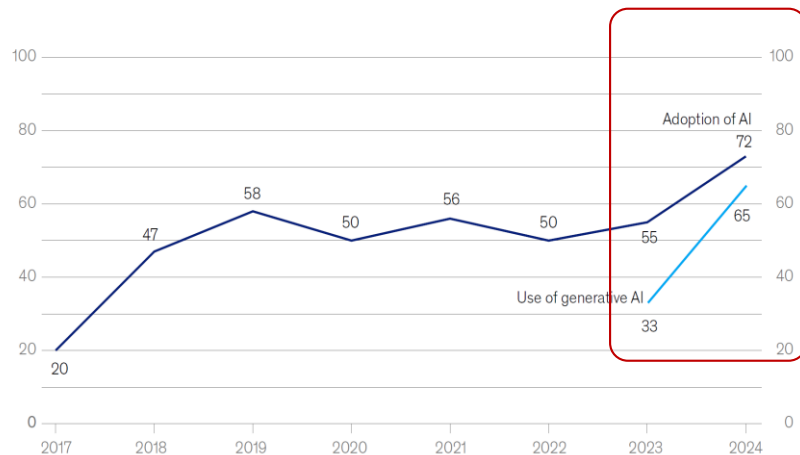
Advanced autonomous driving and robotaxis are becoming a reality

- In April 2024, Huawei unveiled ADS 3.0. The solution's Navigation Cruise Assist (NCA) is just a click away, enabling cars to automatically cruise – either on public or internal roads – to a destination parking lot either above or underground.
- Apollo Go robotaxis are now available in 11 cities in China. The city of Wuhan boasts the world's largest network of robotaxis, with a fleet of about 1000 vehicles by July 2024.

▶ B2B: AI adoption surging in organizations

Latest survey findings suggest that the proportion of companies adopting AI in at least one business function jumped from **55%** in 2023 to **72%** in 2024.

Organizations that have adopted AI in at least 1 business function, % of respondents

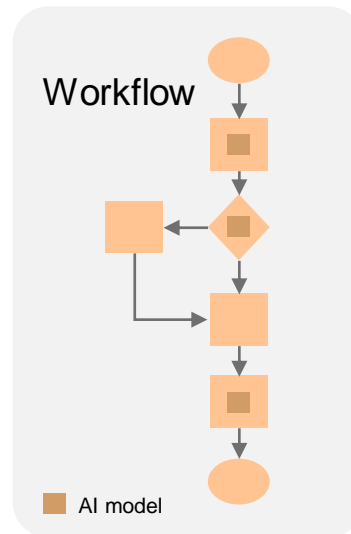


Source: McKinsey & Company

As foundation models become more capable and AI agents more mature, three trends will emerge:

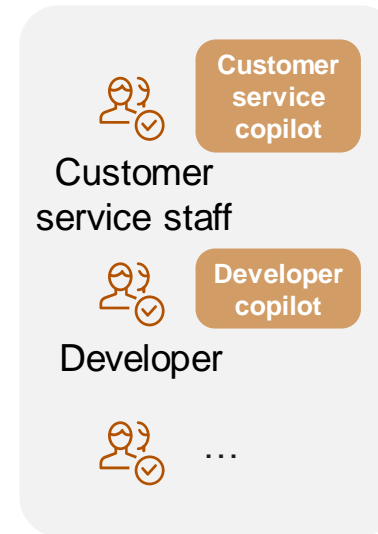
1. Foundation models will be increasingly used to replace traditional algorithms to perform individual tasks like CV and forecasts.
2. Role-based copilots will collaborate with employees in many functional domains, regardless of industry.
3. Businesses will deploy scenario-based AI agents, which is a new use case of AI.

① Embedded AI



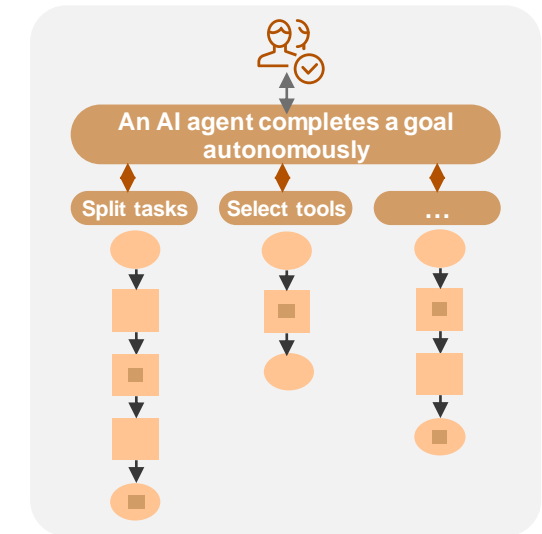
Traditional algorithms are replaced with foundation models to perform individual tasks.

② Copilots



Role-based copilot applications or tools powered by AI models are used.

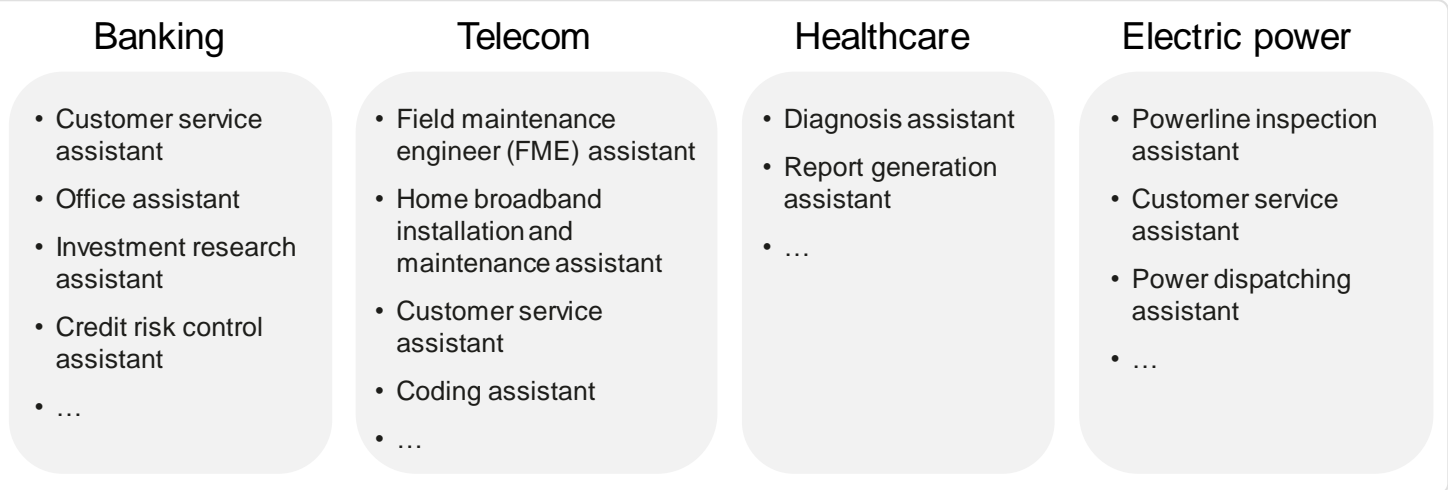
③ AI agents



Scenario-based AI agents complete human-defined goals autonomously.

