中国移动
China Mobile

Model Distribution Network Considerations in China Mobile

March 2025

www.10086.cn

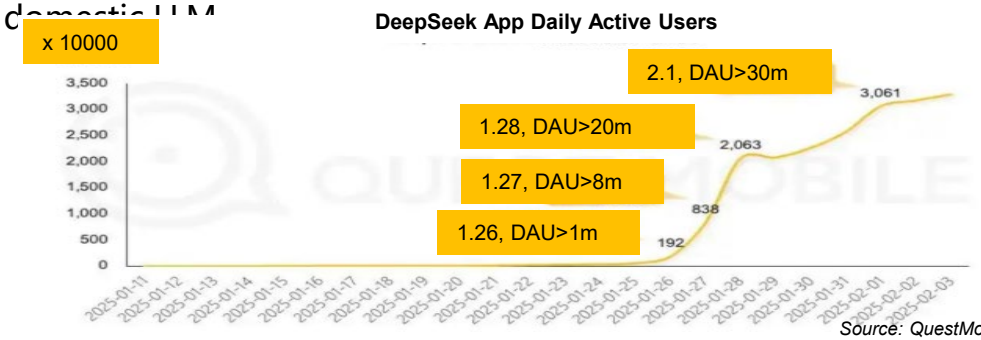
DeepSeek Opens the Era of Inclusive AI with Explosive Growth in Inference Services

DeepSeek-R1 achievs accuracy comparable to OpenAI's o1 model with 70% lower computation requirements, becoming a global focus.

Market Success : Surpassed ChatGPT in downloads across 140 app markets; fastest application to exceed 30 million daily active

Trend1： Paradigm shift from content to AI model delivery

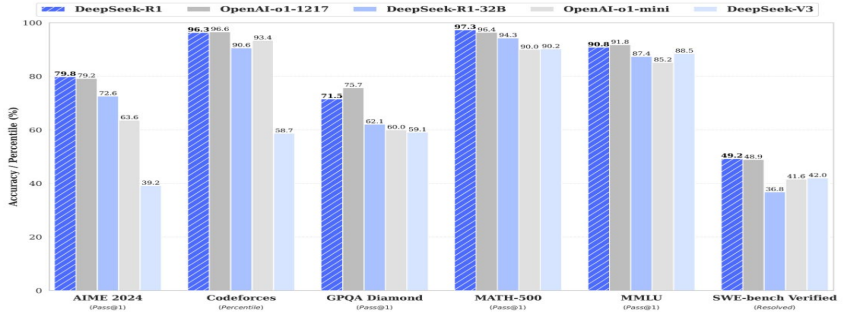
- **DeepSeek Integration:** Over 160 enterprises globally have announced integration with DeepSeek, covering cloud service providers, telecom operators, cybersecurity, automotive, smart hardware, finance, and semiconductor manufacturing.
- **DAU Record:** On Jan. 26, The number of daily active users exceeds 1 million. Two days later the number gets 20 million. Then after another four days it is 30 million, becoming the most popular domestic LLM



Source: QuestMobile

Trend 2: Focus on the "Inference Moment"

- **Performance:** Leading industry performance in mathematics(AIME, MATH-500), knowledge, reasoning(GPQA), and programming tasks(Codefources)
- **Inference Speed:** Increased by 4 times, with API call costs at nearly 3% of GPT-4-Turbo()
- **Model Distillation:** According to Ollama assessment data, distilled smaller models (e.g., 32B) perform very well in benchmark tests
- **Industry Requirement:** Vertical industries have a significant requirem



