Вычисления в памяти для автомобилей

Прорывные применения в автономном вождении

null V1.0

Предисловие: переосмысление вычислительной мощности

Традиционные TOPS: смысл и расчёт

В современной системе оценки производительности чипов искусственного интеллекта TOPS (тера операций в секунду) стал наиболее распространенным показателем вычислительной мощности. Будь то графический процессор NVIDIA, NPU Qualcomm или чипы автономного управления Horizon и Black Sesame, TOPS используется в качестве эталона производительности.

Математическое определение TOPS

Расчет TOPS основан на блоке MAC (умножение-накопление) внутри чипа, который является основным компонентом цифровых вычислений:

Формула расчета:

TOPS = (2 × количество блоков MAC × рабочая частота) / 10^{12}

Объяснение формулы:

• Количество блоков MAC: количество блоков операций умножения-накопления, работающих параллельно внутри чипа (например, NVIDIA Orin-X имеет 2048 блоков MAC INT8).

• Рабочая частота: рабочая частота устройства MAC (например, 1,6 ГГц).

• Коэффициент «2»: блок MAC выполняет одно умножение + одно сложение за один такт, что считается за 2 операции.

• Разделить на 10^{12}: преобразовать операции в секунду (OPS) в единицы TOPS.

Пример расчета (NVIDIA Orin-X):

• Количество блоков MAC: 2048 (точность INT8).

• Рабочая частота: 1,6 ГГц.

• TOPS = (2 × 2048 × 1,6 × 10^{9}) / 10^{12} = 6,55 TOPS (одноядерный)

• Orin-X имеет несколько вычислительных ядер с общей вычислительной мощностью 254 TOPS (INT8).

Физическая реализация блоков MAC

Традиционный цифровой блок MAC реализован на основе схем двоичных логических элементов:

Ввод

• A (вес): 8-битное целое число, например 01101100.

• B (значение активации): 8-битное целое число, например 10110011.

Процесс обработки

Шаг 1: Множитель (Множитель)

• Расширить до 8×8=64 логических элементов И (И-вентиль)

• Сжатие частичных произведений с помощью дерева Уоллеса.

• Вывод 16-битного результата произведения.

Шаг 2: Сумматор (Сумматор)

• Накопить результат продукта в аккумулятор (аккумулятор).

• Используйте сумматор с упреждающим переносом.

• Выведите окончательное накопленное значение

Характеристики энергопотребления:

• Потребляемая мощность при перевороте логического вентиля: ~10 пДж* на операцию MAC (INT8).

• Потребляемая мощность доступа к аккумуляторной памяти: около 20 пДж (цикл чтения-изменения-записи)

• Общая потребляемая мощность: ~30 пДж/MAC.

* Примечание: пДж (пикоджоуль, пикоджоуль) — это единица энергии, 1 пДж = 10^{-12}Джоуль, используемая для измерения минимального энергопотребления на уровне чипа.

Ключевые особенности:

• Численная точность расчетов: результаты полностью достоверны и нет накопления ошибок.

• Высокое энергопотребление: каждая операция требует, чтобы миллионы транзисторов перевернули состояния.

• Ограничено тактовой частотой: улучшение TOPS требует более высокой частоты или большего количества блоков MAC, что приводит к линейному увеличению площади чипа и энергопотребления.

Цена перемещения данных в архитектуре фон Неймана

Даже при наличии мощных массивов ячеек MAC традиционные архитектуры по-прежнему сталкиваются с фундаментальными узкими местами:

Путь потока данных (на примере автономного вождения):

1. Веса моделей хранятся в UFS/SSD (пропускная способность ~3,6 ГБ/с).

2. Загрузите гири в DRAM (LPDDR5, пропускная способность 204 ГБ/с).

3. DRAM → NPU Cache (пропускная способность ~1 ТБ/с)

4. Кэш → Массив MAC (выполнить вычисления)

5. Запишите промежуточные результаты обратно в DRAM.

6. Окончательный результат записывается обратно в хранилище.

[Энергопотребление в каждом звене]

• UFS → DRAM: ~35 пДж/бит

• DRAM → Кэш: около 5 пДж/бит.

• Кэш → MAC: около 1 пДж/бит.

• Расчет ПДК: ~30 пДж/ПДК.

• Общее энергопотребление: обработка данных занимает 60–90 %.

Замкнутый круг «переездного налога»:

• Количество параметров модели AI увеличено: с 10B (GPT-2) → 175B (GPT-3) → 1750B (GPT-4), а количество данных о весе увеличилось в 175 раз.

• Стена пропускной способности памяти: пропускная способность DRAM увеличивается всего в 2–3 раза (с 68 ГБ/с в DDR4 → 204 ГБ/с в LPDDR5).

• Узкое место в энергоэффективности: в сценарии автономного вождения 63,7% потребляемой мощности используется для передачи данных и только 28,5% используется для фактических расчетов.

Вывод: хотя традиционные TOPS обладают высокой вычислительной мощностью, они ограничены тем, что «данные не могут быть перемещены», а фактическое эффективное использование вычислительной мощности составляет всего 20-40%.

iNc: переход к аналоговым вычислениям

Основной принцип: закон Ома + закон Кирхгофа

Технология iNc использует аналоговые вычисления в памяти, полностью выходя за рамки концепции «цифровой MAC + передача данных»:

Физическая реализация (на основе флэш-массива 3D NAND):

Умножение матрицы: $Y = W \times X$

Физическое картирование

• W (весовая матрица): хранится в проводимости (G) ячейки памяти NAND.

→ Отрегулируйте проводимость, контролируя пороговое напряжение (Vth) вспышки.

• X (входной вектор): преобразуется в напряжение (B), приложенное к словесной строке (Word Line).

• Y (выходной вектор): ток, собираемый через битовую линию (I).

[Процесс расчета (выполняется в один этап)]

Шаг 1: Подайте входное напряжение V_1, V_2, …, V_n к каждой строке слова.

Шаг 2: Согласно закону Ома $I = G \times V$

→ Каждая вспышка генерирует ток $I_{ij} = G_{ij} \times V_j$.

Шаг 3: Согласно действующему закону Кирхгофа, ток на битовой линии автоматически суммируется.

→ $I_i = \Sigma(G_{ij} \times V_j) = (W \times X)_i$

Шаг 4. Преобразуйте ток в цифровые результаты с помощью АЦП (аналогово-цифрового преобразователя).

Основные преимущества

• Нулевая передача данных: веса хранятся во Flash, а вычисления выполняются внутри массива.

• Чрезвычайно высокая степень параллелизма: умножение матрицы M×N выполняется за время O(1) (по сравнению с O(M×N) для чисел)

• Повышение энергоэффективности: избегайте переворачивания цифровых логических элементов, что снижает энергопотребление в 10–100 раз.

