In-Memory Computing (Smartphone)

Redefining Smart Terminal Experience with 3D NAND In-Memory Computing

INC-WP-MOBILE-2026-01 V1.0

Preface: understanding iNc computing power

Preface: understanding iNc computing power

Preface: Description of iNc computing power

1. The essence of iNc computing power

Before reading this white paper, it is necessary to explain the computing power of iNc (in-NAND Computing) technology to avoid confusion with the computing power indicators of traditional digital chips:

Not "TOPS" computing power in the traditional sense:

The iNc 3D NAND chip completes vector similarity calculation through the summation of analog domain currents instead of a digital MAC (multiply and accumulate) unit, so the traditional TOPS (trillion operations per second) indicator is not applicable.

The real advantage is "in-place computing":

By directly completing vector matching in the storage array, avoiding data transfer to DRAM/CPU, it achieves a 43-fold improvement in energy efficiency and a 47-fold reduction in latency.

Applicable scenarios:

Data-intensive, bandwidth-sensitive tasks such as vector retrieval, similarity matching, and Embedding search, rather than general matrix operations.

2. Estimation of equivalent computing power of iNc in terminal equipment

Although iNc does not directly provide TOPS values, it can be calculated through equivalent computing power:

Hypothetical scenario: vector similarity retrieval (such as face search, product recommendation)

Solutions	Indicators	Traditional solutions (CPU/GPU)	iNc solutions (in-memory computing)
Processing 170,000 512-dimensional vectors	Computing power requirements	About 10-20 TOPS	Not applicable to traditional TOPS indicators
	Latency	> 20 seconds	< 0.5 seconds (47x acceleration)
	Energy consumption	> 100 J	< 2.5 J (43 times reduction)

Conclusion: The value of iNc does not lie in the TOPS value of general computing power, but in the ultimate energy efficiency and latency optimization of specific tasks.

3. The computing power expression convention in this white paper

To facilitate readers' understanding, the computing power values mentioned in this white paper are all "equivalent computing power", that is, achieving performance equivalent to that of traditional digital chips under specific tasks:

- iNc UFS on mobile: Provides retrieval performance equivalent to 10-20 TOPS, and the energy

efficiency ratio is increased by 5-10 times

- iNc Sensor Hub for smart glasses: Provides visual inference performance equivalent to 50-100 TOPS, power consumption < 500mW

1. Executive summary

1. Executive summary

1. Executive Summary

1.1 Project Vision

Lingcun is committed to upgrading memory from "passive data carrier" to "active computing unit with computing power" through 3D NAND in-memory computing (iNc) technology. In terminal devices such as smartphones, smart glasses, smart watches, etc., Lingcun uses UFS memory + AI operating system to replace the dominant position of the traditional "CPU+DDR" architecture in a large number of AI scenarios, fundamentally reconstructing the terminal's software and hardware stack, and bringing users an unprecedented AI native experience.

1.2 Core value proposition

- Energy efficiency revolution: By completing calculations such as AI inference and vector retrieval within the 3D NAND array, data transfer is reduced by 10-20 times, and system-level power consumption is reduced by 5-10 times, making it possible for complex AI tasks to be resident on the terminal 24 hours a day.

- Experience transition: Achieve end-side large model persistence, millisecond-level response, zero privacy leakage, and support real-time simultaneous interpretation, AI video generation/editing, multi-modal content creation and other scenarios that are difficult for the cloud to handle.

- Cost advantage: In mid-to-low-end and some mid-to-high-end models, using "iNc UFS + lightweight processor" instead of "large CPU + large DDR + traditional UFS" can reduce the overall BOM cost by 10-30%, making high-end AI experience popular on thousand-yuan machines.

- Product Reconstruction: From "App Center" to "AI Agent Center", mobile phones, glasses, and watches are no longer passive tools, but intelligent partners that can actively understand, create, and serve.

1.3 Current Progress

- Technology Maturity: Core iNc technology development has been completed, and in-memory computing prototype verification based on 3D NAND has been achieved.

- Application transplantation: We are entrusting Chongqing University to carry out AI application transplantation work to migrate AI applications based on GPU computing to iNc architecture.

- Industrial Cooperation: Conduct technical docking and solution verification with domestic leading mobile phone manufacturers, UFS controller manufacturers, and 3D NAND foundries.