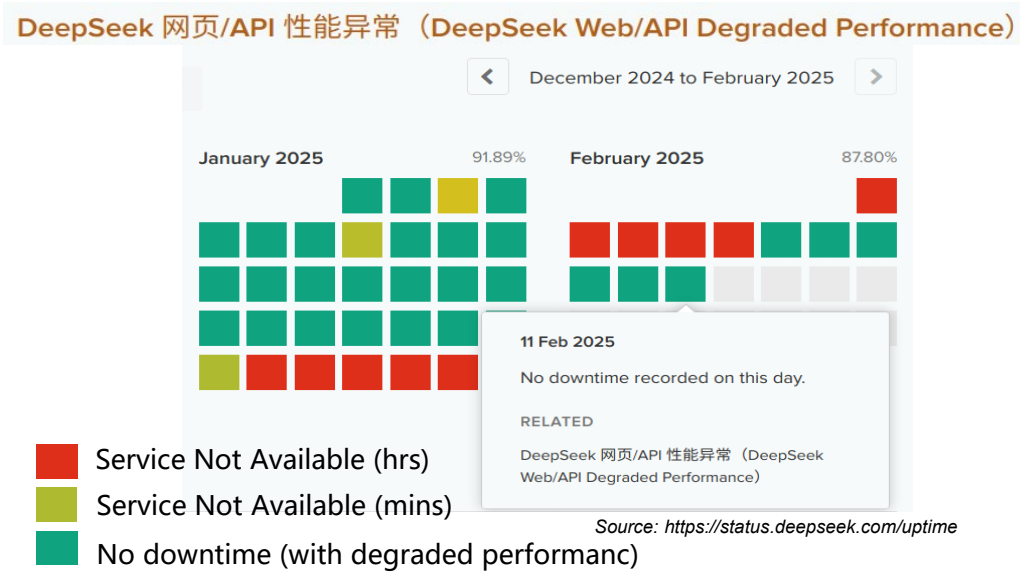
Source: https://api-docs.deepseek.com/news/news250120

AI models are evolving from chat tools to production and daily life tools, forming irreversible new business scenarios, bringing huge opportunities for model distribution and edge inference

Challenges and Opportunities Brought by DeepSeek's Inclusive Inference Services

DeepSeek has experienced multiple outages due to surging access volumes, frequently becoming unavailable. This highlights the inability of the current deployment architecture to handle massive concurrent user access. Operators need to leverage network resources, technical capabilities, and market advantages to seize opportunities in the AI era

- **Phenomenon: "*The server is busy. Please try again later*" issue remain unresolved**
  - Access issues began as early as late January
  - On February 6, DeepSeek had to suspend API service recharges
- **Cause: centralized deployment architecture cannot quickly adapt to the explosive growth of massive concurrent user requirements**
  - **User Scale:** Over 30 million DAUs on February 1
  - **API Integration:** Major applications like Alibaba Dingding, Huawei Xiaoyi, Honor, and WPS have integrated DeepSeek's API
  - **Deployment Method:** Centralized deployment with a single inference pool in China (*single-point bottleneck*)



Comparison with Gemini: Gemini, with a similar DAU scale, uses a distributed architecture, allowing users to access the nearest nodes with much less network latency, without access issues.

- **Carrier Network Opportunity: research a new distributed inference service architecture to handle the challenges of hundreds of millions of users and tens of millions of concurrent accesses for the AI inference age**
  - **Internet era:** CDN distributed web/video content from centralized pools to edge nodes, solving congestion and large-scale user access
  - **Inclusive AI era:** MDN (**M**odel **D**istribution **N**etwork) uses a *distributed* inference architecture to provide ubiquitous low-latency inference services

Deepseek Inference Server: Centralized Deployment

60.204.2.9	(In China)	Shanghai·Huawei Cloud
104.18.27.90	(Oversea)	USA/CA·CloudFlare
104.18.26.90	(Oversea)	USA/CA·CloudFlare

Only **3** IP addresses over the world for *chat.deepseek.com* and *api.deepseek.com*

