Network Innovation and Development Alliance
全球固定网络创新联盟

中国电信
CHINA TELECOM

ИНТЕРНЕТ АССОЦИАЦИЯ КАЗАХСТАНА
INTERNET ASSOCIATION OF KAZAKHSTAN
ҚАЗАҚСТАННЫҢ ИНТЕРНЕТ АССОЦИАЦИЯСЫ

Городская сеть для эпохи ИИ

Альянс Сетевых Инноваций и Развития

Сентябрь 2025

Предисловие

Быстрое развитие индустрии искусственного интеллекта приводит к взрывному росту различных приложений искусственного интеллекта. Городские вычислительные сети (Metropolitan Area Networks - MANs), как критически важная инфраструктура, соединяющая конечных пользователей и вычислительные ресурсы, в настоящее время сталкивается с трансформирующими требованиями к сетевой архитектуре, функциональным возможностям и парадигмам обслуживания.

В 2024 году China Telecom впервые в отрасли представила концепцию «вычислительной сервис-ориентированной городской сети» и выпустила аналитический технический документ (white paper) о вычислительной сервис-ориентированной городской сети, привлекшей внимание аудитории и обсуждение в масштабах всей отрасли. В продолжении концепции «белая книга» представляет углубленный анализ эволюции сети мегаполисов в эпоху искусственного интеллекта. В этой книге впервые анализируется ландшафт развития искусственного интеллекта с точки зрения развития отрасли и макроэкономической политики. Далее проводится углубленный анализ требований к приложениям искусственного интеллекта, чтобы определить основные сетевые возможности, которыми должны обладать городские сети. Затем в настоящем документе рассматриваются цели проектирования, подробно описываются общая архитектура и ключевые технологии городских сетей в эпоху искусственного интеллекта. В заключение предлагаются технические решения, адаптированные к типичным сценариям.

Организации и основные участники, которые внесли свой вклад в подготовку этого настоящего документа:

- Китайский научно-исследовательский институт телекоммуникаций в лице Юнцин Чжу, Цзехуа Ху, Ся Гун и Шичжан Юань.
- Альянс инноваций индустрии новой инфраструктуры в лице Zhongguancun Ultra Cross Connection в лице Бо Юань.
- Huawei Technologies Co. Ltd. в лице Хаобинь Чжао, Цзе Донг и Ли Чжан.
- Корпорация ZTE в лице Вэньцян Тао, Хайдун Чжу и Сяовэй Цзи.

СОДЕРЖАНИЕ

Глава I Тренды развития искусственного интеллекта

1.1 Индустрия искусственного интеллекта вступает в фазу ускоренного роста

Основная движущая сила четвертой промышленной революции индустрия искусственного интеллекта (далее - ИИ) переживает беспрецедентно быстрое развитие, демонстрируя огромный рыночный потенциал. По данным исследовательской компании Grand View Research объем мирового рынка искусственного интеллекта достиг 196,63 млрд в 2023 году и по прогнозам увеличится до 1811,75 млрд к 2030 году при совокупном годовом темпе роста (CAGR) в 37,3% с 2024г. по 2030г. Отчеты об исследованиях в Китае показывают, что масштаб индустрии искусственного интеллекта, как ожидается, увеличится с 398,5 млрд юаней в 2025 году до 1729,5 млрд юаней в 2035 году, при предполагаемом CAGR в 15,6%. Искусственный интеллект, несомненно, стал мощным двигателем глобального экономического роста.

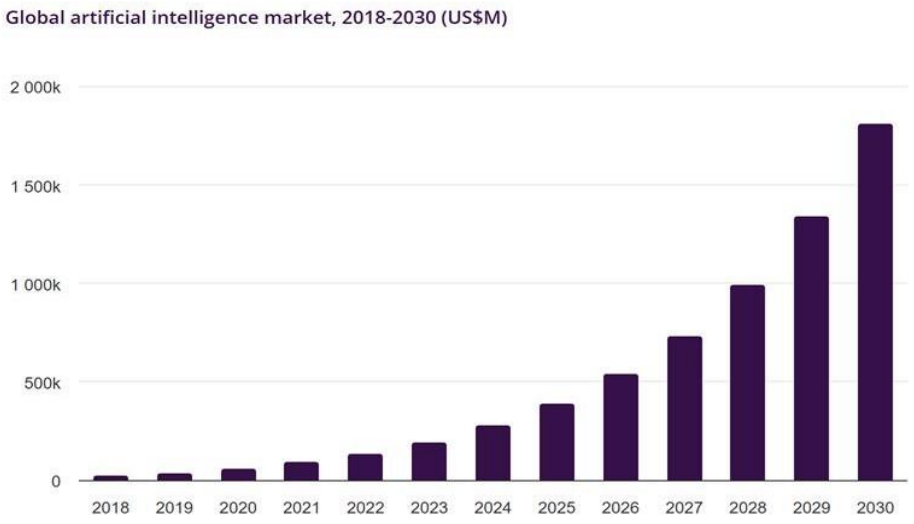

Рисунок 1-1: Глобальный рынок искусственного интеллекта

Мировая индустрия искусственного интеллекта демонстрирует тенденции к развитию «поступательного развития по двум направлениям и разнообразное, многогранное процветание». С одной стороны, глобальные технологические гиганты продолжают увеличивать свои инвестиции в ИИ, такие компании, как Google и Microsoft, углубляют исследования и разработки (R&D) в области основных технологий ИИ; Amazon и Apple продолжают внедрять инновации в

области интеллектуальных облачных сервисов и приложений для конечных устройств, а крупные китайские технологические компании, такие как Baidu, Alibaba, Tencent и Huawei (BATH), также добиваются быстрых успехов в таких ключевых областях, как разработка чипов для ИИ, создание крупных моделей ИИ, компьютерное зрение и воплощенный интеллект. С другой стороны, взрывные прорывы в технологии генеративного искусственного интеллекта вызвали волну инновационных предприятий по всему миру. OpenAI стала пионером в коммерциализации генеративного искусственного интеллекта с помощью ChatGPT, Anthropic и Cohere специализируются на разработке, ориентированной на вертикаль, а в 2025 году китайская DeepSeek значительно ускорила коммерческое применение больших моделей искусственного интеллекта в сценариях логического вывода. Многочисленные развивающиеся компании по цепочке поставок искусственного интеллекта стали инвестиционными центрами, сотрудничающими с лидерами отрасли для формирования синергетической инновационной экосистемы. Эта динамичная модель развития, характеризующаяся конкуренцией и симбиозом между различными игроками, не только ускоряет коммерческое внедрение масштабных языковых моделей в финансах, здравоохранении и производстве, но и обеспечивает надежный импульс для качественного развития цифровой экономики.

Благодаря быстрому развитию индустрии искусственного интеллекта технологии искусственного интеллекта становятся мощными двигателями городского развития, придавая беспрецедентную жизнеспособность различным отраслям городов. В транспортной сфере использование возможностей точного прогнозирования больших моделей искусственного интеллекта оптимизирует транспортные потоки и повышает эффективность поездок. В сфере здравоохранения диагностические технологии с использованием искусственного интеллекта позволяют быстро и точно анализировать медицинские изображения, помогая врачам составлять планы лечения. В сфере образования предоставляется индивидуальный учебный контент, основанный на прогрессе в обучении и характеристиках учащихся, стимулирующий их интерес и потенциал.

Финансовый сектор использует большие модели искусственного интеллекта для оценки рисков и принятия инвестиционных решений, повышая точность и безопасность финансовых услуг. Кроме того, многочисленные области, такие как интеллектуальное производство, интеллектуальные государственные услуги и мониторинг окружающей среды, стали более эффективными, интеллектуальными и устойчивыми благодаря расширению возможностей искусственного интеллекта. Применение технологий искусственного интеллекта обеспечивает жителям более удобную и безопасную жизнь, ведя города к интеллектуальному и цифровому будущему.

1.2 Искусственный интеллект является центром глобальной промышленной политики

Искусственный интеллект стал одной из основных движущих сил городского и социального развития, формируя глобальный консенсус:

- **Соединенные Штаты** запустили "Инициативу Белого дома по созданию умных городов" в 2015 году, использование технологий искусственного интеллекта, больших данных и Интернета вещей (IoT) для оказания помощи городам в решении таких проблем, как пробки на дорогах, управление энергопотреблением и общественная безопасность. К 2025 году это еще больше укрепит инфраструктуру искусственного интеллекта с помощью проекта «The Stargate Project».
- **Европейский союз** предложил «Стратегию Европейского Союза Данных» на 2025 год для продвижения приложений искусственного интеллекта и больших данных в здравоохранении, образовании и городском управлении при поддержке «Программы цифровой Европы» по внедрению искусственного интеллекта в важнейших социальных секторах и сферах жизнеобспечения.
- **Япония** представила концепцию «Супер-Город», объединяющую искусственный интеллект и Интернет вещей для создания «умных городов», основанных на данных.

- **Сингапур** реализовал свою «Национальную стратегию ИИ 2.0», которая сочетает привлечение талантов, промышленное применение, научно-исследовательские инновации и инфраструктуру для создания экосистемы ИИ, улучшающей общественные услуги и конкурентоспособность промышленности.