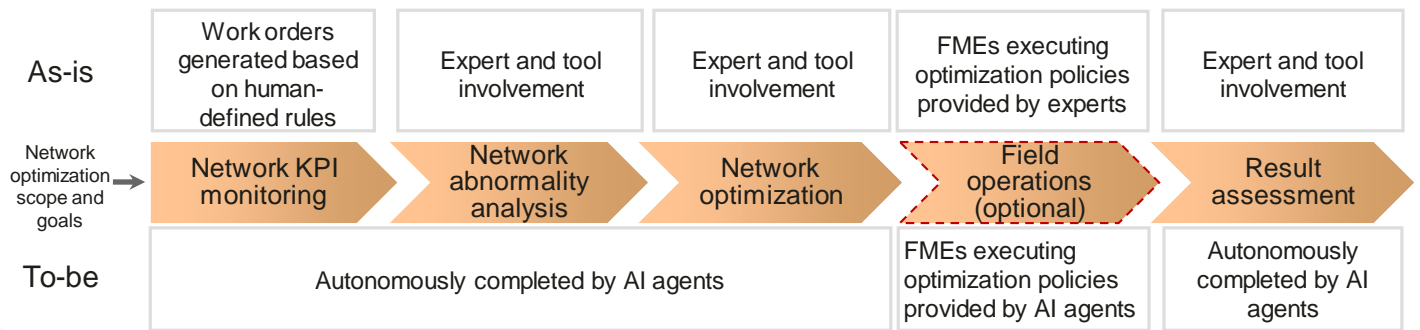
► B2B: Use cases of copilots and agents in industries

Copilot applications are role-based assistants that collaborate with employees to boost efficiency. They are ideal for scenarios with high fault tolerance, and are an important use case of foundation models. They are already being picked up by numerous industries including finance, telecom, healthcare, electric power, government, mining, and transportation.



As an ideal application of foundation models, an agent is able to sense, decide, and act independently. Its defining features include autonomy, adaptability, and interaction, which are key for businesses to achieve highly autonomous operations. Financial institutions, carriers, and manufactures, among others, are already piloting scenario-based agents, such as those for investment and network optimization.

AI agents for wireless network optimization: With an AI agent, one carrier has reduced the number of low-rate cells by 20% and cut the time spent on network optimization from one day to one hour.



02

How enterprises can benefit from All Intelligence

- Vision: 6A enterprises in the age of All Intelligence
- Where to start: Focusing on high-value scenarios for greater value
- Infrastructure: Getting the infrastructure right to meet AI application needs in the long term
- Networks: Reshaping network experience and O&M with autonomous driving networks



► Vision: 6A enterprises in the age of All Intelligence

Enterprises in the age of All Intelligence

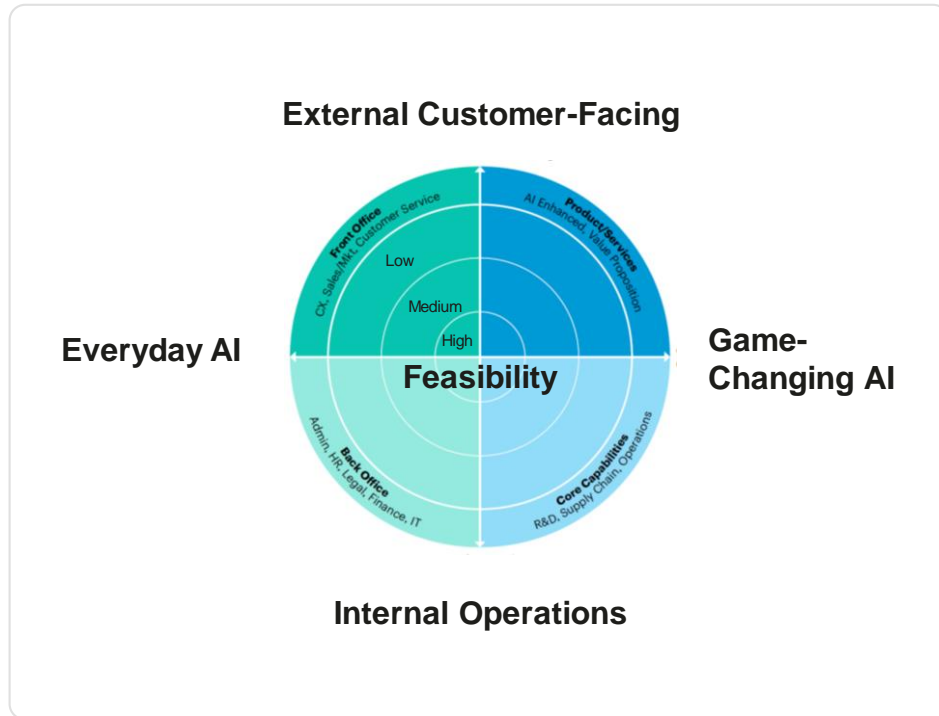
- Adaptive User Experience
- Auto-evolving Products
- Autonomous Operation
- Augmented Workforce
- All-connected Resources
- AI-native Infrastructure

- **Adaptive user experience:** AI systems perceive and understand user behaviors, requirements, interests, tastes, and environmental changes, and then adapt to provide services that best meet user needs. The ability of products to promptly meet massive personalized requirements should be designed from the start, instead of being realized through tailoring. For example, AI tutoring tablets will be able to automatically adjust teaching content and difficulty levels based on student age, learning progress, comprehension skills, and test results, giving each student their own unique learning experience at any time.
- **Auto-evolving products:** Intelligent products will possess the capabilities of self-learning, continuous iteration, adaptability to changes, self-optimization, and self-evolution. For example, self-driving vehicles will be able to learn by themselves – the more they drive, the better they drive.
- **Autonomous operation:** Closed-loop autonomous operations from sensing, planning, and decision-making to execution will become possible, making workflows highly autonomous. For example, intelligent planning platforms deployed at ports will automatically generate operation plans and autonomous container trucks will complete horizontal transportation.
- **Augmented workforce:** Each employee will be equipped with an intelligent assistant that understands them well and can help them complete tasks more efficiently and with high quality. For example, FMEs at base stations will be able to quickly obtain information such as the fault location, root causes, and handling suggestions through an assistant app.
- **All-connected resources:** Every part of an enterprise will be connected, from assets and employees to customers, partners, and ecosystems, and real-time feedback will be available. All objects, processes, and rules will be digitalized. This will drive up the quantity and quality of information, allowing companies to create their data flywheels and gain the upper hand with their information assets.
- **AI-native infrastructure:** AI-native infrastructure involves two aspects. First, the development of intelligent applications must be supported by comprehensive ICT infrastructure (i.e., ICT for Intelligence). Second, the O&M and user experience of ICT infrastructure must be fully intelligent (i.e., Intelligence for ICT).