Location	Fastest	Slowest	Avg
①	上海电信 7ms	福建厦门移动 44ms	22ms
②	广东广州电信 30ms	海南海口联通 57ms	43ms
③	湖北武汉电信 21ms	湖南益阳联通 38ms	29ms
④	天津联通 25ms	内蒙古电信 38ms	29ms
⑤	贵州贵阳电信 35ms	西藏拉萨移动 74ms	49ms
⑥	陕西西安电信 25ms	新疆乌鲁木齐移动 71ms	46ms
⑦	吉林长春移动 37ms	吉林长春电信 51ms	42ms

the fastest pings are over 20ms

Source: https://www.itdog.cn/ping

Doubao LLM Inference Server: Distrubution Deployment

Doubao: A top popular LLM in China and made by Bytedance

Location	Fastest	Slowest	
①	山东济南联通 <1ms	浙江宁波电信 16ms	6ms
②	海南海口电信 <1ms	广东潮州联通 29ms	7ms
③	河南郑州移动 <1ms	河南郑州电信 16ms	4ms
④	山西太原联通 <1ms	内蒙古移动 29ms	10ms
⑤	云南昆明电信 <1ms	西藏拉萨电信 36ms	9ms
⑥	甘肃兰州电信 <1ms	新疆昌吉电信 23ms	3ms
⑦	吉林长春电信 <1ms	黑龙江哈尔滨移动 3ms	1ms

117.21.223.147	0.85%	27.128.209.203	0.85%	183.136.203.179	0.85%	36.154.114.162	0.85%	39.175.118.192	0.85%	36.152.141.143	0.85%
144.7.113.209	0.85%	120.221.236.221	0.85%	112.83.164.210	0.85%	124.225.171.180	0.85%	118.182.226.71	0.85%	182.242.89.196	0.85%
223.221.216.194	0.85%	36.104.151.236	0.85%	119.97.179.164	0.85%	116.253.30.107	0.85%	113.142.76.154	0.85%	116.169.164.177	0.85%
111.6.252.111	0.85%	183.215.58.170	0.85%	219.147.78.84	0.85%	112.46.27.162	0.85%	122.189.220.170	0.85%	117.135.215.225	0.85%
113.201.52.155	0.85%	122.189.220.164	0.85%	116.169.166.243	0.85%	218.98.46.22	0.85%	221.204.58.44	0.85%	42.63.65.39	0.85%
111.161.204.209	0.85%	101.26.38.241	0.85%	182.89.223.236	0.85%	183.93.101.63	0.85%	36.147.28.239	0.85%	116.95.26.191	0.85%
112.29.193.186	0.85%	219.151.19.219	0.85%	101.69.194.212	0.85%	113.5.183.117	0.85%	182.91.252.63	0.85%	116.178.194.238	0.85%
112.47.6.241	0.85%	120.202.141.224	0.85%	111.51.140.205	0.85%	111.32.202.233	0.85%	112.132.36.239	0.85%	111.12.176.13	0.85%
27.221.97.229	0.85%	111.63.180.203	0.85%	43.242.166.216	0.85%	117.177.65.154	0.85%	61.168.100.229	0.85%	183.201.234.237	0.85%
221.229.203.237	0.85%	183.240.133.163	0.85%	120.253.62.232	0.85%	36.248.215.106	0.85%	112.30.217.111	0.85%	42.7.49.135	0.85%
111.29.56.246	0.85%	125.64.109.254	0.85%	122.228.79.237	0.85%	36.159.110.194	0.85%	171.15.36.240	0.85%	60.170.11.103	0.85%
140.249.109.239	0.85%	124.238.231.230	0.85%	36.248.33.202	0.85%	106.116.191.169	0.85%	106.8.159.241	0.85%	113.219.195.190	0.85%
111.43.171.241	0.85%	113.207.77.150	0.85%	222.75.44.192	0.85%	117.85.70.227	0.85%	36.156.169.70	0.85%	61.147.168.161	0.85%
120.233.182.225	0.85%	111.161.204.165	0.85%	36.151.63.235	0.85%	121.11.195.160	0.85%	58.215.50.248	0.85%	110.80.152.240	0.85%
36.189.15.240	0.85%	47.246.29.243	0.85%	118.181.55.70	0.85%	163.181.35.184	0.85%	163.181.201.237	0.85%	111.10.53.153	0.85%
156.59.188.130	0.85%	8.45.52.228	0.85%	163.181.66.243	0.85%	47.246.41.175	0.85%	163.181.201.238	0.85%	47.246.50.180	0.85%
79.133.176.222	0.85%	163.181.202.244	0.85%	42.202.219.193	0.85%	59.36.213.115	0.85%	116.177.228.144	0.85%	116.176.34.29	0.85%
180.130.113.199	0.85%	111.26.55.188	0.85%	120.201.244.115	0.85%	221.8.93.169	0.85%				