• Уменьшение задержки: устранение задержки доступа к памяти DRAM (около 100 нс) и уменьшение задержки преобразования АЦП (около 1 мкс).

Расчетная производительность (чип 3D NAND 512 ГБ):

• Расчет подобия векторов (512 измерений): 50 мкс (по сравнению с цифровым решением 2,35 мс)

• Повышение энергоэффективности: в 43 раза (по сравнению с ЦП/ГП)

• Сокращение задержки: в 47 раз

Почему традиционные TOPS неприменимы к iNc

Сравнение существенных различий:

Критерий	Традиционные цифровые вычисления (MAC)	Аналоговые вычисления iNc
Вычислительный блок	Цифровые логические элементы (И, сумматоры)	Физические элементы (проводимость Flash и суммирование токов)
Точность вычислений	Детерминированные значения (например, INT8, 256 значений)	Непрерывные аналоговые значения (ограничены АЦП; 8 бит — 256 уровней)
Параллелизм	Зависит от числа MAC (например, 2048 MAC)	Параллельность всего массива (матрица M×N за один шаг)
Источник энергоэффективности	Высокочастотное тактирование (например, 1,6 ГГц)	На основе физических законов (естественное сложение токов)
Расчёт TOPS	2 × число MAC × частота / 10^{12}	Неприменимо (нет блоков MAC и тактовой частоты)
Основное преимущество	Универсальность и управляемая точность	Высокая эффективность задач с большим объёмом данных

Почему оценка iNc через TOPS некорректна

1. Физическая архитектура другая: у iNc нет цифрового блока MAC и глобальной тактовой частоты (аналоговые схемы представляют собой системы с непрерывным временем).

2. Различные вычислительные парадигмы: «умножение одной матрицы» iNc соответствует «тысячам операций MAC» цифровой системы, уменьшая временную сложность и

размерность.

3. Цели оптимизации разные: TOPS оптимизирует «количество операций в секунду», а iNc оптимизирует «время выполнения одной задачи и энергопотребление».

Аналогия:

• TOPS = показатель «количества кирпичей, которые рабочий перемещает в час» (подходит для операций на сборочной линии).

• iNc = Заменить «ручное перемещение кирпичей» на «автоматическое перемещение конвейерной ленты» (выход за пределы оценочного измерения «количество перемещенных кирпичей»).

Оценка эквивалентной производительности iNc

Хотя iNc не предоставляет традиционные значения TOPS, он может рассчитать эквивалентную вычислительную мощность через задержку завершения задачи:

Сценарий: поиск векторного сходства (набор данных VGGFace2, 170 000 512-мерных векторов признаков лица)

Традиционное решение (ЦП/ГП):

• Алгоритм: вычислить косинусное сходство для 170 000 векторов один за другим (требуется 512 умножений + 512 сложений × 170 000).

• Всего операций: 512 × 2 × 170 000 = 174 миллиона операций.

• Требуемая вычислительная мощность (время выполнения 20 секунд): $1{,}74\times10^{8}$ / 20 с = 8,7 GOPS ≈ 10 TOPS (с учетом накладных расходов на передачу данных)

• Фактическое энергопотребление: 100 Дж (включая доступ к DRAM + расчет ЦП)

Решение iNc (вычисления в памяти):

• Алгоритм: вектор запроса применяется как напряжение, 170 000 весовых векторов сохраняются как проводимость флэш-памяти, а сходство рассчитывается параллельно.

• Задержка: <0,5 секунды.

• Потребление энергии: < 2,5 Дж (снижение в 43 раза)

• Эквивалентная вычислительная мощность: если преобразовать ее в термины «время выполнения той же задачи», она эквивалентна 10 TOPS × (20 с / 0,5 с) = 400 TOPS (но физическая архитектура совершенно другая).

Вывод:

Ценность iNc заключается не в том, «сколько у вас TOPS», а в:

1. Чрезвычайная энергоэффективность для конкретных задач: для таких задач, как поиск векторов, сопоставление по сходству и поиск встраивания, энергоэффективность увеличивается в 43 раза.

2. Снижение задержки. Устраните узкие места при передаче данных и сократите задержку в 47 раз.

3. Устранение узких мест в пропускной способности: веса находятся во Flash и не требуют использования полосы пропускания DRAM.

Применимые сценарии:

• Задачи с интенсивным использованием данных: крупномасштабный векторный поиск, сопоставление признаков изображения и видео, семантический поиск.

• Задачи, чувствительные к пропускной способности: объединение нескольких датчиков, объединение функций BEV, управление кэшем Transformer KV.

• Сценарии с низким энергопотреблением: носимые устройства, умные очки, периферийные компьютеры в автомобиле.

• Не подходит для: общенаучных вычислений, высокоточного численного моделирования (ограничено точностью АЦП, обычно 8–12 бит).

Автономное вождение: от пиковой мощности к эффективности

Дилемма вычислительной мощности автономного вождения

Порочный круг традиционных решений:

[Требования к автономному вождению L4]

• 8 камер (8 МП при 30 кадрах в секунду) → необработанные данные 760 МБ/с.

• 1 лидар на 128 строк → данные облака точек со скоростью 120 МБ/с.

• Радары 5-миллиметрового диапазона → данные 4D-изображений со скоростью 300 МБ/с

• Всего: поток данных датчика 1,18 ГБ/с.

[Обработка традиционной архитектуры]

Шаг 1: Данные с датчика → SoC DRAM (потребление энергии составляет 40%)

Шаг 2: Расчет DRAM → NPU (потребляемая мощность составляет 30%)

Шаг 3: Расчет NPU (потребляемая мощность составляет 30%)

→ Фактическое эффективное потребление вычислительной мощности составляет < 30 %, а остальные 70 % используются для передачи данных.

[OEM-производители вынуждены накапливать вычислительные мощности]

• Двойной NVIDIA Orin-X (508 TOPS) → Потребляемая мощность 120 Вт

• Фактическое эффективное использование вычислительной мощности: 20–40 %.

• Влияние дальности: 10–15 %

• Стоимость спецификации: > 5000 юаней.

Основное противоречие: дело не в том, что «TOPS недостаточно», а в том, что «TOPS не используется» — узкое место в пропускной способности памяти заставляет вычислительную мощность «умирать от голода».

Как iNc преодолевает узкие места

Философия дизайна: переместите вычисления туда, где находятся данные, а не перемещайте данные в вычислительные блоки.

Архитектурные инновации

Крайовой датчик (концентратор датчиков iNc):

• Встроенная память iN eUFS емкостью 64–128 ГБ на каждый модуль камеры/радара.

• Полное извлечение объектов на месте (первые несколько слоев YOLO/Transformer).