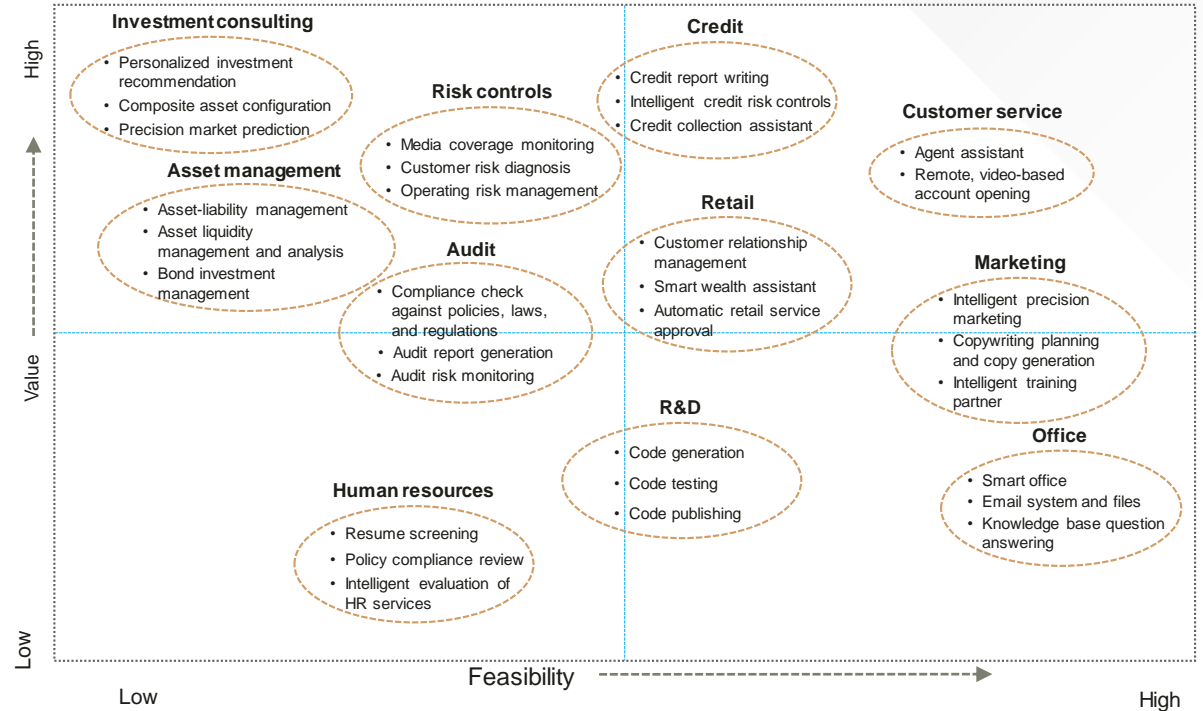
► Where to start: Focusing on high-value scenarios for greater value

To apply AI across the board, enterprises should first flexibly choose the right paths to adopt AI in select scenarios that create higher value, while considering their strategic goals and feasibility. For example, the banking sector will start by realizing Everyday AI in scenarios with high feasibility and high fault tolerance, such as customer service, marketing, and office scenarios, to achieve quick wins. They will then gradually move towards adopting Game-Changing AI in core scenarios with low fault tolerance, such as credit management, risk controls, investment consulting, and asset management.

Gartner AI Opportunity Radar



AI use cases in banking



► Huawei's experience in advancing AI

Huawei has been advancing AI on four fronts: External customer-facing AI, AI for internal operations, everyday AI, and game-changing AI.

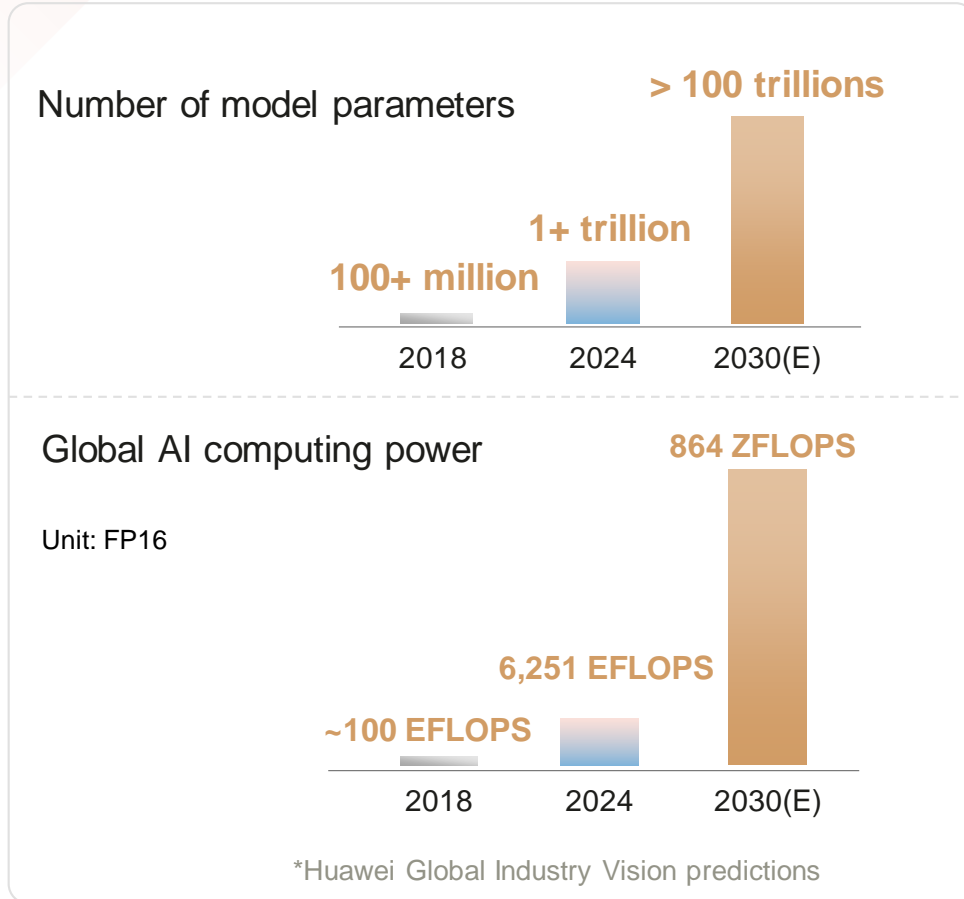
- Products + AI: Huawei's intelligent driving solutions, etc.
- Sales + AI: Ad placement, store site selection, sales training partner, etc.
- Engineering + AI: Supply-demand resource mapping, voice-based ticket filling, intelligent EHS check, automatic acceptance, etc.
- Customer service + AI: Reduced customer problem resolution time, improved customer service efficiency and experience, etc.
- Manufacturing + AI: Intelligent soldering, assembly, testing, and quality inspection for higher productivity and better experience
- Finance + AI: AI-assisted contract verification, contract risk clause identification, efficiency improvement, etc.