- **Китайское правительство** также уделяет приоритетное внимание развитию городов на основе искусственного интеллекта. В 2024 году Национальное управление данных Китая выпустило руководящие принципы по углублению инициатив «умного города», поощряя решения на базе искусственного интеллекта, такие как интеллектуальный анализ, планирование, регулирование и принятие решений с целью всестороннего содействия цифровой трансформации городов.

  Сети стали важнейшей инфраструктурой, поддерживающей развитие глобальной индустрии искусственного интеллекта, и им уделяется первостепенное внимание со стороны стран по всему миру.

- В **Китае** «расширение возможностей вычислений с помощью сетей» было создано в качестве фундаментальный принцип построения умных городов. В октябре 2023 года Министерство промышленности и информационных технологий Китая (MIIT) представило Высококачественный план действий по развитию вычислительной инфраструктуры, целью которого является создание группы эталонных показателей вычислительной мощности городских сетей в ключевых регионах.

- В июне 2023 года правительство **Сингапура** предоставило свой план развития цифровой связанности, в котором предлагается в течение пяти лет построить бесперебойную внутреннюю связь со скоростью 10 Гбит/с, чтобы обеспечить мировой уровень цифровой инфраструктуры Сингапура и определить направление его цифрового будущего.

- В апреле 2024 года Министерство связи **Саудовской Аравии**

опубликовало концепцию «Общества Саудовской Аравии со скоростью 10 Гбит/с», став первыми в мире, кто предложил сквозную высокоскоростную высококачественную сетевую архитектуру Net5.5G для поддержки интеллектуальной трансформации страны.

– В 2025 году в программе **Европейской комиссии** «Цифровая Европа» (DIGITAL) на 2025-2027 годы также была подчеркнута необходимость повышения отказоустойчивости сетей в различных сценариях искусственного интеллекта.

– В **Казахстане** правительство активно продвигает цифровую трансформацию и развитие искусственного интеллекта (ИИ).

– В феврале 2024 года Министерство цифрового развития **Казахстана** опубликовало документ «Концепция развития искусственного интеллекта на 2024-2029 годы». В этом документе планируется создать Национальный центр искусственного интеллекта, интегрировать ресурсы в различных областях и содействовать широкому применению искусственного интеллекта в экономическом и социальном секторах. Для подготовки специалистов в области ИИ Казахстан реформировал свою систему образования. С 2026 года инструменты искусственного интеллекта будут включены в учебные программы начальных и средних школ, многие университеты запустили образовательные программы, связанные с искусственным интеллектом, и в настоящее время исследованиями в области искусственного интеллекта занимаются более 2000 аспирантов.

  Казахстан учредил венчурный фонд в размере 1 миллиарда долларов в Международном финансовом центре Астаны для поддержки перспективных стартапов в области искусственного интеллекта. В июле 2025 года с вводом в эксплуатацию первого суперкомпьютера, вычислительные мощности страны были значительно увеличены, что заложило прочную основу для развития искусственного интеллекта.

  В сентябре 2025 года президент РК Токаев в своем ежегодном

Послании к народу объявил, что о создании Министерство искусственного интеллекта и цифрового развития с целью создания всеобъемлющей экосистемы цифровых активов. Одновременно с этим он предложил создать Государственный Фонд цифровых активов для накопления стратегического крипторезерва из наиболее перспективных активов нового цифрового финансового уклада. Эти меры призваны всесторонне стимулировать цифровое развитие страны с целью превращения Казахстана в полностью цифровую страну в течение трех лет.

С широким распространением крупных моделей ИИ и растущим спросом на такие приложения, как распределенные вычисления, роль сетей в развитии искусственного интеллекта становится все более заметной. Строительство второй «информационной супермагистрали», посвященной искусственному интеллекту, стало глобальным приоритетом.

1.3 Технология искусственного интеллекта развивается взрывными темпами

1.3.1 Технология искусственного интеллекта всесторонне развивается

Развитие технологий искусственного интеллекта демонстрирует заметные тенденции диверсификации сотрудничества, высокоэффективной эволюции и интеграции нескольких экосистем:

- На уровне аппаратного обеспечения значительное увеличение числа сценариев логического вывода привело к быстрому прогрессу в области специализированных чипов искусственного интеллекта, таких как TPU и LPU, в то время как универсальные графические процессоры (GPU) в сочетании с передовыми технологиями, такими как чиплет, 3D-стекинг и квантовые вычисления, предоставляют расширенные возможности для обучения сверхмасштабных моделей искусственного интеллекта.
- В области технологий хранения данных протоколы HBM3 и CXL

обеспечили значительный рост пропускной способности и емкости памяти, в то время как такие архитектуры, как разделение хранилища и вычислений, удовлетворяют спрос на создание частных баз знаний на основе больших моделей искусственного интеллекта.

- Высокоскоростные технологии интерконнекта, такие как UEC, NVLink, UCIe и Falcon, устраняют барьеры в передаче данных, обеспечивая эффективное взаимодействие между распределенными вычислениями и гетерогенными архитектурами.
- В экосистеме программного обеспечения открытые фреймворки, такие как PyTorch и TensorFlow, глубоко интегрируются с автоматизированными цепочками инструментов в сочетании с унифицированным развертыванием облачных и периферийных устройств, что позволяет сквозной оптимизации от обучения до вывода.
- Кроме того, экологически чистые вычислительные технологии, включая жидкостное охлаждение и динамическое управление энергопотреблением, способствуют устойчивому развитию искусственного интеллекта.

1.3.2 Технология больших моделей искусственного интеллекта вступает в фазу быстрого развития

Большие модели искусственного интеллекта стали одной из наиболее широко применяемых ключевых технологий искусственного интеллекта на сегодняшний день. С момента запуска ChatGPT в 2022 году и до появления DeepSeek в 2025 году область больших моделей искусственного интеллекта пережила взрывной рост. Разработка больших моделей демонстрирует многомерные тенденции. С одной стороны, масштаб модели продолжает расширяться с увеличением количества параметров, что позволяет улавливать более сложные закономерности и взаимосвязи для повышения производительности в различных задачах. С другой стороны, мультимодальное слияние стало важным направлением развития, поскольку большие модели

объединяют текст, изображения, речь и другие мультимодальные данные для достижения более полного понимания и генерации информации, расширяя сценарии их применения. Кроме того, большее внимание уделяется безопасности, надежности и интерпретируемости моделей, и исследователи стремятся разрабатывать более надежные архитектуры моделей и методы обучения, чтобы обеспечить стабильную работу и надежное применение больших моделей искусственного интеллекта в сложных средах. Эти тенденции в совокупности стимулируют непрерывное развитие технологии больших моделей искусственного интеллекта, закладывая прочную основу для широкого применения искусственного интеллекта. В настоящее время большие модели искусственного интеллекта развиваются в следующих технических направлениях:

Направление 1: По мере того, как параметры и объем обучающих данных больших моделей искусственного интеллекта продолжают расти, быстро растет потребность в вычислительных мощностях. Одиночные центры обработки данных искусственного интеллекта (AIDC) с 1K+ или 10K+ графическими процессорами едва ли могут удовлетворить требования сверхмасштабного обучения. Например, модель Llama 3.1, выпущенная в 2024 году, имеет 405 миллиардов параметров и требует примерно 15 триллионов токенов для предварительного обучения, а весь процесс обучения требует 39,3 миллиона GPU/часов (H100) вычислительной мощности. **Поэтому внедрение методов распределенного обучения и использование высокопроизводительных сетей для повышения эффективности совместного обучения в нескольких AIDC стало необходимостью для развития искусственного интеллекта.** В настоящее время несколько операторов завершили коммерческое внедрение распределенного обучения, достигнув распределенного обучения для более чем 10 тысяч GPU, 100 миллиардов параметров больших моделей ИИ в AIDC на расстояниях более 100 километров. Среди них China Telecom и Huawei совместно развернули службу распределенного обучения, поддерживающую передачу данных RDMA на расстоянии 120 км без потерь, при этом эффективность обучения достигает более 95%.

Направление 2: **Оптимизация программного обеспечения стала ключевым способом преодоления узких мест в аппаратном обеспечении ИИ, обеспечения рентабельности разработки больших моделей искусственного интеллекта и ускорения внедрения искусственного интеллекта во всех отраслях.** DeepSeek-V3 с открытым исходным кодом в 2025 году завершил предварительное обучение всего за два месяца, используя всего 2048 графических процессоров за счет алгоритмической оптимизации, а модель DeepSeek-R1 еще больше сократила цикл обучения до 2-3 недель. Это «недорогое и открытое» решение значительно снизило технический порог для больших моделей ИИ, что непосредственно привело к двум заметным изменениям. Во-первых, относительно низкие затраты на использование вызвали взрывной рост приложений на основе больших моделей искусственного интеллекта, что привело к увеличению трафика искусственного интеллекта в городах, что требует от сети обеспечения эффективного управления трафиком ИИ. Во-вторых, благодаря оптимизации полного стека программного обеспечения, охватывающего «алгоритм, аппаратное обеспечение и систему», задержка вывода ИИ была сокращена более чем на 60%, что привело к экспоненциальному росту спроса на вывод искусственного интеллекта.