• Сжатие данных 10:1 (передаются только векторы признаков, а не исходные изображения)

• Потребляемая мощность: < 1 Вт/модуль.

Центральная вычислительная платформа (iNc Auto-L4):

• 512 ГБ–1 ТБ iNc UFS (вес модели хранилища + исторические характеристики корпуса)

• Объединение функций нескольких датчиков осуществляется в массиве NAND (суммирование тока аналоговой области).

• Работает с облегченной SoC (50–100 TOPS) для управления потоком управления.

• Энергопотребление системы: < 80 Вт (по сравнению с традиционными 150 Вт)

[Улучшение производительности]

• Коэффициент энергопотребления при передаче данных: снижен с 63,7% до < 10%.

• Эффективное использование вычислительной мощности: увеличено с 20–40% до 70–90%.

• Задержка слияния нескольких датчиков: уменьшена с 80–150 мс до < 30 мс.

• Вывод большой модели на конечной стороне (35B VLA): уменьшено с 200 мс до < 100 мс.

Экономическая ценность

• Стоимость спецификации: снижена на 30–40 % (уменьшается потребность в высокопроизводительной SoC + DRAM большой емкости).

• Повышенная энергоэффективность: энергопотребление снижается на 30–50% при той же производительности, а влияние на запас хода снижается с 15% до 8%

• Цикл НИОКР: от «модульного решения (2 года)» до «комплексной модели (< 1 года)»

Основной вывод: вычисления в памяти iNc обеспечивают реальный технический путь развития автономного вождения от уровня L3 до уровня L5 за счет устранения узких мест при передаче данных, улучшения использования вычислительной мощности и снижения энергопотребления системы, а также значительного снижения порога коммерциализации.

1. Вычислительные проблемы автономного вождения

1.1 Потребность в вычислениях и фактическая загрузка

Согласно данным расчетов соответствующих организаций, требования к вычислительной мощности для каждого уровня автономного вождения следующие:

Уровень автономного вождения	Теоретическая потребность в вычислениях (TOPS, INT8)	Расчётная фактическая загрузка	Типичная мощность
L2 (расширенная система помощи водителю)	2–10 TOPS	60–80%	10–20W
L2+ (расширенная помощь водителю)	10–30 TOPS	40–60%	20–40W
L3 (условная автоматизация)	30–60 TOPS	30–50%	40–80W
L4 (высокая автоматизация)	200–500 TOPS	20–40%	80–150W
L5 (полная автоматизация)	1000–2000+ TOPS	Расчётно < 30%	150–300W

Основные выводы

1. Использование вычислительной мощности снижается с увеличением сложности. Хотя система уровня L4 имеет пиковую вычислительную мощность более 500 TOPS, она ограничена такими факторами, как узкие места в пропускной способности памяти, накладные расходы на передачу данных и рассинхронизация времени нескольких датчиков. Фактический эффективный коэффициент использования составляет всего 20–40%.

2. Передача данных становится черной дырой энергоэффективности: согласно анализу энергопотребления NVIDIA Orin-X, 63,7% потребляемой мощности используется для передачи данных между DRAM и вычислительными блоками (DDR5 ↔ NPU), а доля потребляемой мощности, фактически используемая для операций MAC, составляет < 30%.

3. Цикл «инфляции вычислительной мощности»: чтобы компенсировать низкую загрузку, OEM-производители вынуждены устанавливать больше микросхем (например, двух/четырехъядерный Orin-X), что приводит к резкому росту стоимости, энергопотребления и рассеивания тепла.

1.2 Ограничение пропускной способности памяти в автономном вождении

Характеристики потока данных типичных систем L4:

Источник данных	Разрешение / частота дискретизации	Поток исходных данных	Поток признаковых данных
Восемь камер	8 мегапикселей при 30 кадрах/с	760 MB/s	150 МБ/с (после сжатия)
Один 128-канальный лидар	128 каналов при 10 Гц	120 MB/s	40 MB/s

Источник данных	Разрешение / частота дискретизации	Поток исходных данных	Поток признаковых данных
Пять радаров миллиметрового диапазона	Визуализирующий радар 4D	300 MB/s	80 MB/s
Итого	-	1.18 GB/s	270 MB/s

Узкие места памяти на примере NVIDIA Orin-X

• Теоретическая пропускная способность LPDDR5: 204 ГБ/с (двухканальный 6400 MT/с)

• Эффективная пропускная способность при одновременном доступе к нескольким датчикам: < 100 ГБ/с (загрузка < 50 %).

• Причина:

1. Конфликт банков: несколько потоков данных датчиков одновременно обращаются к одному и тому же банку DRAM, что приводит к ожиданию в очереди.

2. Промах буфера строки. Произвольный доступ приводит к частой активации/предварительной зарядке строки, каждый раз добавляя дополнительную задержку в 30–50 нс.

3. Накладные расходы на переключение чтения и записи: GPU/NPU переключается между чтением входных данных и записью промежуточных результатов, каждое переключение добавляет задержку 20–30 нс.

Последствия:

• Вычислительный блок простаивает: время ожидания поступления данных GPU/NPU составляет 40–60% от общего времени выполнения, а вычислительная мощность «голодает».

• Напрасное энергопотребление: энергопотребление при передаче данных составляет 63,7%, что намного превышает фактическое энергопотребление вычислений (28,5%).

• Неконтролируемая задержка: задержка мультисенсорного слияния ухудшается с теоретических 30 мс до измеренных 80–150 мс, что не соответствует требованиям для экстренного обхода препятствий (< 50 мс).

1.3 Невозможный треугольник больших моделей на устройствах

Автономное вождение развивается от модульного решения (восприятие-планирование-контроль) к комплексной большой модели (VLA: видение-язык-действие). Однако развертывание на конечной стороне сталкивается с серьезными проблемами:

Тип модели	Число параметров	Требуемый объём KV-кэша	Задержка инференса (Orin-X)	Мощность	Проблема
Традиционная модульная конструкция	10 млрд (суммарно по моделям)	< 500 MB	50–80ms	60W	Слабое обобщение и ошибки в редких сценариях
Сквозная VLA	35–70B	2–5 ГБ/кадр	150–300ms	> 120W	Чрезмерные задержка и энергопотребление

Пропускная способность KV-кэша на примере VLA с 35 млрд параметров

• Для каждого кадра вывода требуется доступ к данным KV-кэша объемом 2–5 ГБ (механизм самообслуживания Transformer).

• Ограничение пропускной способности LPDDR5: даже если теоретическая пропускная способность составляет 204 ГБ/с, фактическая доступная пропускная способность составляет < 100 ГБ/с, а доступ к данным объемом 3 ГБ занимает > 30 мс.

• Генерация авторегрессии: инструкции действий должны генерироваться токен за токеном и не могут быть распараллелены, что приводит к линейному накоплению задержек.