12 questions Huawei asks to identify suitable scenarios to apply AI

Business value		1. Can the value of the scenario be clearly measured ?
		2. How long does it take to achieve a positive return on investment (ROI) after implementing AI?
Scenario maturity	Business	3. Are there clearly-defined business owners for the scenario (who are responsible for investment and results)?
		4. Are there clearly-defined processes and rules for the scenario (which ensures that the business has clear boundaries)?
		5. Are there clear user touchpoints of the scenario (meaning the business has been digitalized)?
	Data	6. Is business knowledge/data sufficient to support 0-1 cold starts (clearly-defined, complete, and easily accessible)?
		7. Is business knowledge/data constantly generated, updated, and fed back with operations?
	Technology	8. Can existing technical capabilities support implementation in the scenario (technical feasibility)?
		9. Is there any successful experience that can be replicated within the company?
Continuous operations		10. Are there clear business goals ?
		11. Is there operational data supporting the business goals (measurable)?
		12. Are there organizations, resources, and mechanisms that support continuous operations?

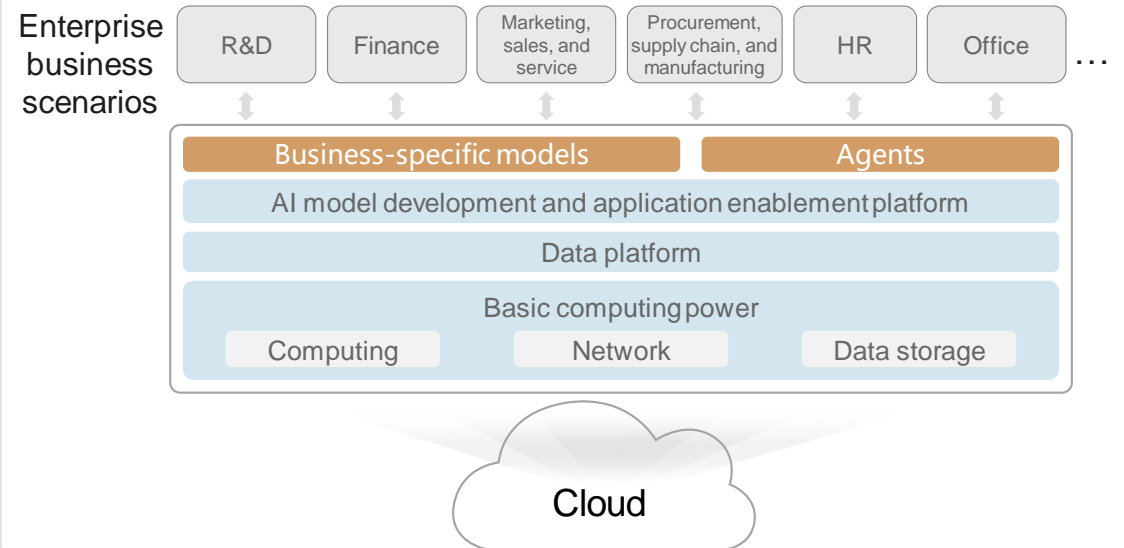
▶ Getting the infrastructure right to meet long-term AI application needs

The scaling law means that foundation model capabilities will continue to improve



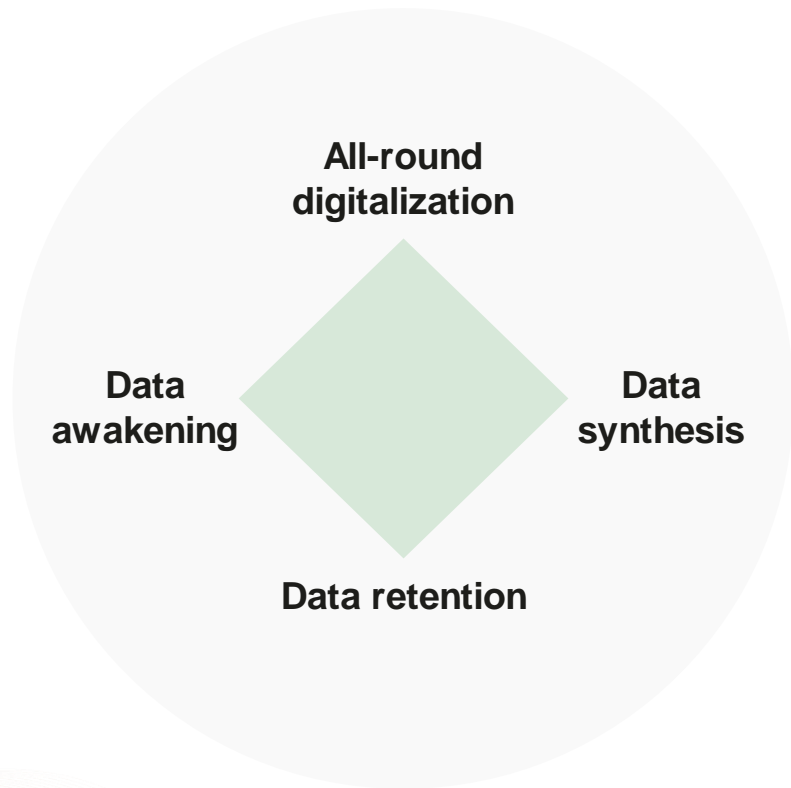
Key directions

1. Abundant, high-quality training data is the key to evolve models and improve enterprise intelligence levels.
2. AI-native computing infrastructure and data infrastructure based on cloud needs to be built to meet enterprises' growing demands for computing power and data.
3. A one-stop AI model development and application enablement platform needs to be developed to meet enterprises' model and application development requirements.