2. Why now is the best time

2. Why now is the time to transform terminal experience

2. Why is now the best time to subvert the terminal experience?

2.1 The wave of terminal-side AI breaks out in full force

2026 is called the "year of device-side AI" by the industry. Artificial intelligence is shifting from "cloud dependence" to "device-generated". Terminal devices can independently complete the entire process of perception, calculation, and decision-making without relying on cloud servers.

Key Trends:

1. Privacy and Security Dimension Upgrade: Sensitive data "not leaving the device" has become standard, and users' trust in AI has been greatly improved.

2. Response delay compressed to millisecond level: Local inference compresses AI response delay from 200-500ms in the cloud to 10-50ms, achieving true "zero-perceived waiting".

3. Get rid of network constraints: AI functions are still available in weak/no network scenarios such as airplane mode, basement, and outdoors.

4. Multi-modal fusion: Multi-modal AI capabilities such as text, images, voice, and video are deeply integrated on the terminal to support more complex creation and interaction scenarios.

2.2 Three major bottlenecks of traditional architecture

Although the demand for end-side AI is booming, the traditional "CPU/NPU+DDR+UFS" architecture faces irreconcilable contradictions:

Bottlenecks	Performance	User pain points
Power Consumption Wall	In the continuous inference scenario, the SoC power consumption is > 5W, the phone is severely heated, and the battery life drops sharply	After using the AI assistant/video generation for 30 minutes, the phone loses 20-30% of its power and is hot to the touch
Bandwidth Wall	Even though the bandwidth of UFS 4.1 reaches 3.6GB/s, data is frequently moved during large model inference, which becomes a performance bottleneck	AI video generation/editing is stuck, and a 1-minute video needs to wait 5-10 minutes
Cost Wall	High-end SoC + large-capacity LPDDR + UFS costs > 1,500 yuan, which is unaffordable for mid- to low-end models	Thousand-yuan phones and wearable devices have almost no local AI capabilities

2.3 iNc technology: the key to breaking the three major bottlenecks

Through 3D NAND in-memory computing, Lingcun fundamentally solves the contradictions of traditional architecture:

Traditional architecture	iNc architecture	Improvement multiple
Data is frequently moved between SoC and UFS	Data is computed locally in the NAND array	Data movement is reduced by 10-20x

Traditional architecture	iNc architecture	Improvement multiple
AI inference power consumption 5-10W	AI inference power consumption 0.5-2W	Power consumption reduced by 5-10x
High-end SoC + large DDR cost > 1,500 yuan	Lightweight SoC + iNc UFS cost < 1,000 yuan	Cost reduction 30-50%
Cloud dependency, latency 200-500ms	Local inference, latency 10-50ms	Response speed increased by 5-20x

3. A disruptive user experience

3. A disruptive user experience with iNc

3. The disruptive user experience brought by iNc technology

3.1 Smartphone: From "Tool" to "Creativity Partner"

3.1.1 Real-time simultaneous interpretation: breaking language boundaries

When the user is on a video call with a foreign customer, the mobile phone translates the other party's English into Chinese in real time and displays subtitles. At the same time, the user's Chinese voice is converted into fluent English and sent to the other party in real time, with no perceived delay (< 300ms).

Technical implementation:

- The three major models of speech recognition (ASR) + machine translation (NMT) + speech synthesis (TTS) are all deployed in iNc UFS

- The inference time of the iNc solution is < 50ms, and the power consumption is < 1Wh/hour

Disruptive Value:

- Completely get rid of dependence on the cloud and can still be used on airplanes and in foreign network-free environments

- Zero privacy leakage, call content is completely processed locally

- Supports 150+ languages

3.1.2 AI video generation and editing: everyone is a creator

Vincent Video Scene: The user said "Generate a 30-second video: the beach at sunset", and the mobile phone generated a 1080p high-quality video in 2 minutes.

Intelligent editing scene: The user said "Delete all passers-by and change the sky to sunset effect", and the mobile phone completed the editing within 3 minutes.

Disruptive Value:

- The creation threshold has been reduced to zero, and professional-level video creation can be completed using natural language.

- Privacy protection, sensitive content is not uploaded to the cloud

- Cost savings, no need to purchase cloud video generation services

3.1.3 Local multi-modal large model: 24-hour personal AI assistant

The large model on the device side is deployed in iNc UFS, with a memory footprint of < 2GB and an inference delay of < 100ms/token, realizing a true "AI native mobile phone".