Направление 3: **Интеллектуальное взаимодействие многосторонних агентов отражает трансформацию технологий искусственного интеллекта от централизованных к распределенным системам и от индивидуального к коллективному интеллекту, что способствует прорывам в производительности ИИ в реальном времени, автономности и совместной работе.** Большие модели ИИ могут быть легко развернуты с помощью таких технологий, как дистилляции моделей, что делает их совместимыми с ограниченными ресурсами, такими как графические процессоры потребительского уровня, мобильные устройства и оборудование Интернета вещей, тем самым способствуя разработке небольших интеллектуальных устройств на базе edge. На уровне программного обеспечения широкое внедрение технологии Multi-Agent

позволяет нескольким терминалам совместно выполнять сложные задачи, что способствует дальнейшему развитию крупномасштабных интерактивных приложений периферийных агентов. Внедрение протоколов Google A2A и MCP для взаимодействия агентов в 2025 году сигнализирует о предстоящем переходе ИИ от архитектуры «облачных вычислений» B2B, B2C и C2C к архитектуре «гранулированных вычислений» A2A, M2M и X2X, при этом частые взаимодействия между интеллектуальными вычислительными частицами предъявляют более высокие требования к надежности и пропускной способности сети.

Направление 4: В сентябре 2024 года OpenAI запустила модель o1 с механизмом Chain-of-Thought (CoT–цепочки размышлений), которая обеспечивает более высокую точность за счет увеличения времени на размышление во время вывода, что знаменует собой смену парадигмы от стремления к скорости ответа к акцентированию вниманию на глубоком рассуждении (Deep Thinking). Эта трансформация привела к сдвигу спроса на вычислительную мощность с предварительного обучения к инференцию (логическому выводу), преодолев ограничения закона масштабирования, который гласит **в то время как предварительное обучение опирается на кластеры масштабирования с более чем 10 тысяч графических процессоров, инференция (логический вывод) может быть реализована с помощью масштабируемых архитектур, состоящих из небольшого числа графических процессоров, что способствует эволюции инфраструктуры ИИ в сторону распределенных и гибко планируемых систем.** Кроме того, значительно возросшие требования к развертыванию системы вывода с использованием искусственного интеллекта повысили требования к улучшениям производительности крупномасштабного вывода. Сетевой распределенный вывод стал ключевым направлением для будущих городских приложений искусственного интеллекта, что требует наличия сетей для поддержки развертывания, распределенного ИИ-вывода. В ответ на это NVIDIA представила платформу (фреймворк) Dynamo Framework, использующую архитектуру с разделением PD для оптимизации планирования ресурсов и эффективности

вычислений при крупномасштабном выводе с помощью искусственного интеллекта.

1.4 Вызовы для городской сети в связи с масштабной коммерциализацией ИИ

Создание комплексной городской экосистемы ИИ стало основным путем модернизации городских систем до уровня передовой интеллектуальной технологии. В этом процессе концепция «город как компьютер» постепенно получила глобальное признание, благодаря глубокой интеграции вычислительных мощностей, хранилищ и терминалов через городские сети (MAN) города превращаются в распределенные сверхмасштабные вычислительные системы, что позволяет осуществлять интеллектуальное управление городом на основе потока данных с точностью до миллисекунд и принятия решений в режиме реального времени. Существующие широкополосные сети, мобильные сети, выделенные правительственные и корпоративные сети, а также облачные сети в городах соединяют различных пользователей. Однако традиционные городские сети с трудом удовлетворяют требованиям по предоставлению городских услуг ИИ, будь то с точки зрения сетевой архитектуры или основных технологий.

1.4.1 Проблемы в области циркуляции

Обучение больших моделей искусственного интеллекта и построение баз знаний обычно требуют объемы данных в масштабе ТБ/ПБ, что предъявляет более высокие требования к пропускной способности сетей передачи данных. Одновременно с этим вычислительный трафик больших моделей демонстрирует значительные эластичные характеристики, требует чрезвычайно высокой надежности сети. Некачественные и недетерминированные сети могут привести к недостаточной пропускной способности передачи данных, чрезмерной задержке или частой потере пакетов, тем самым ставя под угрозу доступность вычислительных ресурсов. Кроме того, итерации версий больших моделей и обновления баз знаний в системах искусственного интеллекта также зависят от стабильной сетевой поддержки. Низкое качество сети может ограничить

реализацию этих функций, что в конечном итоге снизит общую операционную эффективность инфраструктуры искусственного интеллекта.

Быстрое развитие крупномасштабных приложений для вывода и вычислительных парадигм A2A поставило новые задачи перед циркуляцией данных ИИ в городах. С одной стороны, городские сети должны удовлетворять требованиям эффективной передачи данных и взаимодействия между распределенными узлами вывода; с другой стороны, режим A2A привел к экспоненциальному росту высокочастотного трафика взаимодействия между интеллектуальными агентами, что не только значительно увеличивает требования к пропускной способности пограничных сетей, но и требует от городской сети обеспечения надежности информационного взаимодействия между интеллектуальными агентами. Таким образом, для реализации концепции «город как компьютер» необходимо срочно построить новую сверхсвязанную сеть, отличную от традиционных городских сетей, чтобы удовлетворить требования к передаче потоков данных искусственного интеллекта и обеспечить эффективную поддержку эффективной циркуляции вычислительных данных со стороны городских сетей.

1.4.2 Проблемы в области эксплуатации и технического обслуживания

Когда городские сети предоставляют услуги ИИ, управление сетью и техническое обслуживание (O&M) сталкиваются с более серьезными проблемами. С точки зрения модели обслуживания, ИИ изменил структуру сетевого трафика: обучение больших моделей ИИ может вызывать внезапные всплески трафика, а частые взаимодействия между интеллектуальными агентами также генерируют всплески коммуникации, что требует от сетей наличия возможностей прогнозирования и технического обслуживания. Услуги ИИ также требуют более высокой надежности сети, поскольку даже незначительные сбои во время обучения модели могут привести к полной перезагрузке задачи. Когда крупномасштабные услуги инференции заменяют ручные услуги в городах, сети должны обеспечивать качество обслуживания.

Следовательно, традиционные модели управления, которые полагаются на ручное вмешательство и конвергенцию маршрутов для обеспечения базовой доступности сети, больше не могут удовлетворить требованиям к производительности услуг ИИ. Услуги ИИ требуют более высокой скорости самовосстановления после сбоев и более низкой задержки при принятии решений по оптимизации сети, что подталкивает сетевые операции к высокой автономности для удовлетворения таких потребностей, как прогнозируемое обслуживание, информированность об услугах и эластичная оптимизация. Вопрос о том, как оснастить сети высокоинтеллектуальными возможностями управления и эксплуатации, а именно автоматизировать настройку сетевых ресурсов и конфигураций на основе намерений и состояний вычислительных услуг, стал ключевым приоритетом для городской сети, ориентированной на ИИ.

1.4.3 Проблемы безопасности и надежности

С быстрым внедрением крупных моделей огромные объемы городских данных используются для анализа, вычислений и обработки. Данные от предприятий, домохозяйств и частных лиц составляют трафик частной сферы, что создает значительные риски для безопасности для домохозяйств и частных лиц трафик частной сферы включает конфиденциальные данные, такие как личная информация и данные о потреблении, утечка которых может привести к нарушению конфиденциальности. Для предприятий трафик частной сферы включает данные о научных исследованиях и разработках, производственные данные и операционные данные, нарушение безопасности которых может подорвать конкурентоспособность или даже вызвать юридические споры. Поскольку передача данных подвержена потенциальным угрозам, таким как кража, подделка и потеря, городские сети должны обладать надежными средствами защиты данных для обеспечения их конфиденциальности, целостности и доступности.

Традиционные системы AAA Аутентификация-Авторизация-Учет (Authentication, Authorization and Accounting) и технологии шифрования данных, основанные на потоках трафика, с трудом удовлетворяют требованиям

безопасности и доверия, предъявляемым к сценариям ИИ. Однако новые технологии, такие как блокчейн и квантовое шифрование, предлагают инновационные решения для надежной циркуляции данных. Блокчейн обеспечивает неизменяемые, отслеживаемые от начала до конца механизмы доверия для потоков данных ИИ через распределенные реестры и смарт-контракты; квантовое шифрование использует такие прорывные технологии, как квантовая распределение ключей, чтобы существенно улучшить возможности защиты от прослушивания при передаче данных. Городские сети должны интегрировать эти инновационные механизмы, чтобы создать надежную основу для крупномасштабного внедрения ИИ в городах, обеспечивая критически важную инфраструктурную поддержку для широкого внедрения городских ИИ-сервисов.

Глава II Требования к городским сетям, основанным на ИИ

2.1 Инновации в ИИ-приложениях продолжают ускоряться

В начале 2025 года DeepSeek возглавила волну преобразований в генеративном ИИ, движимая своей исключительной производительностью и ведущей в отрасли экономичностью в области обучения LLM и логического вывода, что ускорило коммерциализацию технологий искусственного интеллекта. Сегодня приложения искусственного интеллекта вступили в стадию масштабируемого развертывания, обслуживая различные сценарии в домашних условиях (toH), для потребителей (toC) и бизнеса (toB), с проникновением во множество вертикальных отраслей, включая СМИ, юридические услуги, образование и производство.