• Результат: задержка вывода VLA на стороне терминала составляет > 150 мс, что не соответствует требованиям L4 к реальному времени (< 50 мс).

2. Технологический путь iNc

2.1 Основные технические принципы

Технология Lingcun iNc принципиально исключает обработку данных, выполняя вычисления аналоговой области внутри массива 3D NAND:

Сравнение архитектуры:

Тип архитектуры	Путь данных	Узкое место пропускной способности	Энергоэффективность	Задержка
Традиционная архитектура	Хранилище → DRAM → вычисления → DRAM → хранилище	Ограниченная пропускная способность DRAM (204 ГБ/с)	Базовый уровень	Базовый уровень
Вычисления вблизи памяти	Хранилище → HBM → вычисления → HBM → хранилище	Пропускная способность HBM (4,9 ТБ/с)	0.6x	2–3x
Вычисления в памяти iNc	Вычисления внутри хранилища (без перемещения данных)	Нет узкого места (аналоговое суммирование токов)	0.02x	43x

Расчётная производительность кристалла iNc 3D NAND 512 Гбит

Тип задачи	Набор данных	Задержка iNc	Задержка традиционного решения	Ускорение	Повышение энергоэффективности
Поиск сходства векторов	VGGFace2 (170 000 лиц, признаки размерности 512)	50 мкс (одно сопоставление)	2.35 msCPU	47x	43x
Возврат 600 лучших результатов после пулинга	Как указано выше	< 500 мс (поиск по всей базе)	> 20 s	40x	35x
Объём передаваемых данных	Как указано выше	Снижение в 14,3–33,3 раза (выборка с пулингом)	Базовый уровень	-	-

Данные о производительности

Потеря точности: всего 0,43% (93,39% против идеального значения 93,82%), влияние на практическое применение незначительно.

2.2 Архитектурная роль iNc в автономном вождении

Основная концепция дизайна: iNc не заменяет SoC, а служит гетерогенным сопроцессором и сотрудничает с SoC:

Тип задачи	Характеристики вычислений	Ответственный блок	Причина
Решения управления потоком	Много ветвлений и сложная логика	SoC CPU50–100 DMIPS	CPU эффективен в предсказании ветвлений и переключении

Тип задачи	Характеристики вычислений	Ответственный блок	Причина
			контекста
Планирование траектории	Небольшие матричные операции и оптимизация в реальном времени	SoC NPU50–100 TOPS	Требуется детерминированное выполнение с малой задержкой
Векторный поиск	Масштабные вычисления сходства (например, семантические карты)	iNc	Большие объёмы данных и зависимость от пропускной способности
Объединение признаков	Объединение признаков BEV от нескольких камер	iNc	Доступ к множеству прошлых кадров разных датчиков
Управление KV-кэшем	Чтение и запись KV-кэша трансформера	iNc	Частые обращения к памяти, ограничение пропускной способности
Загрузка весов	Загрузка и повторное использование параметров моделей 35–70 млрд	iNc	Не занимает DRAM

2.3 Показатели на основе данных вычислительного кристалла памяти

Обратитесь к данным определенной модели внутренней памяти и вычислительного интегрированного чипа:

Показатель	Неназванная отечественная модель	iNc Auto-L4 (прогноз STORLEAD по аналогичным архитектурам)	Основание оценки улучшения
Техпроцесс	12nm	12 нм (автомобильный класс)	-
Физическая вычислительная мощность	256 TOPS (INT8)	Эквивалент 500–800 TOPS (векторный поиск)	Аналоговые вычисления iNc обходят ограничения цифровых MAC
Типичная мощность	35W	< 60W	Мощность перемещения данных iNc < 10% (традиционно: 63,7%)
Показатель энергоэффективности	7.3 TOPS/W	10–15 TOPS/W	15,6 TOPS/Вт по аналогии с академическими работами
Фактическая пропускная способность	ResNet-50: 8700 кадров/с (пакет=1)	BEVFormer: расчётные 150–200 кадров/с	iNc сокращает обращения к памяти при объединении признаков
Проверенная задача	Сети обработки облаков точек и BEV	Как выше + KV-кэш большой модели VLA	-

iNc Auto-L4 в сравнении с двумя NVIDIA Orin-X

Показатель	Два кристалла Orin-X	iNc Auto-L4 + облегчённая SoC	Улучшение
Пиковая вычислительная производительность	508 TOPS (INT8)	SoC: 50 TOPS + iNc: эквивалент 500–800 TOPS	Ускорение в 5–10 раз для отдельных задач
Фактическая загрузка	20–40%	70–90%	iNc устраняет ограничение пропускной способности
Мощность системы	120 Вт (два кристалла)	< 80WSoC 20W + iNc 60W	Снижение на 33%

Показатель	Два кристалла Orin-X	iNc Auto-L4 + облегчённая SoC	Улучшение
Доля мощности на перемещение данных	63.7%	< 10%	Снижение в 6 раз
Задержка объединения BEV	80–120 мс (измерено, ограничение DRAM)	< 30 мс (локальное объединение)	Снижение в 3–4 раза
Задержка инференса VLA	150–300ms	< 80 мс (локальный доступ к KV-кэшу)	Снижение в 2–4 раза
Стоимость компонентов	> 5000 юаней (два кристалла + LPDDR5 16 ГБ)	< 3500 юаней (одна SoC + iNc UFS)	Снижение на 30–40%

3. Преобразующие сценарии применения

3.1 Объединение данных датчиков: детерминированная работа в реальном времени

3.1.1 Временная катастрофа традиционных решений

Типичная система L4 (8 камер + 1 лидар + радар 5 мм):

[Время поступления данных датчика (фактическое измерение)]

• Камера 1: T0 = 0 мс

• Cam2: T0 + 3 мс (разница во времени срабатывания)

• Камера 3: T0 + 6 мс

• ...

• LiDAR: T0 + 15 мс (задержка обработки облака точек)

• Радар1: T0 + 20 мс.

• ...

Процесс обработки данных (традиционная архитектура):

1. Данные каждого датчика → DRAM (задержка 5–15 мс).

2. Выравнивание временных меток + извлечение признаков → NPU (задержка 30–50 мс)

3. Объединение функций BEV → NPU (задержка 20–40 мс, ограничена пропускной способностью DRAM)

4. Обнаружение цели + прогнозирование траектории (задержка 20–30 мс)

Общая задержка: 80–150 мс (фактическое измерение)

Вопрос:

• Неопределенность времени: время поступления данных датчиков не синхронизировано, а выравнивание временных меток требует буферизации нескольких кадров, что увеличивает задержку.

• Конфликт за полосу пропускания: несколько датчиков одновременно получают доступ к DRAM, что приводит к конфликтам банков и фактическому использованию полосы пропускания < 50%.