More than 100+ IP addresses in China. So it can use near inference servers to distrubute the inference requests

the fastest pings are all less 1ms

Average pings are less than 10ms

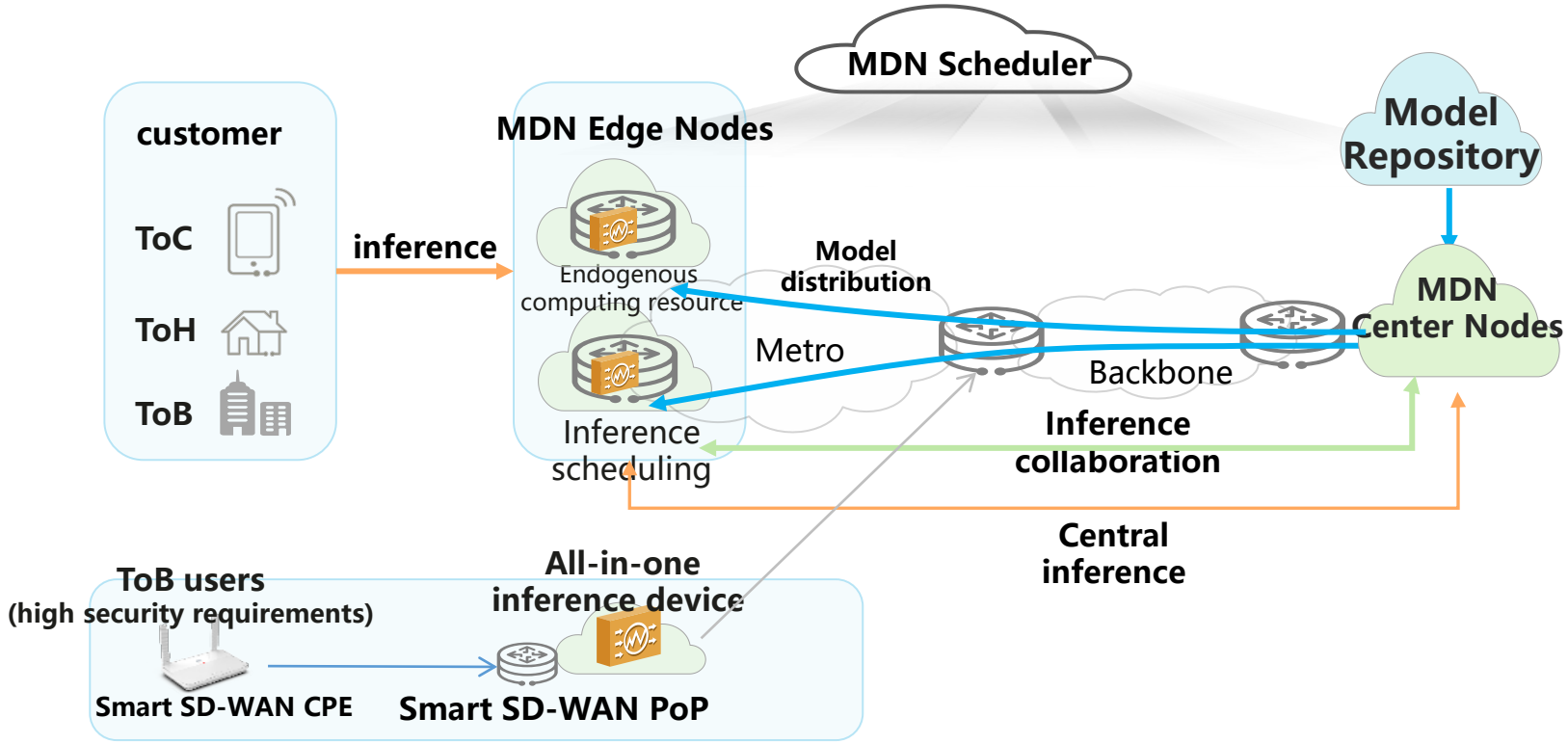
"Server busy" issue is never happened in Doubao