2.1.1 Сценарии AItoH (ИИ для дома)

ИИ значительно повышает профессионализм, интерактивность и персонализацию домашних услуг, обогащая домашние сценарии. В настоящее время в отрасли постепенно достигается консенсус относительно построения интегрированной экосистемы умного дома, сочетающей связь, вычислительную мощность и интеллект. Через взаимодействие «облако-сеть-периферия-устройство» (cloud-network-edge-device), предоставляя пользователям широкополосного доступа интеллектуальные облачные сервисы, поддерживаются различные сценарии AItoH, включая умный дом и домашние помощники:

– Умный дом включает в себя умные телевизоры, умные холодильники и другие продукты для умного дома, которые используют технологии искусственного интеллекта, такие как распознавание голоса и компьютерное зрение. Теперь устройства поддерживают интеллектуальные возможности, включая взаимодействие на естественном языке, изучение привычек пользователя и адаптацию к контексту. Эти продукты для умного дома могут динамически регулировать освещение, температуру и влажность в зависимости от предпочтений пользователя, используя при этом технологии

распознавания лиц и анализа поведения для повышения безопасности дома.

– Помощники по дому включают в себя продукты умных домашних помощников (smart home assistant), умные колонки и домашних роботов, которые используют обработку естественного языка и другие технологии искусственного интеллекта для обеспечения гармоничного диалога человека и машины. Эти продукты обеспечивают точное понимание намерений при выполнении задач, включая напоминания по расписанию и поиск информации, одновременно предоставляя контекстуализированные сервисы, такие как управление устройствами и мониторинг безопасности, благодаря бесшовной совместимости IoT.

2.1.2 Сценарии AItoC (ИИ для потребителя)

Искусственный интеллект революционизирует взаимодействие между потребителем и сервисом, улучшая пользовательский опыт и стимулируя рыночные инновации. Инновационный ландшафт искусственного интеллекта является свидетелем быстрого распространения различных вертикальных приложений. Крупные игроки отрасли активно внедряют решения AItoC в интеллектуальных терминалах, персонализированных сервисах и сферах цифрового образа жизни, используя столичные сервисы искусственного интеллекта для улучшения пользовательского опыта и удержания клиентов. Текущие приложения AItoC в основном охватывают следующие категории:

– Повышение производительности (productivity enhancement) обеспечивается за счет приложений искусственного интеллекта, таких как интеллектуальный поиск, автоматическое обобщение, генерация контента и помощь в написании кода, значительно повысили эффективность как для частных лиц, так и для организаций. Эти приложения оптимизируют сложные рабочие процессы, позволяя пользователям сосредоточиться на более ценных стратегических инициативах, одновременно способствуя инновациям и конкурентным преимуществам.

– Креативная генерация (Creative Generation) обеспечивается за счет приложений искусственного интеллекта, включая автоматизацию дизайна, генерацию изображений, синтез видео и сочинение музыки, которые революционизируют индустрию создания контента. Эти приложения дополняют творческие идеи создателей контента.

– Развлечения (Entertainment) обеспечиваются за счет приложений с искусственным интеллектом, такие как ИИ-камеры и виртуальные компаньоны, которые трансформируют пользовательский опыт благодаря новым парадигмам взаимодействия и повышению вовлеченности. Эти приложения используют профилирование пользователей для предоставления персонализированных развлекательных услуг, повышая удовольствие от использования цифровых технологий.

2.1.3 Сценарии AItoB (ИИ для бизнеса)

Искусственный интеллект демонстрирует огромные возможности в области анализа данных и поддержки принятия решений, позволяя предприятиям добиваться значительного повышения операционной эффективности и существенного снижения затрат. Кроме того, искусственный интеллект обладает исключительными возможностями в обработке данных и генерации контента, позволяя предприятиям получить доступ к новым возможностям для бизнеса. Технологическая конвергенция искусственного интеллекта, 5G и передовых вычислений ускоряет интеллектуальную трансформацию промышленности, создавая ценность замкнутого цикла система «высокоскоростного подключения+вычисления в реальном времени+интеллектуальное принятие решений», которая меняет целые процессы от производства до технического обслуживания:

– Ускорение разработки продукта (Accelerating Product Development) на этапе анализа требований ИИ использует обработку естественного языка и анализ тональности для быстрого анализа массивных

пользовательских отзывов и рыночных данных, позволяя точно идентифицировать скрытые потребности и болевые точки. На этапе концептуального проектирования ИИ автоматически создает сотни жизнеспособных решений на основе исторических данных и технических характеристик для оценки инженерами, значительно сокращая циклы проектирования. Для инженерной валидации системы моделирования на основе ИИ с учетом физических законов (physics-informed AI) точно предсказывают параметры производительности продукта, существенно снижая затраты на верификацию.

– Повышение операционной эффективности (enhancing operational efficiency) ИИ позволяет предприятиям достигать интеллектуальных и высокоэффективных операций через автоматизацию процессов, оптимальное распределение ресурсов и усиление управления цепочками поставок. Например, системы мониторинга на основе ИИ проводят мониторинг узлов цепочки поставок в реальном времени, прогнозируя потенциальные сбои и колебания спроса для динамической оптимизации уровней запасов и планирования логистики. Кроме того, системы технического обслуживания на основе ИИ анализируют данные датчиков и исторические записи о техническом обслуживании, чтобы точно прогнозировать модели отказов, позволяя осуществлять упреждающее планирование технического обслуживания, что значительно сокращает незапланированные простои.

2.2 ИИ-приложения демонстрируют различные модели развертывания

Развертывание ИИ-приложений требует соблюдения дифференцированных требований к времени отклика, одновременно учитывая критические аспекты, включая безопасность данных, эластичное масштабирование ресурсов и системное обслуживание. Благодаря координации городских ИИ-центров обработки данных (AIDC), городских сетей (Metropolitan Area Networks - MAN) и различных моделей развертывания может быть построена иерархическая и

совместная система обеспечения возможностей ИИ в масштабе города. Распространенные модели развертывания включают: облачное развертывание (cloud deployment), локальное развертывание (on-premises deployment), гибридное развертывание (hybrid deployment), федеративное развертывание (federated deployment) и периферийное развертывание (edge deployment).

Облачное развертывание (Cloud Deployment), провайдеры интернет-услуг обычно предлагают облачное развертывание для обеспечения быстрого предоставления AI-приложений и широкого покрытия пользователей. Ведущие предприятия обычно строят собственные AIDC для поддержки своих собственных сервисных требований, одновременно предлагая услуги аренды вычислительных мощностей. Для малых и средних предприятий создание собственных AIDC влечет за собой высокие инвестиции и затраты на обслуживание, что делает их более склонными арендовать вычислительные мощности для быстрого развертывания и итерации ИИ-приложений.

Локальное развертывание (On-Premises Deployment) особенно подходит для таких отраслей, как финансы, здравоохранение и производство, которые требуют строгой безопасности данных и соответствия нормам. Этот подход позволяет предприятиям сохранять полный контроль над своими данными, обеспечивая всю обработку и хранение данных в пределах их внутренних сетей, одновременно обеспечивая сверхнизкую задержку доступа к приложениям. Однако непрерывное масштабирование ИИ-моделей приводит к непомерно высоким затратам и требовательным операционным требованиям для локального развертывания.

Гибридное развертывание (Hybrid Deployment) сочетает преимущества облачного и локального развертывания, позволяя предприятиям обрабатывать конфиденциальные данные локально, одновременно используя облачные ресурсы для обработки неконфиденциальных данных. Предприятия могут обрабатывать критичные к задержке задачи локально, одновременно стратегически передавая ресурсоемкие или второстепенные workloads в облако, оптимизируя как инвестиции в локальное оборудование, так и операционные

расходы. Гибридное развертывание предоставляет предприятиям сбалансированное решение по производительности, безопасности и стоимости, что делает его одним из все более предпочтительных подходов к развертыванию ИИ-приложений.

Федеративное развертывание (Federated Deployment) использует распределенные вычисления, чтобы позволить нескольким предприятиям совместно обучать более эффективную глобальную модель без обмена приватными данными. Конкретно, каждый участник обучает ИИ-модель локально, затем передает зашифрованные параметры модели на центральный сервер для агрегации, генерируя улучшенную глобальную ИИ-модель, которая впоследствии распространяется среди всех участников. Федеративное развертывание способствует совместному обучению нескольких участников при сохранении конфиденциальность данных, предоставляя инновационный и практичный подход к развертыванию для ИИ-приложений.

Периферийное развертывание (Edge Deployment) предназначено для сценариев, требующих обработки в реальном времени и быстрого реагирования, таких как автономное вождение, промышленные системы управления и «умный дом». Например, в автономном режиме периферийно развернутые ИИ-приложения позволяют анализировать в реальном времени данные с камер, радаров и датчиков для обеспечения мгновенного принятия решений, гарантируя быструю реакцию на изменения окружающей среды. В промышленных системах управления периферийно развернутые ИИ-приложения поддерживают непрерывную работу даже без стабильного сетевого соединения, гарантируя бесперебойное производство.

2.3 ИИ-приложения предъявляют новые требования к городским сетям

Городские сети объединяют разнородные вычислительные ресурсы и различные пользовательские терминалы в регионе, обеспечивая подключение для различных моделей развертывания, включая облачное развертывание и гибридное развертывание, и служит в качестве критически важной

инфраструктуры для устойчивого развития искусственного интеллекта. Развитие городских сетей по аналогии с городскими электросетями или водопроводными сетями для обеспечения «одна точка доступа, вычисления по требованию» (one-point access, on-demand computing) постепенно становится общепринятым в отрасли.