• Недостаточная производительность в режиме реального времени: общая задержка > 100 мс, невозможность выполнения требований по предотвращению аварийных препятствий (< 50 мс).

3.1.2 Предобработка на периферии и объединение данных в памяти

Архитектурный дизайн

Модуль Edge Sensor (iNc Sensor Hub, eUFS 64–128 ГБ):

• iNc Sensor Hub встроен в каждый модуль камеры/радара

• Завершено на месте: шумоподавление, извлечение признаков, обнаружение цели (YOLO/Transformer).

• Вывод: передается только вектор признаков (512 измерений) + целевой кадр, а объем данных сжимается до 10 % от исходного.

• Задержка: < 10 мс (изображение 8 миллионов пикселей → 512-мерные элементы)

Центральная вычислительная платформа (iNc Auto-L4, UFS от 512 ГБ до 1 ТБ):

• Получать векторы признаков каждого датчика и сохранять их в iNc UFS.

• Объединение функций BEV: векторно-взвешенное объединение выполняется внутри массива NAND (суммирование тока аналоговой области).

• Задержка: < 20 мс (по сравнению с традиционным решением 40–60 мс)

• Результат: объединенные функции BEV → SoC NPU для обнаружения целей и прогнозирования траектории.

Повышение производительности

Этап	Задержка традиционного решения	Задержка решения iNc	Улучшение
Предобработка данных датчиков	5–15 мс (передача исходных данных)	< 10 мс (выделение признаков на периферии)	1.5–3x
Передача данных	1.18 GB/s → DRAM	270 МБ/с → центральный iNc (сжатие 4×)	Снижение сетевой нагрузки на 75%
Объединение признаков BEV	40–60 мс (узкое место DRAM)	< 20 мс (объединение в памяти)	2–3x
Общая задержка системы	80–150ms	< 50ms	2–3x
Мощность	150 Вт (30 Вт датчики + 120 Вт центр)	< 90 Вт (15 Вт датчики + 75 Вт центр)	Снижение на 40%

Проверка в реальных сценариях

По результатам испытаний одной отечественной модели на беспилотном автомобиле:

• Вывод сети BEV: частота кадров увеличивается с 8 кадров в секунду (Orin-X) до 25 кадров в секунду (H30), а задержка снижается в 3 раза.

• Обработка облака точек: сеть PointPillars обрабатывает 128-строчный лидар, при этом задержка снижается с 50 мс до 18 мс.

3.2 Сквозные модели VLA в автомобиле

3.2.1 «Тройная дилемма» большой модели VLA

Модель Vision-Language-Action (VLA) является ключевой технологией для автономного вождения L5, которая может генерировать команды управления непосредственно на входе датчика без необходимости использования промежуточных модулей.

Типичные характеристики модели VLA (параметры 35B):

Компонент	Число параметров	Использование памяти (FP16)	Временные характеристики инференса
Визуальный кодировщик	5BViT-H	10 GB	Восемь параллельных потоков камер
Языковой декодировщик	25BLLaMA-30B	50 GB	Авторегрессионный вывод по одному токену
Модуль формирования действий	5B	10 GB	Выходы угла руля, акселератора и тормоза
KV-кэш	-	2–5 ГБ/кадр	Обращение ко всей истории KV для каждого кадра

Три проблемы традиционных решений на Orin-X

1. Задержка загрузки параметров:

• Вес модели 35B (70 ГБ FP16) невозможно поместить в LPDDR5 (32 ГБ), и их необходимо загружать партиями.

• Загрузка параметров размером 70 ГБ из UFS 4.0 (3,6 ГБ/с) занимает >19 секунд, что делает вывод в реальном времени невозможным.

2. Черная дыра пропускной способности кэша KV:

• Для каждого кадра вывода требуется доступ к данным кэша KV объемом 2–5 ГБ.

• Фактическая пропускная способность LPDDR5 составляет < 100 ГБ/с, а доступ к 3 ГБ занимает > 30 мс.

• Для вывода 30 кадров в секунду требуется, чтобы один кадр выполнялся каждые 33 мс, а доступ KV исчерпал временной бюджет.

3. Задержка авторегрессионной генерации:

• Инструкции к действию должны генерироваться жетон за жетоном (например, «повернуть налево на 15°, ускориться до 40 км/ч» требуется 20+ жетонов).

• Для каждого поколения токена требуется полный вывод (включая доступ к кэшу KV), а общая задержка составляет > 150 мс.

Результаты: Традиционный Orin-X использует модель VLA 35B с задержкой вывода > 200 мс и энергопотреблением > 150 Вт, что не соответствует требованиям L4/L5.

3.2.2 Постоянные веса модели и локальное управление KV-кэшем

Проект архитектуры (iNc Auto-L5, UFS 1–2 ТБ):

[Метод развертывания модели]

• Веса моделей 35–70B (смешанная точность FP16/INT8), хранящиеся в iNc UFS (нет необходимости загружать в DRAM)

• Во время вывода веса представляют собой моделируемую матрицу домена, умноженную непосредственно на значения активации в массиве NAND.

• Промежуточные значения активации (< 2 ГБ) хранятся в LPDDR5 малой емкости

(достаточно 8 ГБ).

[Управление KV-кешем]

• Кэш KV (2–5 ГБ/кадр), хранящийся в iNc UFS.

• Когда вычисляется самовнимание Трансформатора, вектор запроса непосредственно выполняет расчет подобия с вектором KV в массиве NAND (скалярное произведение области моделирования).

• Избегайте загрузки кэша KV в DRAM, устраняя узкие места в полосе пропускания.

• Задержка: < 50 мс (по сравнению с традиционным решением> 150 мс)

Прирост производительности: теоретические и академические оценки

Этап	Традиционное решение (Orin-X)	iNc Auto-L5	Улучшение
Загрузка модели	19+ секунд (из UFS в DRAM)	Нулевая загрузка (веса постоянно находятся в iNc)	Неограниченный
Доступ к KV-кэшу	30 мс/кадр (ограничено пропускной способностью DRAM)	< 10 мс/кадр (вычисления в памяти)	3x
Авторегрессионная генерация	150–300 мс (на токен)	< 80 мс (параллельная оптимизация и быстрый доступ к KV)	2–4x
Общая задержка инференса	200–350ms	< 100ms	2–3.5x
Мощность	150W	< 80W	Снижение на 47%
Потребность в DRAM	32 ГБ (ограниченный запас)	8 ГБ (достаточный запас)	Снижение в 4 раза

3.2.3 Техническая осуществимость полной автономности L5

Сценарий: Транспортное средство находится в незнакомой горной местности, в которой оно никогда раньше не бывало, и сталкивается со сценариями с длинным хвостом, которых нет в обучающих данных, такими как временные оползни, переходы диких животных и команды жестами.