Source: https://www.itdog.cn/ping

Model Distribution Network (MDN) Target Architecture and Key Technologies

Model Distribution Network (MDN): A distributed architecture with edge intelligence, dynamic scheduling, and inference service recognition to achieve efficient model distribution and edge-cloud collaborative inference

1. **Model Repository**: Stores and manages different LLM versions
2. **MDN Center Nodes**: Deployed centrally to obtain models from the repository and distribute them to edge nodes
3. **MDN Edge Nodes**: Deployed in various geographical locations to receive and deploy models from center nodes, providing local edge inference services
4. **User Side Client**: End-user devices/apps launch inference requests
5. **MDN Scheduler**: Dynamically adjusts inference task distribution based on network conditions and device load



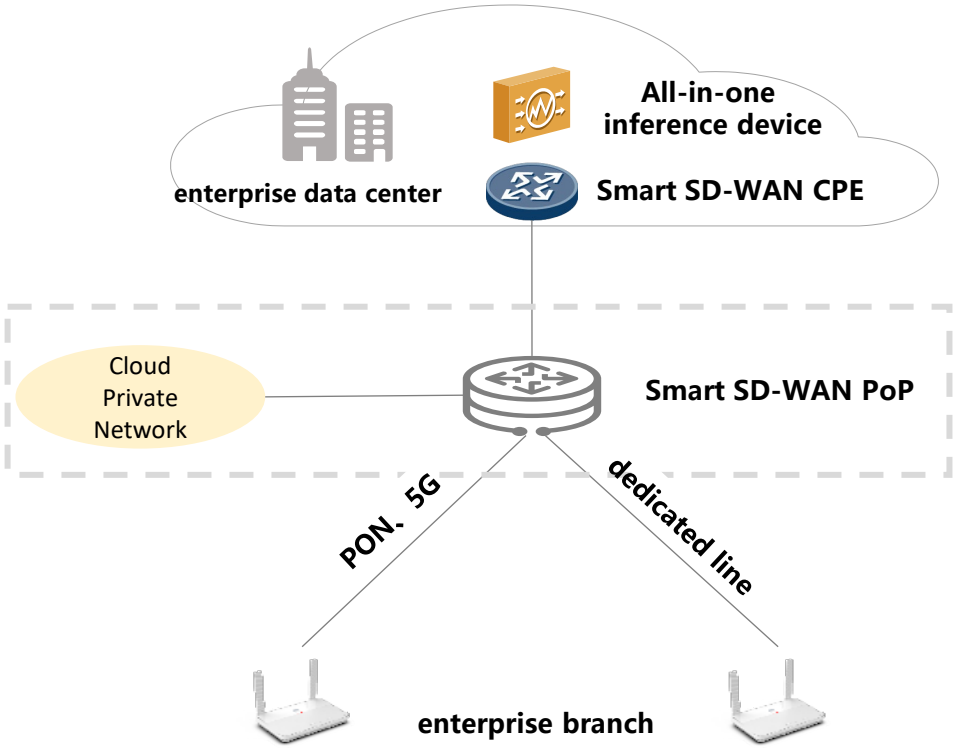
Key technologies:

1. Model inference with AI computing resource endogenous
2. Service identification
3. Collaborative inference
4. Models distribution
5. Large scale users certainty guarantee
6. Security

MDN Deployment Scenarios (1/4)

Scenario 1: 2B Standalone Deployment

Integrated Secure Inference Service with Computing and Network

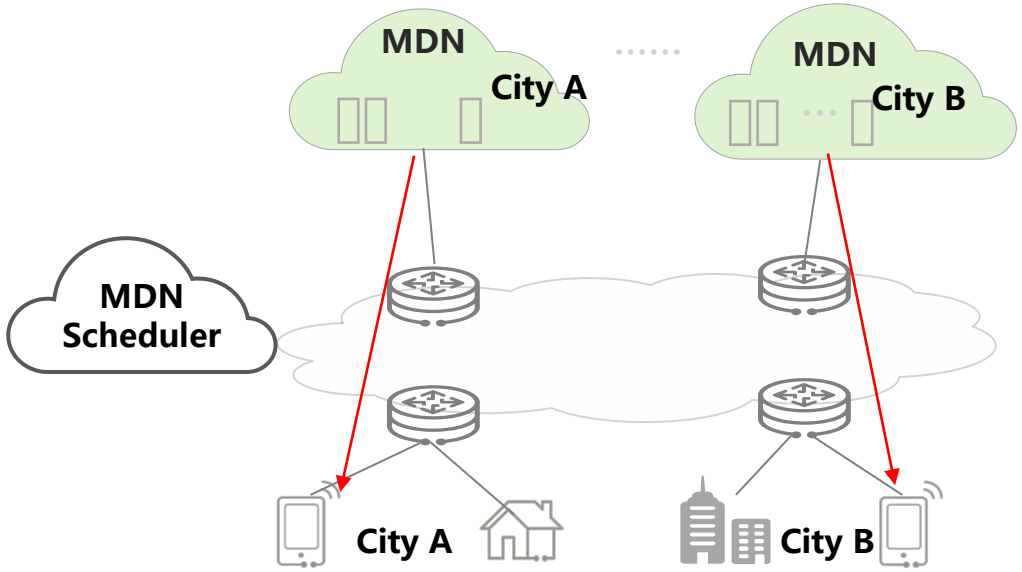


- For high-security 2B sectors, deploying an all-in-one inference device within the enterprise will protect the data.
- It is needed to protect the network accessing for enterprise branch accessing the private inference server from public internet or dedicated lines.
  - **"All-in-One Inference Server + Smart SD-WAN"** can provide secure inference service and **secure accessing**

MDN Deployment Scenarios (2/4)