ИИ-модели, требуемые для различных сценариев применения, демонстрируют значительные различия, которые можно классифицировать по масштабу на два различных типа: большие модели (large models) и малые модели (small models). Малые ИИ-модели обычно относятся к моделям с меньшим количеством параметров и менее глубокими слоями, характеризуются их облегченной архитектурой, вычислительной эффективностью и гибкостью развертывания. Эти модели специально оптимизированы для выделенных задач и вертикальных доменов, с репрезентативными реализациями, включающими DistilBERT, TinyBERT и MobileNet. Большие ИИ-модели относятся к моделям с огромным количеством параметров и сложными вычислительными архитектурами, демонстрируя расширенные репрезентационные возможности и превосходную точность для решения более сложных задач, с репрезентативными примерами, включающими Deepseek, GPT-4, Qwen. Разнообразные ИИ-модели предъявляют значительно дифференцированные требования к городским сетям.

2.3.1 Требования больших моделей ИИ

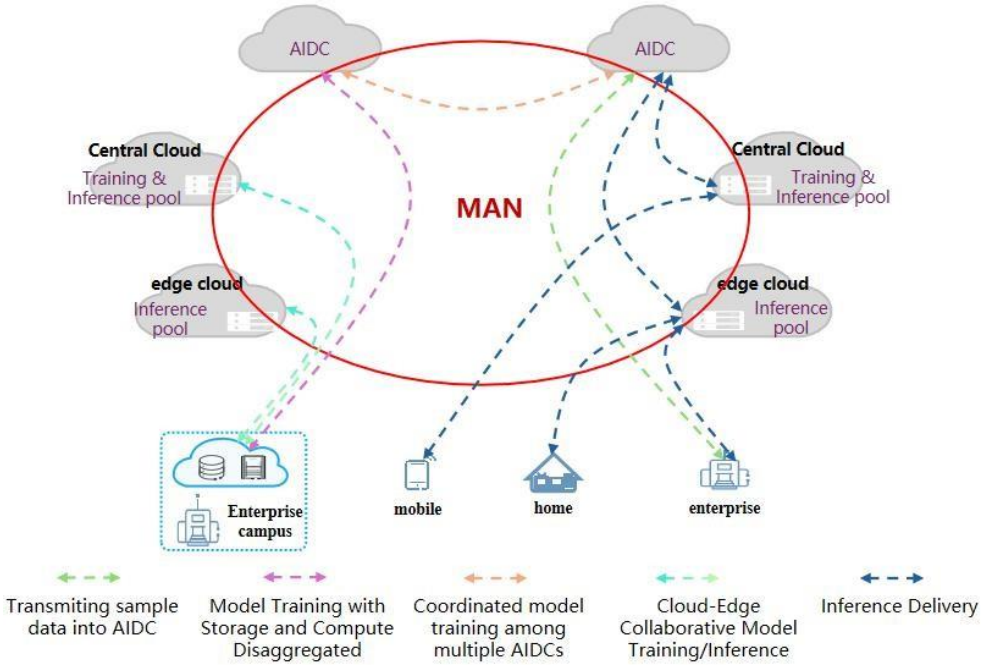

Рисунок 2-3-1: Требования для больших моделей ИИ

Жизненный цикл крупных ИИ-моделей охватывает несколько этапов, включая передачу выборочных данных, обучение модели и инференс (логический вывод) модели, каждый из которых предъявляет различные характеристики передачи данных с точки зрения объема и шаблонов, что, в свою очередь, предъявляет повышенные требования к городским сетям.

1. Передача выборочных данных в AIDC.

С быстрым развитием больших моделей искусственного интеллекта объем данных растет беспрецедентными темпами. Согласно отчету Global DataSphere 2023, опубликованному IDC (International Data Corporation), объем данных в Китае достиг примерно 30 ЗБАЙТ в 2023 году и, по прогнозам, увеличится до 76,6 ЗБАЙТ к 2027 году. В настоящее время многие предприятия по-прежнему полагаются на доставку физических жестких дисков для передачи образцов данных. Такой подход «ручное копирование+физическая доставка» не только неэффективен, но и сопряжен с риском потери данных. Существующие сетевые решения имеют существенные ограничения: традиционные службы выделенной

линии используют модели ежемесячной / годовой подписки с фиксированной пропускной способностью, в то время как предприятиям обычно требуется только периодическая выборочная передача данных, что приводит к высоким затратам по сравнению с фактическим использованием. Городские сети требуют повышения возможностей до предоставлять более эффективные и оптимизированные по затратам услуги по передаче выборочных данных.

Городская сеть должна поддерживать балансировку нагрузки в масштабе сети для достижения устойчивой сверхвысокой пропускной способности, превышающей 90% по всем каналам связи, обеспечивая эффективную почасовую передачу выборочных данных в терабайтном (ТБ) масштабе от предприятия к AIDCs. Одновременно городские сети должны обладать высокоэластичными и гибкими сервисными возможностями, предлагая предприятиям гибкую полосу пропускания по требованию за счет предоставления услуг ‘точно в срок’ на основе задач, одновременно предоставляя многоуровневые услуги передачи данных (на уровне минут, часов и суток) для удовлетворения разнообразных запросов пользователей. Кроме того, городские сети должны обладать интеллектуальными возможностями управления вычислительной мощностью для динамического подбора оптимальных ресурсов вычислительной сети и путей передачи на основе характеристик обслуживания, включая происхождение, тип и зону покрытия, создавая тем самым более гибкую и эффективную систему предоставления вычислительной мощности.

2. Обучение модели с разделением хранения и вычислений

Многие отрасли обрабатывают конфиденциальные данные с критическими требованиями безопасности, такими как экспериментальные данные и данные об инцидентах в автомобильной промышленности, или записи о транзакциях потребителей и персонально идентифицируемая информация в финансовом секторе. При аренде ресурсов облачных вычислений эти организации или предприятия требуют строгого локального хранения данных и гарантированной защиты от утечки данных во время обучения модели. Для удовлетворения этих требований безопасности данных обучение модели требует архитектуры с разделением хранения и вычислений (с вычислительными узлами, развернутыми

в облаке, и узлами хранения, остающимися на территории предприятия), где данные обучения подгружаются в память по требованию без записи на диски вычислительных узлов.

В этом сценарии выборочные данные записываются непосредственно с узлов хранения в память вычислительных узлов через городские сети с помощью технологии RDMA. Современные основные протоколы RDMA полагаются на механизмы повторной передачи Go-Back-N, что делает их чрезвычайно чувствительными к задержкам и потере пакетов (даже уровень потерь пакетов в 0,1% может снизить вычислительную производительность на 50%). Поэтому городские сети должны не только поддерживать высокоустойчивую и высокопроизводительную передачу данных, но и включать точное управление потоком для гарантирования передачи RDMA без потерь, обеспечивая ухудшение вычислительной эффективности менее чем на 5% в городских доменах протяженностью 100-500 км. Более того, городские сети должны развертывать надежные механизмы шифрования данных для обеспечения безопасности передачи данных.

3. Скоординированное обучение модели между ИИ-ЦОД

Закон масштабирования (Scaling Law) для больших AI-моделей сохраняется, при этом потребности в вычислительной мощности выросли примерно в миллион раз за последнее десятилетие, и, по прогнозам, сохранят годовой темп роста, превышающий 400%. Масштабируемость отдельных пулов вычислительных ресурсов ограничена ограничениями физической инфраструктуры, включая пространство и энергоснабжение. Скоординированное обучение модели между несколькими ИИ-ЦОД (AIDC) позволяет эффективно объединять географически распределенные вычислительные мощности, поддерживая обучение крупных AI-моделей в масштабах более 100000 графических процессоров (GPU). Вычислительная мощность существующих ИИ-ЦОД (AIDC), как правило, невелика (в Китае ИИ-ЦОД (AIDC) с 100-300 ПФлопс (PFlops) составляют более 70% от общего количества). Следовательно, интеграция распределенных вычислительных мощностей между центрами обработки данных, научно-исследовательскими учреждениями и поставщиками облачных услуг поможет

преодолеть географические, инфраструктурные и ограничения поставщиков для создания единой и высокоэффективной платформы услуг вычислительных мощностей.

В этом сценарии данные синхронизации плоскости параметров передаются через городские сети, в то время как данные плоскости выборок остаются храниться на территории предприятий, эффективно изолируя потенциальные риски утечки данных. Это решение предъявляет строгие требования к пропускной способности сети и задержкам, обязывая городские сети развертывать каналы 400G/800G с передачей RDMA без потерь для гарантии нулевой потери пакетов во время обучения модели. Синхронизация параметров между графическими процессорами в основном основана на коллективных коммуникационных операциях AllGather/AllReduce, что создает значительные проблемы в условиях высококонкурентных и пакетных моделей трафика. Если взять в качестве примера обучение модели с 1000 миллиардами параметров, то один цикл синхронизации параметров в кластере ИИ с 16 тыс. графических процессоров генерирует более 1,6 ПБ одновременного трафика. Следовательно, городские сети требуют обновление возможностей устройств (включая буферы портов на уровне ГБ и организацию очередей на уровне арендаторов) для оптимизации обработки пакетного трафика и коллективного планирования связи, а также создание сетевой архитектуры с высоким коэффициентом сходимости (4:1, 8:1, 16:1) для баланса между вычислительной эффективностью и затратами на развертывание. Более того, сбои сети, вызывающие критические проблемы, такие как прерывания задач обучения, значительно снижают эффективность обучения. Городские сети должны реализовать изоляцию сегментации сети на уровне арендаторов и включить технологии моделирования и самовосстановления сети для реализации автономной сети 4-го уровня, гарантирующей контролируемый масштаб последствий сбоев и быстрое восстановление обслуживания.