Традиционное модульное решение:

• Необходимо заранее написать правила для каждого сценария (например, «Обнаружен уклон → замедлить + объезд»).

• Степень охвата сцен с длинным хвостом составляет < 80%, а автоматическое вождение отключается сразу при обнаружении неизвестных сцен.

План iNc + VLA:

• Модель 70B VLA работает в реальном времени на iNc Auto-L5 (задержка < 100 мс, потребляемая мощность < 80 Вт).

• Модель обладает возможностями «рассуждения на основе здравого смысла» благодаря крупномасштабному предварительному обучению (например, на основе 1 миллиона часов

видеороликов о вождении).

• При столкновении с оползнем модель понимает, что «дорога впереди непроходима» → автоматически планирует маршрут объезда.

• Способность к генерализации повышается в 3–5 раз, а степень охвата сцен с длинным хвостом составляет >95%.

Ценность для бизнеса:

• Попрощайтесь с зависимостью от высокоточных карт: стоимость сбора и обновления высокоточных карт составляет > 1 млрд юаней в год на автомобильную компанию, а модель VLA может снизить затраты на 80%.

• Ускорить реализацию L5: от «необходимости охватить все сценарии» (невозможно) до «позволить модели научиться рассуждать» (возможно).

3.3 Интеллектуальный салон: помощь с учётом контекста

3.3.1 Локализованное развертывание мультимодальных больших моделей

Сценарий: Во время вождения пользователь сказал: «Помогите мне найти ближайший ресторан с детскими развлечениями. Вокруг должно быть тихо. Забронируйте столик на 4 человека на 19:00».

Традиционный план:

• Голос загружается в облако (задержка > 2 секунд) → Cloud GPT-4 понимает намерение → Вызывает API карты → Возвращает результат.

• Проблема: недоступно в туннелях/горных районах без сети; риск утечки конфиденциальной информации в деликатных разговорах (например, деловых переговорах).

План iNc (iNc Auto-L3, 256–512 ГБ UFS):

• Мультимодальная большая модель (VLM) 7B находится в iNc UFS и не требует облачных вызовов.

• Процесс рассуждения:

1. Распознавание речи ASR (задержка < 100 мс)

2. VLM понимает замысел: «ресторан» + «детские площадки» + «тихо» + «4 человека» + «19:00».

3. Поиск векторных данных из локальной библиотеки карт (хранится в iNc): соответствие функциям «удобно для детей и тихо»

4. Чтобы забронировать номер, позвоните в Meituan API.

5. Голосовая трансляция TTS: «Для вас зарезервирован XX ресторан. Он находится в 3 километрах и ожидается через 10 минут».

• Общая задержка: < 3 секунд (незаметно для пользователя)

• Потребляемая мощность: < 2 Вт (по сравнению с облачным решением, требующим поддержания соединения 4G/5G, потребляемая мощность> 5 Вт)

3.3.2 Мониторинг водителя и распознавание эмоций

Сценарий: Автомобильная камера фиксирует, что водитель зевнул три раза подряд + частота моргания снизилась на 30%.

Традиционный план:

• Модель DMS (500M параметров) работает на процессоре SoC NPU и выводит оценку усталости каждые 100 мс.

• Потребляемая мощность: 5–8 Вт (непрерывная работа).

• Частота ложных тревог: 10–15% (без учета индивидуальных различий).

Решение iNc (iNc Auto-L3):

• Модель DMS развернута в iNc, с задержкой вывода < 20 мс и энергопотреблением < 1,5 Вт.

• Персонализированное моделирование: создайте персонализированную модель усталости для каждого пользователя (например, «нормальная частота моргания человека составляет 15 раз в минуту, а если она меньше 10 раз в минуту, он утомлен»).

• Распознавание эмоций: распознавайте такие эмоции, как гнев, тревога, счастье и т. д., с помощью микровыражений лица + интонации голоса и регулируйте атмосферу в кабине (музыка, освещение, ароматы).

• Активное вмешательство: после обнаружения усталости немедленно воспроизводится динамичная музыка + увеличивается температура кондиционера + голосовая подсказка «Рекомендуется отдохнуть в передней зоне обслуживания».

Социальная ценность:

• Усталость вождения является причиной >20% дорожно-транспортных происшествий. DMS на базе iNc может заранее предупредить за 5–10 минут, спасая десятки тысяч жизней каждый год.

3.4 Непрерывное обучение на протяжении срока службы автомобиля

3.4.1 Индивидуальное обучение стилю вождения

Сценарий: Пользователи привыкли к агрессивному вождению (быстро ускоряются, следуют за автомобилем впереди, решительно меняют полосу движения), но стратегия автоматического вождения автомобиля при выходе с завода является консервативной (медленное ускорение, большое расстояние между автомобилями), в результате чего пользователи часто берут на себя управление.

Традиционный план:

• Данные загружаются в облако для обучения (стоимость трафика > 1 юань/ГБ, период обучения > 7 дней).

• Риск конфиденциальности: данные о поведении вождения (например, места частого посещения, стиль вождения) собираются в облаке.

Решение iNc (iNc Auto-L4, поддерживает встроенное обучение FP8/BF16):

• Онлайн-обучение: сцена (изображение камеры, данные радара, действия рулевого колеса) каждый раз, когда пользователь берет на себя управление, автоматически записывается как отрицательный образец.

• Ночное обучение: когда автомобиль заряжается, iNc локально выполняет дополнительное обучение модели стратегии вождения (потребление энергии < 50 Вт, завершение за 3 часа).

• Вступает в силу на следующий день: обновленная модель развертывается автоматически, а скорость внедрения снижается на 80 % (с 10 раз на 100 километров до < 2 раз).

Техническая реализация (рассчитана на основе научных исследований):

• iNc поддерживает расчет градиента обратного распространения сигнала (точность FP8/BF16) с коэффициентом энергоэффективности 15,6 TOPS/Вт.

• Объем обучающих данных: каждое дублирование записывает 10 секунд данных сцены (приблизительно 1 ГБ), а 1000 дублей накапливают 1 ТБ данных.

• Время обучения: данные объемом 1 ТБ используются для точной настройки модели 7B на iNc Auto-L4, что занимает < 3 часов (по сравнению с 1 днем для загрузки в облако + 1 днем для обучения).

Прорывная ценность:

• Тысячи людей, тысячи лиц: стратегия вождения каждого автомобиля уникальна и идеально адаптирована к привычкам пользователя.

• Конфиденциальность данных: данные о вождении не загружаются в облако и соответствуют Закону о защите личной информации.

• Непрерывная эволюция: автомобиль продолжает учиться на протяжении всего своего жизненного цикла (более 10 лет) и становится тем умнее, чем больше он используется.