▶ Generating abundant, high-quality business data

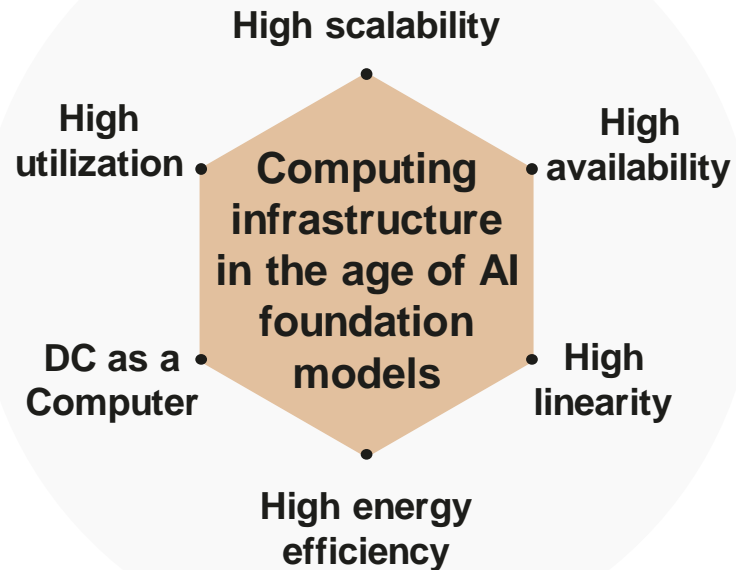
There are four areas that businesses can focus on to generate abundant, high-quality data for their AI systems.



- **Data awakening:** Cold data and historical data, such as business archives and records, exist in large quantities and are rich sources of information. Data awakening – the process of turning cold data into warm data, and ultimately, into actionable insights – is crucial for AI development.
- **All-round digitalization:** This serves to digitalize every aspect of a business, from its business objects and rules to its processes. Businesses must keep their AI application needs in mind when determining how often they collect data as well as the resolution and formats of the data collected. The more data that is collected, the better for AI development.
- **Data synthesis:** Businesses can generate synthetic data by using algorithms, statistical models, or generative AI. There is virtually no limit on how much synthetic data can be created, and such data is conducive to protecting privacy and reducing biases. According to Gartner, by 2026, 75% of businesses will use synthetic data, up from less than 5% in 2023.
- **Data retention:** In the past, data was retained for the purposes of future queries and legal compliance. Now, when businesses determine how long data should be retained, they also need to account for AI development. The longer data is retained – to the extent allowed by applicable law – the better for AI development.

▶ Building AI-native computing infrastructure

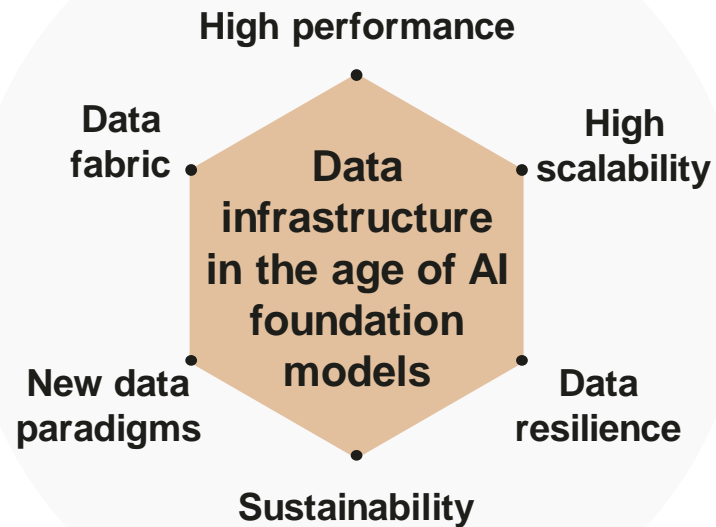
Six features will determine whether computing infrastructure can meet long-term business needs for AI adoption.



- **High scalability:** The computing infrastructure can continuously scale up to meet growing demand for AI computing.
- **High availability:** Designed to support hyperscale clusters, the computing infrastructure can ensure stable training over long periods of time.
- **High linearity:** The computing infrastructure supports compute workloads by design and minimizes resource overhead. This is key to the linear performance growth of hyperscale clusters.
- **High utilization:** All types of processors, memory, and gears are managed in a centralized pool to provide unified resource scheduling and processing, which is especially crucial for hyperscale clusters.
- **DC as a Computer:** A data center (DC) works as a computer, employing a unified network architecture, a unified bus, and memory semantic communication.
- **High energy efficiency:** Liquid cooling reduces the energy use of the computing infrastructure.

► Building AI-native data infrastructure

Six features will determine whether data infrastructure can meet long-term business needs for AI adoption.