3.1.4 AI photo album and memory management

Pinpoint the target photo from 50,000 photos within 1 second and automatically generate a

personalized memory video.

3.2 Smart Glasses: From "Display" to "Second Brain"

3.2.1 Real-time environment understanding and AR navigation

The glasses can recognize street signs, buildings, and shops in real time, superimpose AR navigation arrows and multi-language translations, have a latency of <50ms, power consumption of <1W, and can work independently for 10+ hours.

3.2.2 Real-time gesture recognition and interaction

Gestures can control music playback, adjust volume, and switch AR interfaces. The inference delay is < 20ms, the power consumption is < 500mW, and the battery life is > 10 hours.

3.3 Smart Watch: From "Health Monitor" to "Life Guardian"

3.3.1 Real-time health abnormality warning

The watch continuously monitors heart rate, blood oxygen, and respiratory rate. When abnormal heart rhythm is detected, it immediately vibrates to wake the user. The delay is < 1 second and the power consumption is < 200mW.

3.3.2 Offline AI Voice Assistant

When hiking in mountainous areas (completely without network), the watch can record logs and answer questions, with power consumption < 300mW and battery life > 48 hours.

4. Product transformation

4. Product transformation

4. Product disruptive changes brought about by iNc technology

4.1 Three major revolutions in mobile phone product forms

4.1.1 From "App Center" to "AI Agent Center"

Traditional model: Users need to switch between hundreds of apps, and each app is an island of information.

New mode for iNc driver:

- Unified AI Agent: Users only need to say "Help me book a ticket to Shanghai tomorrow", and the AI Agent will automatically complete the entire process.

- Cross-application collaboration: AI Agent can access photo albums, address books, calendars, and emails at the same time

- Personalized Learning: AI Agent learns user habits and proactively provides suggestions

Technical basis: iNc UFS provides retrieval performance equivalent to 10-20 TOPS and supports 7-35B parameter end-side large models.

4.1.2 From "cloud dependence" to "device-cloud collaboration"

- 80% of AI tasks are completed locally, and only 20% of ultra-large-scale tasks are moved to the cloud

- Intelligent offloading: Privacy-sensitive tasks are forced to be processed locally

- AVAILABLE OFFLINE: Core AI functionality fully localized

4.1.3 From "high-end exclusive" to "universal access"

- Thousand yuan machine can also run large models: BOM cost reduced by 30-50%

- Performance democratization: Mid- to low-end users can also enjoy high-end features

- Reduce the digital divide in society

4.2 Product Redefinition of Smart Glasses

4.2.1 From "mobile phone accessories" to "independent terminals"

- Work independently: Built-in iNc Sensor Hub, providing visual reasoning performance equivalent to 50-100 TOPS

- All-day battery life: Power consumption < 500mW, battery life > 10 hours

- Complex scene support: Real-time translation, AR navigation, and gesture interaction are all localized

4.2.2 From "monitor" to "cognitive enhancer"

- Environment Understanding: The glasses understand the surrounding environment and proactively provide help

- Situation Awareness: Automatically adjust function priorities based on scenarios

- Memory Enhancement: Record what the user sees and hears throughout the day

4.3 Product transition of smart watches

4.3.1 From "Health Monitor" to "Life Guardian"

- Active warning: Immediate warning when abnormality is detected, which may save lives

- Personalized Modeling: Build a personalized health model for each user

- Medical Grade Accuracy: Run complex algorithms natively via iNc

4.3.2 From "mobile phone attachment" to "independent intelligence"

- Completely independent: Can independently complete voice assistant, health analysis, and schedule management

- Extreme scenes available: Become the only smart terminal in mountain climbing, diving, marathon and other scenarios

5. Technical solutions and product definition

5. Technical solutions and product definition

5. Technical solutions and product definitions

5.1 iNc (in-NAND Computing) technology core

Technical Principle:

A large number of vector/matrix operations in AI reasoning are "moved" into the 3D NAND array, and digital calculations are performed directly on the Flash storage unit to avoid frequent data transfer between the SoC and the memory.