3.4.2 Региональная адаптация через федеративное обучение

Сценарий: Транспортное средство перебрасывается из Пекина в Тибет для использования в условиях гипоксии на плато, сложных горных дорог, экстремальных погодных условий и других сценариев, которых нет в данных обучения в Пекине.

Традиционный план:

• Автомобильным компаниям необходимо отдельно собирать данные, обучать модели и продвигать OTA для Тибета (цикл > 6 месяцев).

• Скорость поглощения выросла до 20+ раз на 100 километров.

Схема iNc (федеративное обучение):

• Совместное обучение в мастерской: когда по Тибету ездят несколько транспортных средств, iNc локально рассчитывает градиент модели.

• Агрегация градиентов: каждый автомобиль загружает градиенты (< 10 МБ) в облако для агрегирования вместо необработанных данных (> 1 ГБ).

• Обновление модели: оптимизированная глобальная модель создается в облаке и передается на все транспортные средства OTA.

• Временной цикл: сокращен с 6 месяцев до < 1 месяца, скорость передачи снижена до < 2 раз/100 километров.

Социальная ценность:

• Одна модель может адаптироваться к любому региону мира (левостороннее/праворульное управление, различные правила дорожного движения, сложные дорожные условия) без необходимости отдельного обучения для каждого рынка.

• Ускорить внедрение L5: быстро накопить данные сценариев с длинным хвостом посредством совместного обучения глобального парка транспортных средств.

4. Преобразование продукта

4.1 Централизованные вычисления и интеграция памяти с вычислениями

Больные точки традиционной архитектуры:

• Автомобиль оснащен более чем 100 ЭБУ, которые управляют питанием, шасси, кабиной и ADAS соответственно. Они независимы друг от друга и их трудно координировать.

• Длина ремня > 5 километров, вес > 60 кг, стоимость > 5000 юаней.

Новая архитектура на базе iNc:

[Центральная вычислительная платформа (iNc Auto-L4/L5 + облегченный SoC)]

• 1–2 iNc Auto-L4/L5 (для хранения данных + вычислительная мощность искусственного интеллекта)

• 1–2 легких SoC (отвечают за поток управления + принятие решений в реальном времени, 50–100 TOPS)

• Объединение памяти в пул через межсоединение CXL 3.0 (пропускная способность > 1 ТБ/с).

• Количество ЭБУ уменьшено до < 10 (сохраняется только резервирование безопасности и управление исполнительными механизмами)

[Архитектура контроллера домена]

• Интеллектуальная область вождения: iNc Auto-L4 + SoC (отвечает за распознавание, планирование и контроль).

• Домен кабины: iNc Auto-L3 (отвечает за мультимодальное взаимодействие, развлечения, DMS)

• Область питания: легкий микроконтроллер (отвечает за управление двигателем и аккумулятором).

• Домен шасси: легкий MCU (отвечает за торможение и рулевое управление).

Оптимизация производительности и стоимости

Показатель	Традиционная архитектура	Архитектура iNc	Улучшение
Число ЭБУ	100+	< 10	Снижение на 90%
Длина проводки	> 5 км	< 1 км	Снижение на 80%
Мощность системы	> 300W	< 200W	Снижение на 33%
Стоимость компонентов	> 20 000 юаней (кристаллы + ЭБУ + проводка)	< 12 000 юаней	Снижение на 40%
Цикл обновления ПО	> 2 лет (тесная связь с аппаратурой)	< 6 месяцев (программно определяемая система)	Ускорение в 4 раза

Пример: Tesla FSD

• Tesla использует архитектуру центральных вычислений + OTA, чтобы сократить цикл итерации функции с 2–3 лет для традиционных автомобильных компаний до 3–6 месяцев.

• Ожидается, что решение iNc снизит затраты еще на 30% на этой основе (снижение требований к DRAM + high-end SoC).

4.2 От продажи автомобилей к вычислениям как услуге

4.2.1 Автономное вождение по подписке

Традиционный режим:

• Пользователи платят единовременный платеж за стоимость оборудования для автономного вождения (> 20 000 юаней) при покупке автомобиля.

• Автомобильные компании не имеют последующей выручки и имеют низкую рентабельность (< 10%).

Новый режим для драйвера iNc:

• Стандартная аппаратная конфигурация: все модели с завода оснащены iNc Auto-L4 (стоимость < 3500 юаней).

• Подписка на программное обеспечение:

◦ Функция L2+: бесплатно (высокоскоростной NOA, автоматическая парковка)

◦ Функция L3: 199 юаней/месяц (городское NOA, двухточечное автономное вождение)

◦ Функция L4: 399 юаней/месяц (полностью без водителя, ограниченная территория)

• Доход автомобильной компании: доход от подписки на жизненный цикл велосипеда (10 лет) > 20 000 юаней, валовая прибыль > 80%.

Случай:

• Плата за подписку Tesla FSD составляет 199 долларов США в месяц, количество подписчиков в 2025 году составит > 500 000, а годовой доход составит > 1,2 миллиарда долларов США.

4.2.2 Аренда неиспользуемых вычислительных ресурсов

Сценарий: Когда автомобиль заряжается в ночное время (8 часов), вычислительная мощность iNc Auto-L4 (эквивалентная 500–800 TOPS) простаивает.

Бизнес-модель:

• Аренда вычислительной мощности: после разрешения владельца автомобиля вычислительная мощность iNc сдается в аренду платформам облачных вычислений (например, фермам рендеринга и обучению искусственному интеллекту).

• Распределение доходов: владельцы автомобилей получают доход в размере 20–50 юаней за ночь (стоимость электроэнергии < 5 юаней).

• Годовой доход: 7 000–18 000 юаней в год, что может компенсировать часть стоимости покупки автомобиля.

Техническая осуществимость:

• Потребляемая мощность iNc < 60 Вт (потребляемая мощность < 0,5 градуса в течение 8 часов в ночное время, стоимость < 0,3 юаня).

• Подключайтесь к облаку через 5G/Wi-Fi, получайте задачи рендеринга/обучения и загружайте выходные результаты.

4.3 Роботакси и автономная логистика

4.3.1 Автономные такси: сократите транспортные расходы на 70 %

Традиционное такси:

• Стоимость водителя: 150 юаней в день (8 часов).

• Стоимость топлива: 80 юаней/день (200 километров × 0,4 юаня/км).

• Амортизация автомобиля: 50 юаней в день.

• Общая стоимость: 280 юаней в день, стоимость километра 1,4 юаня.

Автономное такси (iNc Auto-L5):

• Стоимость водителя: 0 юаней (беспилотное вождение).

• Плата за электричество: 40 юаней в день (200 километров × 0,2 юаня/км).