4. Совместное облачно-периферийное (Cloud-Edge) обучение/инференс модели

Резкое снижение затрат на обучение и инференс больших моделей позволило предприятиям быстро внедрять AI-приложения путем локального развертывания серверов AI Training & Inference. Однако локальные пулы вычислительных ресурсов предприятий сталкиваются со значительными проблемами при расширении мощностей и высокими операционными затратами на обслуживание, что делает их неадекватными для удовлетворения растущих потребностей в дообучении (fine-tuning) моделей и инференсе. Для решения этого, облачно-периферийное взаимодействие между локальными пулами предприятий и облачными пулами вычислительных ресурсов представляет собой более эффективный, гибкий и экономичный подход для реализации эластичного масштабирования вычислительных мощностей. Это решение использует методы параллельных вычислений, включая pipeline parallelism и expert parallelism, для разделения крупных AI-моделей между локальными и облачными пулами вычислительных ресурсов. Путем реализации локального развертывания входных/выходных embedding слоев, оно обеспечивает строгое локальное хранение выборочных данных, тем самым выполняя требования безопасности данных для строго регулируемых секторов, таких как финансовый и здравоохранение.

Для этого сценария городские сети должны поддерживать передачу RDMA без потерь, чтобы предотвратить значительное снижение вычислительной эффективности, вызванное потерей пакетов. Одновременно городские сети требуют сетевого сегментирования (network slicing) на уровне арендатора для обеспечения эффективной изоляции сервисов, удовлетворяя требованиям SLA, одновременно предотвращая влияние сбоев других сервисов. Кроме того, городские сети должны обладать возможностями интеллектуального планирования вычислительных мощностей для динамического выбора оптимальных периферийных пулов ресурсов на основе местоположения пользователя и сервисных потребностей, обеспечивая эффективные процессы дообучения модели/инференса.

5. Доставка инференса (Inference Delivery)

ИИ-инференс позволяет применять крупные ИИ-модели в реальных сценариях, являясь критически важным шагом для коммерциализации. К 2027 году примерно 70% новых приложений, как ожидается, будут включать модели ИИ-инференса, при этом параллельные транзакции между ИИ-приложениями и пулами ресурсов, как ожидается, достигнут порога в миллион масштаба. Доставка инференса состоит из двух основных процессов. Доставка модели, означающая развертывание моделей ИИ-инференса в нескольких периферийных облаках и доставка результата, обозначающая взаимодействие между пользователями и моделями ИИ-инференса для генерации требуемых выходных данных.

В этом сценарии городские сети должны обеспечивать возможности передачи данных с низкой задержкой и высокой пропускной способностью с повсеместным покрытием, и беспрепятственным доступом для обеспечения качества обслуживания приложений искусственного интеллекта. Городские сети также должна включать возможности детерминированного обслуживания, позволяющие точно идентифицировать трафик и оптимизировать выбор маршрута для повышения детерминированности передачи и надежности.

2.3.2 Требования малых AI-моделей

Малые ИИ-модели отличаются компактной архитектурой, низкими вычислительными потребностями и возможностями быстрого реагирования. Эти модели обычно разработаны для специализированных задач и демонстрируют уникальные преимущества в средах с ограниченными ресурсами, таких как смарт-терминалы и IoT-устройства. С их широким развертыванием в системах «умного дома», промышленных IoT-приложениях и на мобильных платформах, они предъявляют больше требований к городским сетям.

1. Выдача выводов

В режиме реального времени ИИ-инференса малые ИИ-модели преимущественно развертываются на периферийных устройствах вблизи источников данных, обеспечивая мгновенную обработку входных данных и

генерацию прогнозных результатов для достижения сверхнизкой задержки. Для сценариев, требующих большей вычислительной мощности, периферийные узлы взаимодействуют с облаком в гибридной архитектуре развертывания. Периферийные устройства обрабатывают рутинные высокочастотные запросы инференса локально, в то время как computationally intensive или аномальные случаи передаются через городские сети в облако для глубокого анализа. Городские сети должны обеспечивать гарантированную пропускную способность, детерминированный путь с низкой задержкой и возможности интеллектуального оркестрирования трафика для обеспечения надежного предоставления услуг ИИ-инференса в режиме реального времени.

2. Федеративное обучение

Федеративное обучение - важнейшая парадигма обучения для малый AI-моделей. Оно использует федеративное развертывание, которое значительно повышает эффективность модели, одновременно обеспечивая сохранение конфиденциальности локальных данных, предъявляя три критических требования к городской сети. Во-первых, синхронизация параметров в реальном времени требует гарантированного периодического подключения для участников для поддержания непрерывности обучения. Во-вторых, безопасность передачи данных требует сквозного шифрования для параметров модели для предотвращения любой потенциальной утечки модели. В-третьих, городские сети должны включать возможности динамического распределения ресурсов, выделяя большую пропускную способность участникам с более высоким приоритетом на основе их различного прогресса обучения.

2.3.3 Требования гибридных AI-моделей

Гибридные AI-модели развертывают облегченные малые AI-модели на периферии, в то время как крупные AI-модели с расширенными возможностями понимания и логического рассуждения размещаются в облаке. Благодаря эффективному взаимодействию между этими моделями, гибридные AI-модели полностью используют преимущества малой модели в response с низкой задержкой и персональной адаптации, одновременно используя возможности

крупной модели в многомодальном понимании и обобщенном интеллекте

Координация между крупными AI-моделями и малыми AI-моделями в основном отражается в двух критических аспектах: взаимодействие данных и обновление модели. Для взаимодействия данных, развернутые на периферии малые AI-модели выполняют локальный сбор и предварительную обработку данных перед передачей критических данных в развернутые в облаке крупные AI-модели для анализа, с последующей доставкой результатов вычислений обратно на периферийные устройства для исполнения. Этот процесс требует, чтобы городские сети предоставляли детерминированные сервисные возможности, обеспечивающие передачу данных с низкой задержкой, высокой пропускной способностью и высокой стабильностью. Для обновления моделей, развернутые в облаке крупные AI-модели могут распределять оптимизированные параметры или модели на периферийные устройства с помощью таких методов, как дистилляция знаний (knowledge distillation), обеспечивая непрерывную итерацию малых AI-моделей. Этот процесс опирается на способность балансировки нагрузки на уровне сети городских сетей для реализации высокой пропускной способности, особенно во время одновременных обновлений на массивных количествах периферийных устройств.

2.4 ИИ-приложения ведут развитие городских сетей к следующему поколению

Быстрое развитие AI-приложений предъявляет все более строгие требования к городским сетям. На архитектурном уровне городские сети должны поддерживать эффективное направление трафика север-юг и восток-запад для удовлетворения требований облачно-периферийной и межоблачной координации, одновременно обеспечивая эластичную масштабируемость для достижения повсеместного доступа пользователей. На техническом уровне городские сети должны включать в себя такие возможности, как балансировка нагрузки в масштабе сети, точное управление потоками на уровне потоков и сетевые решения с высоким коэффициентом переподписки для поддержки сверхбольшого трафика вычислений ИИ, обеспечивая при этом интерактивный

опыт в режиме реального времени посредством детерминированных сервисов и сегментации сети на уровне арендаторов. На операционном уровне городские сети должны укреплять сервисно-ориентированные возможности для предоставления гибких и проворных услуг вычислительных мощностей для пользователей, одновременно усиливая возможности интеллектуальной эксплуатации и технического обслуживания (O&M) для обеспечения высокой стабильности и надежности сервисов.

Глава III Архитектура городских сетей в эпоху ИИ

3.1 Задачи проектирования городских сетей

Проектирование городских сетей в эпоху ИИ ориентировано на создание сетевой инфраструктуры следующего поколения с глубокой интеграцией вычислений и сети, которая является интеллектуальной, эффективной, безопасной и надежной. Основными направлениями являются следующие:

1. Интегрированные вычисления и сеть, конвергентная транспортировка

На основе пула вычислительных ресурсов городская вычислительная сеть может интегрировать разнородные вычислительные мощности общих вычислений, интеллектуальных вычислений и суперкомпьютеров. Такие технологии, как SRv6, и другие технологии используются для единообразного планирования и разумного управления сетевыми, облачными и вычислительными ресурсами, устраняя физическую изоляцию. Передача гетерогенных вычислительных мощностей между доменами без потерь и совместное обучение нескольких центров оказания помощи поддерживается для создания основы для инноваций в области синергии облачных сетей. Единый доступ к фиксированным, мобильным и облачным сервисам и конвергентный носитель множества сервисов обеспечивает повсеместный доступ пользователей. Городские сети создают интеллектуальную, гибкую, безопасную и надежную высококачественную сетевую инфраструктуру для эффективной поддержки эффективной совместной работы многомерных служб и предоставления комплексных услуг подключения по всем сценариям для цифровой трансформации и интеллектуальной модернизации отраслей.