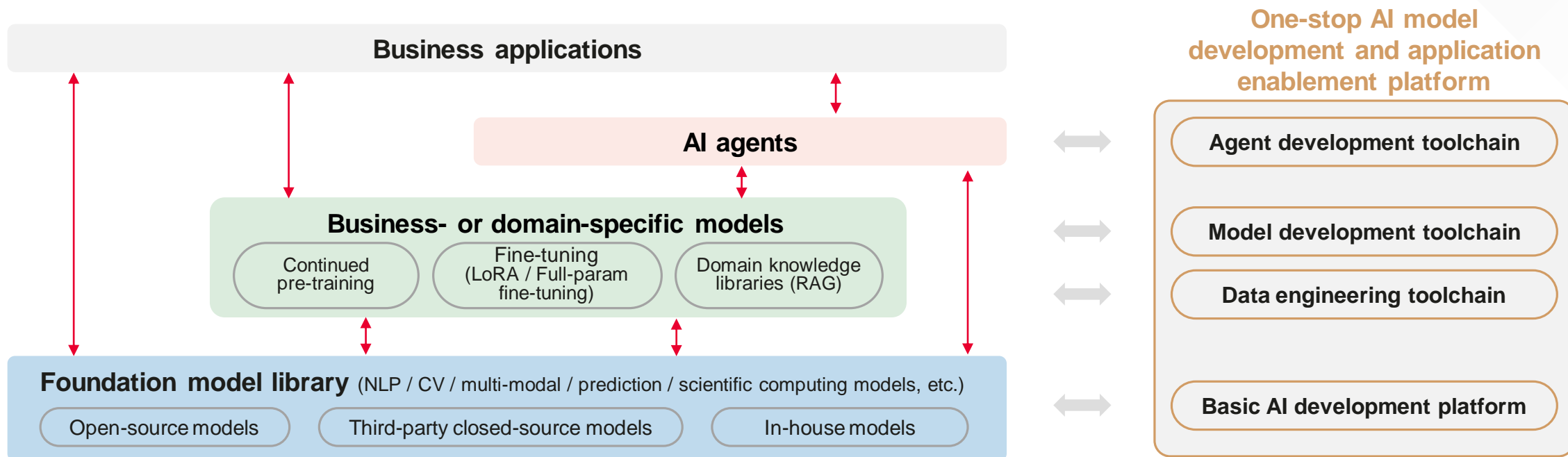


- **High performance:** All-flash storage is the ideal solution to fast data processing. The combination of all-flash storage and innovative architecture can increase processing efficiency by at least tenfold.
- **High scalability:** An innovative architecture of storage-compute decoupling allows for exabyte-scale capacity expansion. As capacity grows, performance increases linearly.
- **Data resilience:** Redundancy by design provides at least 99.9999% availability. Multi-layer security helps protect against at least 99.99% of ransomware attacks. The checkpoint (CKPT) mechanism ensures data recovery within seconds.
- **Data fabric:** With a global data map, it's possible to visualize and manage massive amounts of heterogeneous data from multiple sources, and use this data to create greater value.
- **New data paradigms:** Vectors, KV-Cache, and other new data paradigms are employed to optimize the semantics of data access, and enable life-long memory for inference as well as high-precision and high-performance retrievals.
- **Sustainability:** New storage media and hardware innovations promise high energy efficiency (less than 1 watt per TB) and storage density (greater than 1 PB/U).

► Creating a one-stop AI model development and application enablement platform

Given that foundation models are being rapidly iterated, enterprises build their foundation model libraries with either open-source models or closed-source models of third parties, or by developing their own models. Then, the enterprises create their business- or domain-specific models with their own data to meet their own needs. AI agents are an ideal application of AI models, and enterprises are scrambling to develop AI agents on top of their AI models. During this process, the enterprises need a mix of capabilities, which can be provided through a one-stop AI model development and application enablement platform.

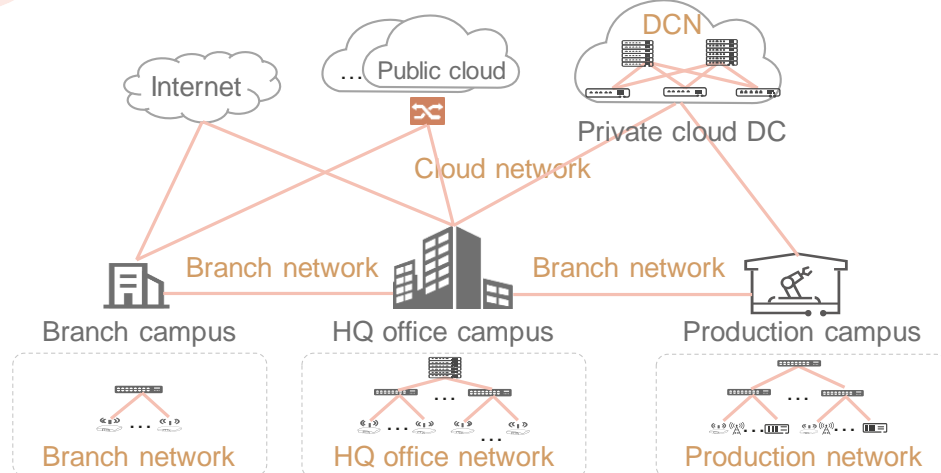
The one-stop AI model development and application enablement platform includes a basic AI development platform (developing and managing in-house models); a data engineering toolchain and a model development toolchain (using foundation models and high-quality data to rapidly build and deploy business- or domain-specific models through fine-tuning, continued pre-training, and other means); and an agent development toolchain (supporting rapid development of AI agent applications).



► Reshaping network experience and O&M with autonomous driving networks

Network challenges

- It's difficult to guarantee good employee experience in a fully wireless office where there is huge demand for cloud-based applications, video, collaboration, and smart services.
- O&M workloads grow in scope and complexity as networks grow increasingly larger (office & production networks, data center networks, branch networks, and cloud networks), with more and more types of network elements (NEs).
- Designing, integrating, verifying, and launching a new service takes a long time if it involves multiple network domains.
- The exponential growth of virus variants and the fact that attackers are increasingly using AI to further their agenda will make it more challenging to safeguard operations and manage unknown threats.



Value of scenario-based agents and role-based copilots, powered by the Telecom Foundation Model and digital twins:

Zero service delays

- Guaranteed experience for all users, all NEs, and all applications

Zero network disruptions

- Real-time visibility into how all NEs on the network are working; proactive risk identification and mitigation; automated recovery of non-hardware faults

Zero-wait service provisioning

- Real-time provisioning of new services, thanks to automated network configuration generation, simulation, and verification

Zero security risks

- Ever-evolving ability to detect virus variants and unknown threats, and automated handling of security threats

Agent for network optimization

Agent for network fault recovery

Copilot for customer service

Agent for configuration generation

Copilot for security

...

03

| Implications of AI for networks

- Network for AI: Evolving networks to advance AI
- AI for Network: Embarking on Autonomous Network Level 4 for greater value



▶ Evolving networks to advance AI

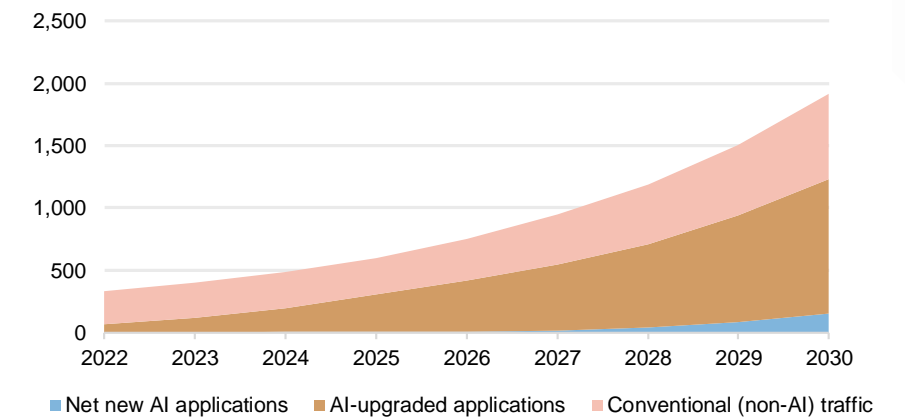
Networks for individual users, vehicles, machines, and businesses must evolve to meet the needs of AI applications.