• Амортизация транспортного средства: 50 юаней/день (стоимость одного транспортного средства <200 000 юаней, срок службы 10 лет)

• Общая стоимость: 90 юаней в день, стоимость километра 0,45 юаней.

• Плата за пользование: 0,8 юаня/км (по сравнению с традиционным такси — 2 юаня/км).

Коммерческая жизнеспособность:

• Транспортное средство работает 24 часа в сутки (водитель не отдыхает), среднее количество заказов, полученных в день, составляет > 50, а ежедневный доход составляет > 600 юаней.

• Валовая прибыль: > 75%, период безубыточности < 18 месяцев (по сравнению с традиционным такси > 5 лет).

Объем рынка:

• Мировой туристический рынок составляет > 10 триллионов долларов США. Когда проникновение роботакси достигнет 30%, размер рынка превысит 3 триллиона долларов США.

4.3.2 Автономная логистика последней мили

Традиционный экспресс:

• Стоимость работы курьера: 5–8 юаней/заказ.

• Амортизация электрического трехколесного велосипеда + плата за электроэнергию: 1 юань за заказ.

• Общая стоимость: 6–9 юаней/заказ.

Беспилотный автомобиль доставки (iNc Auto-L4):

• Стоимость рабочей силы: 0 юаней.

• Плата за электроэнергию + амортизация: <1 юаня/заказ (стоимость одного автомобиля <100 000 юаней, доставка 200 заказов в день).

• Общая стоимость: < 1 юань/заказ.

Объем рынка:

• Ежедневные заказы на экспресс-доставку в Китае составляют >300 миллионов. Когда уровень проникновения беспилотной доставки достигнет 50%, годовая экономия средств составит> 200 миллиардов юаней.

5. Влияние на отрасль и общественная ценность

5.1 Преобразование автомобильной производственной цепочки

Этап	Традиционное решение	Решение iNc	Влияние
Датчик	Лидар: 50 000–100 000 юаней за штуку × 1–2	Сокращение до 0–1 (дополнение V2X)	Снижение стоимости на 50–80%
Вычислительный кристалл	Два кристалла Orin-X > 5000 юаней	Один iNc Auto-L4 < 3000 юаней	Снижение стоимости на 40–60%
Хранилище	LPDDR5 16 ГБ + UFS 256 ГБ > 1000 юаней	iNc UFS 512 ГБ < 800 юаней (интегрированные вычисления)	Снижение стоимости на 20–30%
Разработка ПО	Модульное решение, цикл > 2 лет	Сквозная модель, цикл < 1 года	Снижение затрат на разработку на 50%

5.2 Общественная ценность

1. Безопасность дорожного движения:

• Человеческие ошибки являются причиной 94% причин несчастных случаев, а система L4/L5, управляемая iNc, снижает уровень несчастных случаев на 90%.

• Спасенных жизней в год: сотни тысяч (смертность в результате дорожно-транспортных происшествий во всем мире >1,3 миллиона в год).

2. Устранение пробок:

• Интеллектуальная диспетчеризация + сотрудничество между автомобилем и дорогой повышает эффективность городского движения на 30–50 % и сокращает время в пути на 20–30 %.

3. Энергосбережение и сокращение выбросов:

• Электромобили + оптимизация автономного вождения (плавное ускорение, координация автопарка) снижают потребление энергии на 20–30%.

• Сокращение выбросов углекислого газа: миллиард тонн в год (рассчитано на основе 1 миллиарда транспортных средств по всему миру).

4. Трансформация занятости:

• Сокращение рабочих мест: водители (>100 миллионов во всем мире)

• Новые должности: тренер по искусственному интеллекту, удаленный сотрудник по безопасности, управление автопарком, диспетчеризация вычислительных мощностей (оценка > 50 миллионов).

6. Дорожная карта и внедрение технологии

6.1 Развитие технологии

Этап	Время	Техническая цель	Область применения
Первое поколение	2026–2027	iNc Auto-L3: эквивалент 100–300 TOPS для поиска	Магистральная NOA, интеллектуальный салон, DMS
Второе поколение	2027–2028	iNc Auto-L4: эквивалент 500–800 TOPS, поддержка VLA 35 млрд	Городская NOA и опытная эксплуатация роботакси
Третье поколение	2028–2030	iNc Auto-L5: эквивалент 1000+ TOPS, модели 100+ млрд и обучение на кристалле	Полная автономность и масштабные сети роботакси

6.2 Состояние на январь 2026 года

• Техническая проверка: Прототип iNc Auto-L3 прошел стендовые испытания, и производительность векторного поиска достигла стандарта (эквивалент 200 TOPS, потребляемая мощность < 30 Вт).

• Миграция приложений: Университет Чунцина осуществляет миграцию основных алгоритмов автономного вождения (BEVFormer, UniAD, PointPillars) на архитектуру iNc.

• Сотрудничество с автомобильными компаниями: подписаны соглашения о технической оценке с двумя ведущими компаниями, производящими автомобили на новых источниках энергии, и планируется завершить проверку установки автомобилей в первом полугодии 2027 года.

• Эталонный случай: определенная отечественная модель была развернута на беспилотном транспортном средстве для проверки возможности интегрированной архитектуры хранения и вычислений.

7. Заключение: переосмысление автомобиля

Заключение: переосмысление автомобиля

Технология вычислений в памяти Lingcun iNc фундаментально решает три основные дилеммы автономного вождения, устраняя узкие места при передаче данных, улучшая использование вычислительной мощности и снижая энергопотребление системы:

• От «инфляции вычислительной мощности» к «эффективности вычислительной мощности»: увеличьте эффективное использование вычислительной мощности с 20–40 % до 70–90 % и снизьте энергопотребление на 30–50 % при той же производительности.

• От «облачной зависимости» к «автомобильному интеллекту»: большая модель VLA 35–70B работает на автомобиле в реальном времени (задержка < 100 мс, потребляемая мощность < 80 Вт), обеспечивая настоящее автономное вождение уровня L5.

• От «закалки на заводе» к «эволюции на протяжении всей жизни»: благодаря непрерывному обучению на стороне транспортного средства + интегрированному обучению автомобиль постоянно оптимизируется на протяжении всего своего жизненного цикла и становится умнее, чем больше он используется.

В 2026 году, «первом году автономного вождения», Lingcun завершила разработку основных технологий и тесно сотрудничает с Университетом Чунцина и ведущими автомобильными компаниями. Мы искренне приглашаем партнеров отрасли принять участие в этой столетней трансформации автомобильной промышленности, используя технологию вычислений в памяти, чтобы сделать безопасные, эффективные и интеллектуальные путешествия полезными для всех, а также воплотить в жизнь идею «привнесения общего интеллекта в физический мир».

Получите залог и переопределите автомобиль, начиная с расчета залога.