Key Benefits:

1. Bandwidth bottleneck elimination: Computation is completed directly on the storage side

2. Energy efficiency improvement: The energy efficiency of 512Gb in-memory computing 3D NAND chip is 5-10 times higher than that of traditional architecture

3. Cost structure optimization: iNc logic units only occupy a small part of the additional area and have low marginal costs

5.2 Product form

Version	Target device	Capacity	Equivalent computing power	Power consumption	Typical applications
Ultimate Edition	High-end mobile phone/XR	512GB-1TB	Equivalent to 10-20 TOPS	< 2W	Large model inference, video generation
Popular version	Mid-to-high-end mobile phones	256GB-512GB	Equivalent to 5-10 TOPS	< 1.5W	Real-time translation, AI photography
Wearable version	Glasses/watches	64GB-128GB	Equivalent to 50-100 TOPS	< 500mW	Image feature extraction, gesture recognition

5.3 AI operating system (AI OS)

Core features:

1. Task Scheduling Engine: Intelligent judgment task allocation

2. Integrated storage and calculation memory management: Unified management of model weights and calculations

3. Developer SDK: Provides unified API

4. Device-cloud collaboration: Seamless switching between local and cloud

5.4 Replacing the boundaries of traditional architecture

Traditional architecture	iNc architecture	Applicable scenarios
Large CPU + Large DDR + Traditional UFS	Lightweight processor + iNc UFS	AI inference, vector retrieval
NPU integrated in SoC	iNc as an independent AI acceleration unit	Continuous reasoning, low power standby
Cloud large model	End-side large model (iNc bearer)	Privacy protection, low-latency interaction

Don’t do: Complex control flow, large amounts of branch logic, OS scheduling

Key points: Large-scale matrix/vector operations, model weight storage and reasoning, and Embedding library retrieval

6. Market opportunities and industry trends

6. Market opportunities and industry trends

6. Market opportunities and industry trends

6.1 Explosion of the terminal-side AI market

- 2026 is called the "year of device-side AI": Global consumer electronics manufacturers fully demonstrate their device-side AI capabilities, marking the comprehensive transformation of AI from "cloud dependence" to "device endogenous"

- Market Size: The global storage and computing integrated AI chip market size will be US$231 million in 2025, and is expected to reach US$44.33 billion in 2032, with a compound annual growth rate of 112.4%

- Chinese market: China's high-end AI chip market will grow by more than 60% year-on-year in 2026, and the local AI chip market share is expected to reach 50%

6.2 Terminal equipment shipments

- Smartphones: Global annual shipments are 1.2-1.4 billion units, the Chinese market is about 300 million units, and mid-to-high-end models account for 35-40%

- Smart Glasses/XR: Global annual shipments are expected to exceed 50 million units, with CAGR > 50% in the next 3 years

- Smart watches/wearables: Global annual shipments exceed 200 million units, and the Chinese market is about 60-80 million units

6.3 User pain points and needs

Pain points	User needs	iNc solutions
AI functions rely on the cloud, with high privacy risks	Sensitive data is processed locally	80% of AI tasks are completed locally
AI function consumes a lot of power and the phone gets hot	Long battery life, low heat	Power consumption reduced by 5-10 times
AI is not available in weak network/no network environments	Available offline	Core AI functions are fully localized
Video generation/editing is slow and costly	Fast and free generation	Local generation speed increased by 3-5 times
Language barriers limit cross-border communication	Real-time simultaneous interpretation	Latency < 300ms, supports 150+ languages

7. Competition and differentiation

7. Competition and differentiation

7. Competitive landscape and differentiated advantages

7.1 Comparison of existing solutions

Solution type	Advantages	Disadvantages
SoC integrated NPU	Mature ecosystem, high performance	High power consumption, data transfer bottleneck, high cost
Independent AI chip	Dedicated acceleration	Mainly oriented to data centers, low end-side penetration
Traditional LPDDR+UFS	Large-scale mass production, mature ecosystem	No integrated storage and computing capabilities, obvious power consumption bottleneck
Storage and computing integrated chip (data center)	High energy efficiency	Mainly focusing on servers/acceleration cards, with few terminal products

7.2 Differentiated advantages of Lancun

1. The world’s first 3D NAND in-memory computing solution for terminals: Focus on consumer terminals such as mobile phones/glasses/watches, with a market size 10-100 times that of data centers

2. UFS ecological compatibility: Based on the mainstream UFS interface and 3D NAND process, the OEM introduction threshold is low