- Networks for individual users: For AI assistants to provide an incredible or even human-like experience, they must rely on real-time, multi-modal interactions, which in turn require extremely low network latency and high uplink bandwidth. Providing a human-like experience requires an overall latency down to 400 ms, or an E2E network latency of about 70 ms if the latencies of the application and device are taken out of the equation. If an app or web user wishes to see an AI-generated 3D model of an item in real time, the network's downlink bandwidth must reach 80–400 Mbps, depending on the size of the model. Additionally, more high-quality AI-generated videos will become available, and AI recommendation systems will boost video content distribution and views. Network traffic will surge as users spend more time watching videos in higher definition.
- Networks for vehicles: Robotaxi security monitoring and remote takeover raise the bar for network latency and uplink bandwidth. In the case of Apollo Go, remote takeover means the videos of six cameras (12–24 Mbps) must be uploaded via a 5G network, and an E2E network latency greater than 100 ms can adversely affect control precision.
- Networks for machines: The number of IoT devices with multi-modal sensing capabilities will grow, increasing demand for network uplink bandwidth and creating new network traffic.
- Networks for businesses: Transmitting massive amounts of sample data for training requires elastic private lines with a minimal bandwidth of 10 Gbps.

According to Omdia, global network traffic will increase at a compound annual growth rate of 25%, thanks to AI-upgraded applications and net new AI applications.

AI Applications		E2E Network Latency	Uplink Bandwidth	Downlink Bandwidth	Traffic
Networks for individual users	AI assistants: Real-time multi-modal interactions <small>(Incredible or even human-like experiences)</small>	70–200 ms	≥ 20 Mbps	-	New traffic
	High-quality AI-generated content (HD / 4K / 8K videos; AI recommendation systems that promote content consumption)	-	-	Est. bandwidth for 8K videos > 100 Mbps	Traffic growth
	3D models of items (Real-time display of 3D models)	-	-	80–400 Mbps	Traffic growth
Networks for vehicles and machines	Robotaxis <small>(Security monitoring, remote takeover, and data uploading)</small>	< 100 ms	12–24 Mbps	-	New traffic
	IoT devices with multi-modal sensing	-	≥ 1 Mbps	-	New traffic
Networks for businesses	Uploading of sample data used for training	-	≥ 10 Gbps <small>(Elastic bandwidth)</small>	-	-

Projected global network traffic growth, 2023–2030
(Exabytes per month)



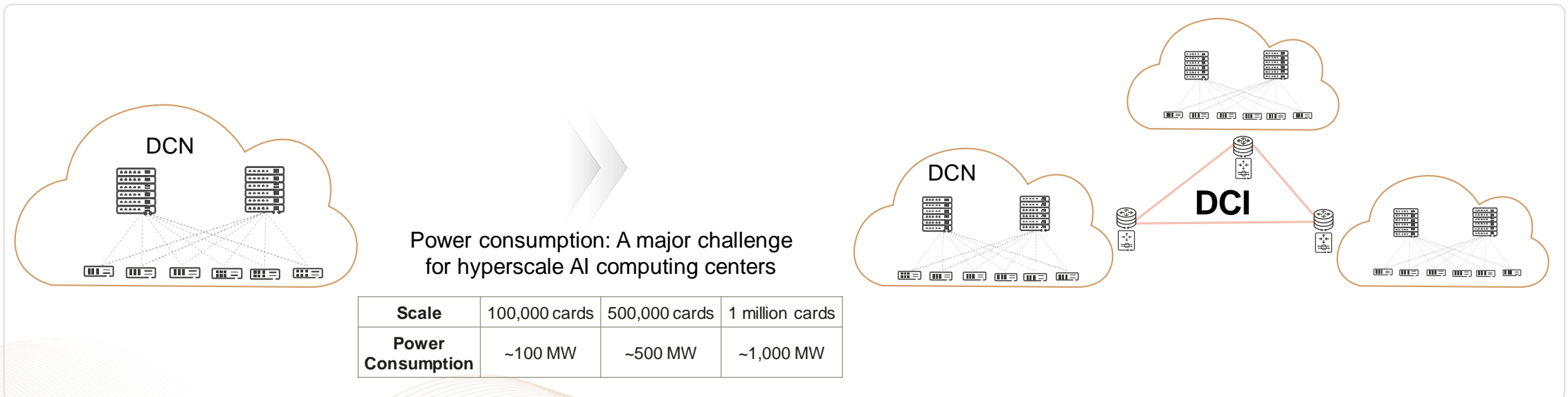
Source: Omdia

© 2023 Omdia

▶ Supporting AI training using 100,000+ cards across data centers

Nowadays, training an AI model requires 100,000 cards, if not more. A single AI computing center may not be able to support this requirement due to equipment room and power restrictions. Therefore, it makes sense to combine the AI computing power of multiple DCs to achieve model training across the DCs. However, this poses challenges to interconnection between the DCs.