Scenario 2: 2C&2H Dynamic Delivery

Balanced Inference Scheduling Service

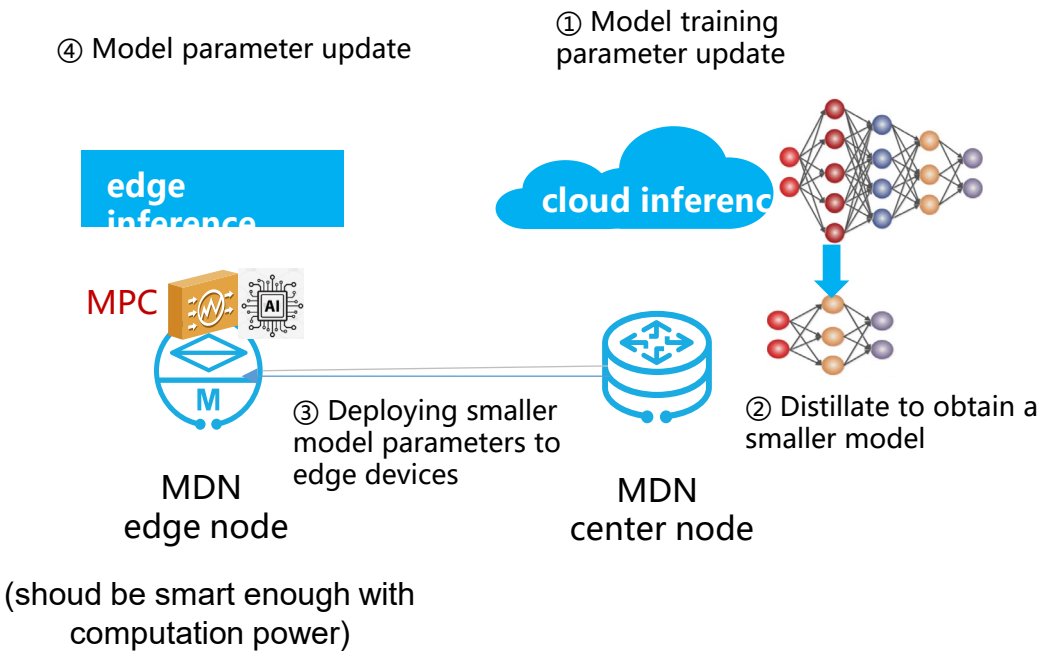


- Deploy the full size model (e.g., DeepSeek-R1 671B) in distributed MDN inference centers
- Use the **MDN scheduler** (DNS is a option) to **distribute inference tasks**
- **Dynamic Delivery**: the scheduler needs be aware inference resource load and user proximity for *multi-weighted factors* dynamic load balancing

MDN Deployment Scenarios (3/4)

Scenario 3: Collaborative Inference (Full size and Distilled)

Collaborative Inference Service for Large and Small Models

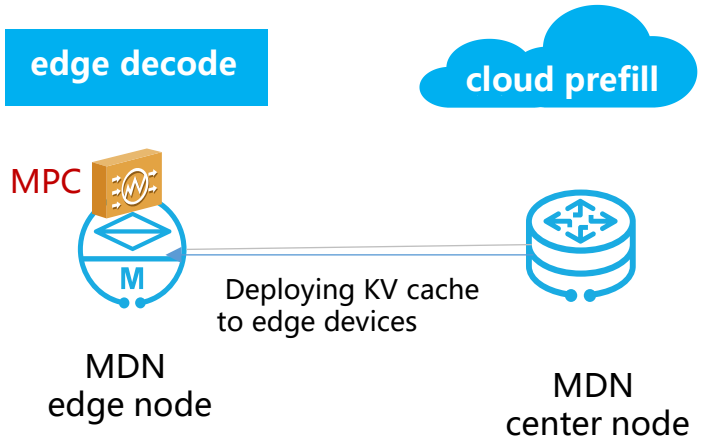


- Deploy the full size model (e.g., DeepSeek-R1 671B) in the inference service center
- Deploy distilled small models (e.g., 7B/70B) on edge nodes.
  - edge nodes can **schedule inference tasks to either the small models in edge or the full size model in the cloud** to reduce total computing resources
    - Edge node must support dynamic scheduling based on task type, complexity, and user priorities. So the edge node is not only a packet forwarding router (BNG) but also with computation and storage (*intelligent BNG*: **iBNG**)
  - Timely package the edge node inference tasks and results to center node for checking and updating.

MDN Deployment Scenarios (4/4)

Scenario 4: Collaborative Inference (Disaggregating Prefill and Decode)

Collaborative Inference Service for Prefill stage and Decode stage



- Deploy the full size model (e.g., DeepSeek-R1 671B) in the inference service center
  - Prefill phase (which is comput-intensive task) is executed on center node.
  - The center node focuses on generating the precomputed KV Cache, and then sends to MDN edge.
- Edge node leverages the precomputation results to perform decoding operations
  - Decode phase (which is memory-intensive task) is executed on edge node
- Disaggregating P/D **reduces** total computing resources, power consumption and network latency *without* loss any accuracy.
  - The network between center and edge should *zero packet loss, large*

中国移动
China Mobile

Thank You for Listening!

Corrections and Suggestions Welcome!