3. Software and hardware integration solution: Provide a complete solution from chip to AI OS/SDK, reducing OEM development costs by more than 50%

4. Both advantages in cost and energy efficiency: The cost per unit of computing power is 30-50% lower than that of traditional NPU, and the power consumption per unit of computing power is 5-10 times lower

5. In-depth cooperation with top universities such as Chongqing University: AI application transplantation verification is in progress to accelerate productization

8. Technology roadmap and milestones

8. Technology roadmap and milestones

8. Technology roadmap and key milestones

8.1 Key Milestones

Milestone	Time	Success Sign
M1: Technical verification completed	2026 Q1 (completed)	Core iNc technology development completed, and the first batch of AI application transplants completed with Chongqing University
M2: The first UFS chip package	2026 Q2	The first iNc UFS chip was successfully packaged, and the developer kit was released
M3: Head OEM introduction	2026 Q3-Q4	Sign a joint development agreement with 1-2 head phone manufacturers
M4: mass production and launch	2027 Q1	The first batch of flagship mobile phones/smart glasses equipped with iNc UFS are launched
M5: Ecological Construction	2027 Q2-Q4	Developer community established, 100+ applications adapted to iNc
M6: Standardization	2028 Q1	Obtained a seat in the JEDEC/UFS Association, iNc interface becomes an industry standard

8.2 Current Progress (January 2026)

- Core technology: iNc architecture design, 3D NAND array computing unit, and controller logic have been developed and verified through simulation

- AI application transplantation: We are entrusting the School of Big Data and Software of Chongqing University to carry out the transplantation work of GPU-based AI applications to iNc architecture, including:

- CNN/Transformer inference operators (convolution, matrix multiplication, attention mechanism)

- Vector retrieval algorithms (FAISS, ANN)

- End-to-end applications such as image recognition, speech recognition, and video generation

- Industrial Cooperation: Technical docking with domestic head UFS controller manufacturers, 3D NAND foundries, and mobile phone manufacturers

9. Social value and long-term vision

9. Social value and long-term vision

9. Social value and long-term vision

9.1 Technology inclusiveness

- Lower the threshold for AI: Through iNc technology, even thousand-yuan phones can enjoy high-end AI experience and narrow the digital divide.

- Barrier-free communication: Real-time simultaneous interpretation function makes language no longer a communication barrier and promotes global cultural exchanges

9.2 Energy Saving and Emission Reduction

- Energy efficiency improvement: Compared with traditional architecture, in-memory computing reduces system-level energy consumption by 5-10 times, saving billions of kilowatt-hours of electricity globally every year.

- Comply with the "double carbon" goal: Support the national green computing power strategy

9.3 Data security and privacy protection

- Device-side AI: 80% of AI tasks are completed locally, and data does not need to be uploaded to the cloud, effectively protecting user privacy

- Compliance with regulations: Meets privacy protection requirements such as the Personal Information Protection Act

9.4 Long-term vision

2030: Lingcun iNc technology will become the standard configuration of more than 30% of smart terminals in the world, completely changing the way of human-computer interaction:

- Mobile phones become "creativity partners" and users can complete professional-level content creation using natural language

- Smart glasses become the "second brain", enhancing users' cognitive and decision-making abilities in real time

- Smart watches become "life guardians", proactively warning of health risks and saving millions of lives

Ultimate goal: Let AI change from "a luxury product in the cloud" to "everyone's basic right" and achieve true AI democratization

Conclusion

Conclusion

Conclusion

In-memory computing (iNc) technology represents a paradigm shift from "Von Neumann architecture" to "integrated storage and computing". We not only provide technical solutions, we are also committed to redefining the user experience of smart terminals:

- Make real-time simultaneous interpretation a standard feature of every mobile phone, breaking language boundaries

- Let AI video generation/editing change from a professional tool to a creative capability available to everyone

- Let smart glasses truly become the "second brain" and enhance human cognition

- Let smart watches evolve from "health monitors" to "life guardians", saving countless lives

In 2026, when device-side AI explodes in full force, Lingcun has completed core technology development and is conducting AI application transplantation verification with top universities such as Chongqing University. We sincerely invite industry partners, developers, and users to participate in this revolution in terminal experience and use in-memory computing technology to transform AI from a "luxury product in the cloud" to a "basic right for everyone."

Collect deposits and redefine the AI terminal, starting from deposit calculations.