2. Эластичность, манёвренность, гибкость и эффективность

На основе модульной архитектуры Spine-Leaf и технологической основы IPv6 Enhanced может быть достигнуто гибкое расширение сети и предоставление услуг за считанные минуты. Интеллектуальная идентификация больших потоков (elephant flows) и планирование на уровне сетевого потока обеспечивают балансировку нагрузки на сетевом уровне и усовершенствованное управление потоками услуг, обеспечивая высокую пропускную способность и низкую задержку при передаче данных, реализуя быструю обработку трафика и удобный

доступ для пользователей, а также всесторонне повышая общую эффективность передачи данных по сети.

3. Точное управление и динамическая конвергенция

На основе технологии интеллектуальной идентификации потока и точного управления потоком, а также механизма детерминированной передачи данных с задержкой и оптимизации конвергенции сети, построена высокопроизводительная архитектура межсоединений RDMA без потерь. Благодаря возможности интеллектуального планирования на уровне потока и гибкой сетевой архитектуре, ориентированной на вычислительную мощность, поддерживается динамическое сотрудничество между вычислительными мощностями предприятия и центрами-концентраторами вычислительной мощности. Адаптация потока обслуживания по требованию и точное управление ресурсами эффективно поддерживают требования к пропускной способности данных на уровне ТВ для обучения и вывода больших моделей искусственного интеллекта, достигая оптимального баланса между затратами на построение сети и эффективностью вычислений.

4. Интеллектуальное управление, безопасность и надежность

Интеллектуальная система управления, основанная на искусственном интеллекте, используется для создания интеллектуальных возможностей по вводу в эксплуатацию, таких как оптимизация планирования на уровне потока, самовосстановление неисправностей и сетевое моделирование. Архитектура двухуровневого резервирования и междоменный механизм аварийного восстановления обеспечивают высокую доступность системы. Разделение сети, управление потоками на уровне клиента и стандартные интерфейсы безопасности используются для создания многоуровневой системы изоляции безопасности. Интеграция с такими технологиями, как гарантия передачи без потерь пакетов и сквозное обнаружение деградации качества опыта (QoE), обеспечивает надежную передачу по всем путям данных и резервную защиту для многоплоскостных вычислительных границ, гарантируя безопасный и надежный запуск сервиса на протяжении всего жизненного цикла услуги

3.2 Общая архитектура городских сетей

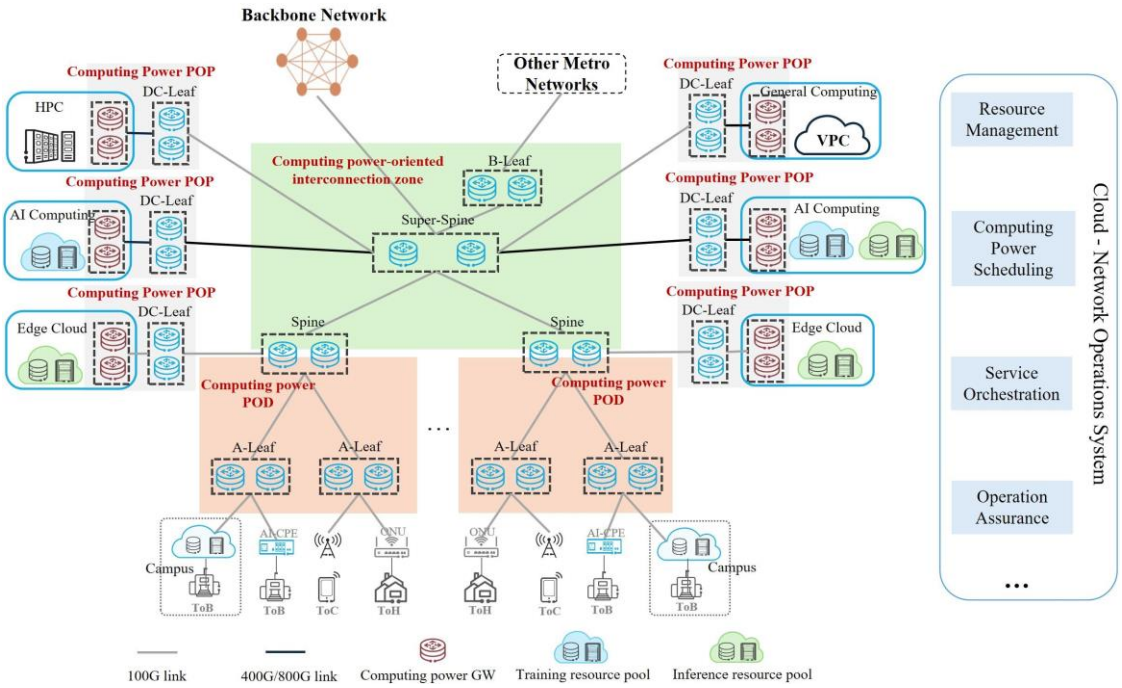

Рисунок 3-2: Архитектура городских сетей для эпохи ИИ

Архитектура городской сети для эпохи искусственного интеллекта состоит из трех основных модулей: зона POD (Point Of Delivery), ориентированная на вычисления; зона POP (Point Of Presence), ориентированная на вычисления; зона взаимосвязей, ориентированная на вычисления. Эти три модуля беспрепятственно взаимодействуют с облачной операционной системой «облако-сеть» через стандартные интерфейсы и протоколы. Используется настраиваемая и поддерживающая «горячую замену» блочная (building-block) архитектура для достижения эластичного масштабирования вычислительных ресурсов по требованию.

– Зона POD, ориентированная на вычислительную мощность Этот модуль вводит POD в центр обработки данных в городскую сеть. Основанный на модульной архитектуре «позвоночник-лист» (spine-leaf modular architecture), модуль обеспечивает эффективный доступ к клиентским

терминалам и филиалам предприятия через оптоволоконные сети, PON и 5G, а также поддерживает обмен данными большой емкости, быструю конвергенцию и перенаправление трафика фиксированных и мобильных услуг, а также интеллектуальных вычислительных сервисов в регионе. Технологии SRv6 и EVPN используются для предоставления нескольких услуг унифицированным образом. Облачная и сетевая операционная система используется для автоматического предоставления услуг и интеллектуального обслуживания, и эксплуатации. Технология сетевой нарезки (network slicing technology) обеспечивает настраиваемую пропускную способность и безопасность для интеллектуальных вычислений, промышленности и государственных услуг.

– Зона POP, ориентированная на вычислительные мощности. В качестве точки соединения между облачной сетью и опорной сетью, этот модуль стандартным образом соединяется с пулом вычислительных ресурсов для реализации планирования по требованию и гибкого распределения вычислительных ресурсов для поддержки интегрированных вычислительных сетевых услуг. Выполняет функцию сетевого якоря пула вычислительных ресурсов, подключается к провинциальным/региональным магистральным узлам, открывает межвычислительные каналы и поддерживает междоменную совместную работу с ресурсами и восстановление после сбоев. Взаимодействует с зоной POD, ориентированной на вычислительную мощность, для обеспечения сквозного соединения без потерь между пользователями в разных POD и пулами вычислительных ресурсов.

– Зона интерконнекта, ориентированная на вычислительную мощность: Этот модуль служит хабом между городской сетью, магистральной сетью, Интернетом и промышленными частными сетями. Это упрощает соединение между городскими сетями и внешними сетями, а также между различными пулами вычислительных ресурсов, реализует гибкое расширение компонентов и эффективно перенаправляет трафик между

компонентами. Используются высокоскоростные каналы связи 400G/800G, балансировка нагрузки на сетевом уровне и технологии SRv6/EVPN для достижения эффективной переадресации междоменного трафика и оптимизации путей. Технология сетевого сегментирования обеспечивает дифференцированную пропускную способность и гарантию безопасности интеллектуальных вычислительных служб, обеспечивая стабильное взаимодействие служб и удобство работы пользователей.

Три модульные зоны вместе составляют архитектуру городской сети, ориентированную на эпоху искусственного интеллекта. Каждая модульная зона играет определенную роль, чтобы обеспечить эффективную транспортировку AI-сервисов в городской сети. Зона POD, ориентированная на вычисления, функционирует как точка входа доступа пользователей и соединяется с зоной POP, ориентированной на вычисления, через магистральные устройства (spine), чтобы построить эффективные каналы передачи между пользователями и пулом ресурсов вычислительных мощностей. Зона интерконнекта и зона POP, ориентированные на вычислительную мощность, подключены к сети взаимодействия междоменных пулов вычислительных мощностей для реализации интеллектуального планирования ресурсов вычислительных мощностей. Три модуля используют стандартные технологии, такие как SRv6 и EVPN, для обеспечения согласованности логики сквозного обслуживания и предоставления высококачественных сетевых возможностей передачи данных для служб искусственного интеллекта.