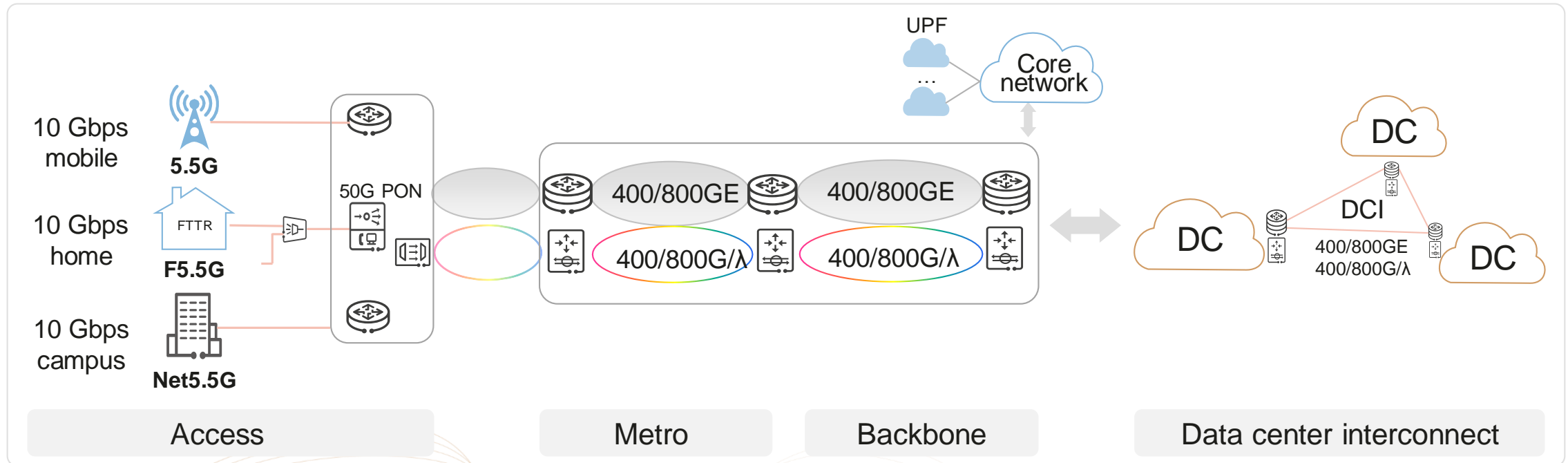
- **Zero packet loss of DCI:** AI training workloads are highly sensitive to network packet loss, and training efficiency may be halved even if only 0.1% of packets are lost.
- **Network utilization:** When elephant flows occur, the conventional 5-tuple-based load balancing method may fail. Unbalanced loads can affect the utilization of the entire network.
- **Traffic surges:** In a cluster with tens of thousands of cards, burst traffic and concurrent traffic are commonplace. In extreme cases, peak concurrent traffic may suddenly rise above 1,000 Tbps.



► Evolving towards 5.5G for optimal AI application experiences

2024 is the first year of 5.5G commercial deployment, and carriers that evolve their networks to 5.5G will be well-poised to support AI applications. Specifically, carriers can:

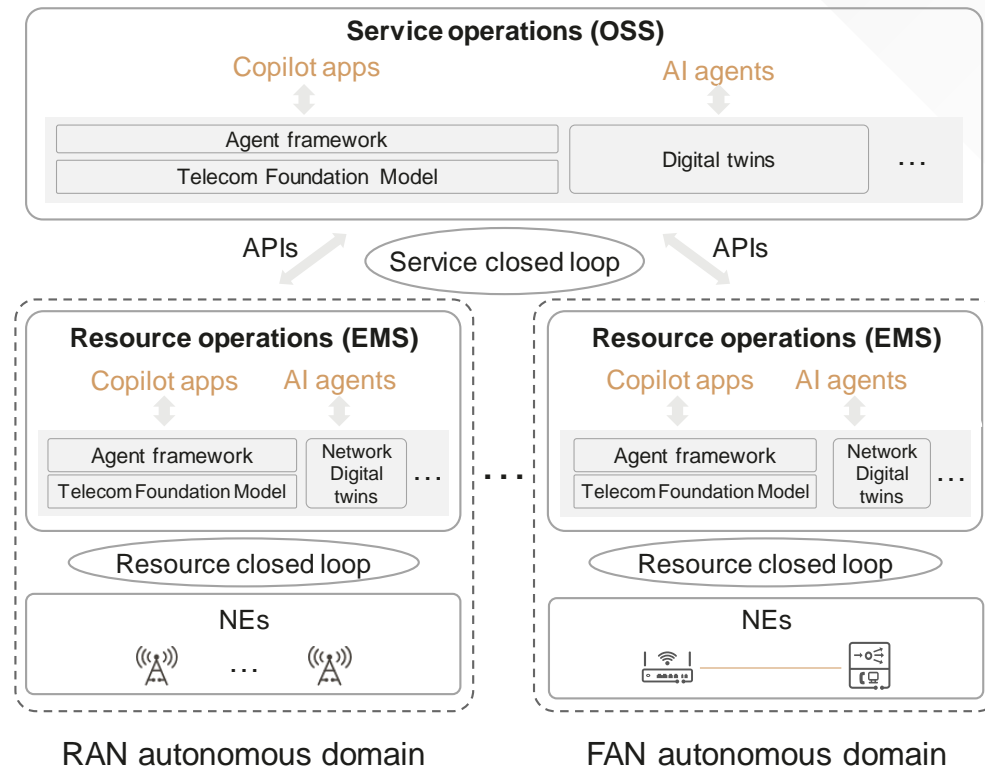
- Provide 10 Gbps access networks for mobile subscribers, homes, and campuses to enable ubiquitous access to AI services.
- Combine IP and optical technologies to deliver ultrafast, low-latency, intelligent, and elastic connections on metro and backbone WANs, thus empowering incredible AI application experiences.
- Combine IP and optical technologies to support ultrafast and reliable DCI connections and intelligent resource scheduling, thus underpinning cross-DC training that requires massive AI computing power.



▶ AI for Network: Embarking on Autonomous Network Level 4 for greater value

- In June 2024, the TM Forum published the Autonomous networks: Level 4 industry blueprint. This white paper sets out the guidelines for realizing Level-4 autonomous networks (AN L4) in high-value scenarios, specifies the first batch of 15 AN L4 high-value scenarios, and provides a reference architecture for AN L4.
- It makes sense for carriers to consider L4 as the starting point of their AN journey. The next step is to have the right strategy in place, select the right scenarios, and revamp business processes with AI agents and copilot apps. This is how carriers can minimize the impact of AN implementation on their organizations. Carriers can first target quick wins, and then iterate progressively to reach new heights.

AN L4 High-Value Scenarios		Efficacy Indicators	
1	Operations	Mobile service assurance	Subscriber complaint rate and service availability
2		Mobile service complaint handling	Complaint handling timeliness
3		HBB service assurance	Subscriber complaint rate and service availability
4		HBB complaint handling	Complaint handling timeliness and onsite repair time
5		Private line service provisioning	Success rate of automated E2E service provisioning
6		Private line service assurance	SLA fulfillment rate
7–11	Maintenance	Troubleshooting for five domains (wireless network, fixed access network, core network, IP network, and optical transport network)	Number of fault tickets MTTR Onsite self-service
12		Core network changes	Zero major faults
13	Optimization	Wireless network optimization	Optimization automation rate and network indicators
14		Wireless network energy consumption optimization	Year-on-year decrease in energy consumption per GB
15		IP network optimization	Optimization automation rate and network indicators



Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and
organization for a fully connected,
intelligent world.

**Copyright©2024 Huawei Technologies Co., Ltd.
All Rights Reserved.**

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