На основе концепции иерархической развязки и совместного проектирования, архитектура создает интегрированную сеть вычислительных услуг, обеспечивающую периферийный доступ, планирование ядра и междоменную совместную работу. Она использует операционную систему «облако-сеть» для реализации единого управления и контроля, а также интеллектуального планирования ресурсов по всей сети. Операционная система облачной сети сосредоточена на четырех основных модулях: управление ресурсами, планирование вычислительных мощностей, оркестрирование

сервисов и обеспечение операционной деятельности (operation assurance). Управление ресурсами интегрирует сетевые и вычислительные ресурсы для достижения глобальной видимости и управления, а планирование вычислительных мощностей динамически оптимизирует распределение ресурсов на основе сервисных требований. Оркестрирование сервисов реализует быстрое развертывание сервисов и сквозную интеграцию через автоматизированные процессы. Обеспечение операционной деятельности использует технологии интеллектуального мониторинга и анализа для обеспечения стабильной работы системы и пользовательского опыта. Совместная работа модулей обеспечивает прочную основу для требований к вычислительной мощности в эпоху ИИ: высокой параллельной обработки (concurrency), низкой задержки и высокой надежности.

3.3 Ключевые модули городской сети

3.3.1 Зона POD (Point Of Delivery), ориентированная на вычисления

Зона POD является периферийным уровнем доступа к городским сетям и обеспечивает конвергентный доступ для клиентских терминалов (2C), филиалов предприятия (2B) и домашних пользователей (2H). Агрегирует трафик по уровням через базовые станции, CPE, конечные узлы (leaf) и магистральные узлы (spine) для формирования широкого охвата и гибкой системы предоставления вычислительных услуг. Кроме того, глубокие (deep) и поверхностные (shallow) вычисления могут быть установлены по требованию, предоставляя заказчикам вычислительные услуги с низкой задержкой и высоким уровнем производительности. Его основные функции включают:

- Конвергентный доступ поддерживает несколько режимов доступа, таких как оптоволокно, PON, и 5G, реализующие «однострочность для многочисленных вычислений». Одна линия может удовлетворить требования к доступу в Интернет, облачные сервисы и пулы с несколькими вычислительными мощностями.

– Эластичная пропускная способность обеспечивает возможности гибкого доступа от 0 до 100 Гбит / с, адаптируясь к изменениям требований клиентов к вычислительной мощности.

– Планирование объединения ресурсов поддерживает объединение вычислительных мощностей на глубоких и поверхностных границах и планирование между блоками, гибкое покрытие в зависимости от масштаба обслуживания, обеспечивая эффективную передачу вычислительных ресурсов.

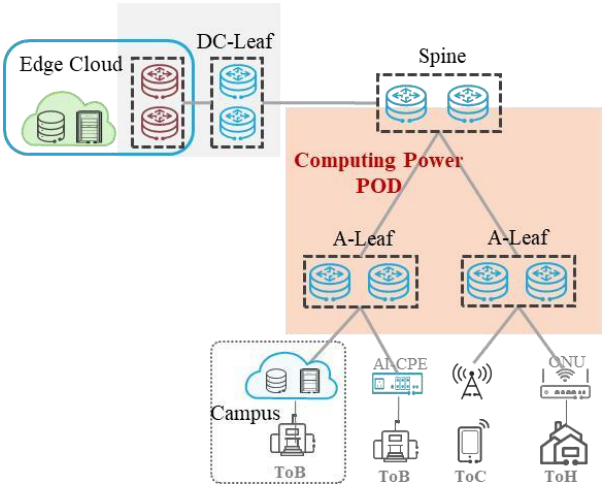

Рисунок 3-3 Модульная зона, ориентированная на вычислительную мощность

POD-зона, ориентированная на вычислительную мощность, использует сетевую архитектуру с широким охватом и многоуровневой конвергенцией, чтобы динамически сбалансировать использование ресурсов вычислительной мощности при одновременном снижении затрат на покрытие сети. Точное планирование и контроль на уровне потока обеспечивают качество передачи данных на основе протокола RDMA и эффективно поддерживают передачу данных на большие расстояния без потерь в сценариях, где быстро обрабатываются большие выборки и разделены хранилище и вычисления. Кроме того, операционная система обладает возможностью гибкой настройки ресурсов. Благодаря динамической настройке сетевых путей и ресурсов полосы пропускания операционная система эффективно справляется с колебаниями

трафика услуг и обеспечивает непрерывность и стабильность обслуживания.

3.3.2 POP-зона, ориентированная на вычислительную мощность

Зона POP, энергориентированная на вычислительную мощность, соединяет городскую сеть и пул вычислительных ресурсов через шлюз вычислительной мощности, реализуя стандартизированное и быстрое соединение между сетью плоскости выборки и сетью плоскости обслуживания городской сети, и пулом вычислительных ресурсов. Зона POP вычислительной мощности предоставляет стандартизированные политики соединения функциональных зон и руководство по развертыванию, поддерживая интегрированное несение и планирование ресурсов нескольких сервисов. Его основные функции включают:

- Модульная сеть: стандартные модули подключаются к гетерогенным пулам вычислительных ресурсов (собственных или сторонних) для реализации объединения ресурсов и единого управления.
- Сквозное соединение без потерь: подключается к провинциальным/региональным магистральным узлам и связывается с несколькими POD вычислительной мощности, чтобы обеспечить соединения с низкой задержкой и высокой надежностью между пользователями в разных POD и пулах вычислительной мощности.
- Поддержка интеллектуальных вычислительных сервисов: точное управление потоком на уровне потока используется для удовлетворения требований сервисов, таких как расчет входных данных выборки, обучение моделей с распределением хранения и вычислений по AIDC, а также совместное обучение между кластерами.

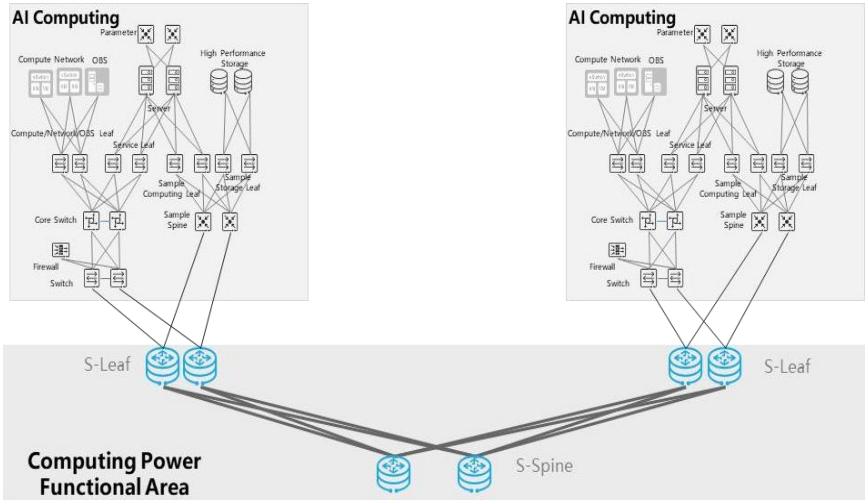

Рисунок 3-3-2: POP-зона, ориентированная на вычислительную мощность

Зона POP, ориентированная на вычислительную мощность, поддерживает конвергенцию нескольких сервисов и различные услуги, такие как общие вычисления и интеллектуальные вычисления. Являясь центром планирования трафика между северной и южной частями, вычислительный шлюз интеллектуально соединяется с сетями плоскости выборки и плоскости обслуживания пула вычислительных ресурсов посредством стандартных политик взаимосвязи. POP вычислительной мощности устанавливает высокоскоростные соединения с провинциальными или региональными магистральными узлами и несколькими POD, образуя сквозное соединение между пользователями, вычислительными POD, вычислительными POP и пулами вычислительных ресурсов.

3.3.3 Зона интерконнекта, ориентированная на вычислительную мощность

Являясь основным узлом-концентратором городской сети, зона интерконнекта, ориентированная на вычислительную мощность, соединяет POD-зону, ориентированную на вычислительную мощность, и POP-зону, ориентированную на вычислительную мощность, с магистральной сетью и выходом в Интернет по высокоскоростным каналам связи 400G/800G. Технология разделения сети обеспечивает дифференцированную пропускную способность и гарантию безопасности интеллектуальных вычислительных

сервисов, обеспечивая стабильное взаимодействие служб и удобство работы пользователей. Ее основные функции включают:

- Дифференцированные услуги: на основе таких технологий, как точная идентификация трафика для классификации и маркировки различных потоков услуг, с целью обеспечения дифференцированного качества обслуживания для различных типов услуг и удовлетворения различных требований к задержке, пропускной способности и коэффициенту потери пакетов различных услуг. Обеспечение приоритетной обработки и лучших сетевых ресурсов для критически важных и высокоценных услуг.
- Планирование и управление трафиком: планирует и управляет трафиком в каждой функциональной зоне городской сети единым образом и направляет трафик на различные каналы и пути в зависимости от нагрузки сети, требований к услугам и заранее определенных политик. Таким образом, трафик распределяется равномерно, и можно улучшить использование сетевых ресурсов.
- Высокоскоростное межсетевое соединение: в качестве центра соединения городской сети с магистральными сетями, другими аналогичными сетями, POP, ориентированными на вычислительную мощность, и POD, ориентированными на вычислительную мощность, использует каналы 400G/800G для реализации высокоскоростного соединения между различными сетями. Обменивается маршрутной информацией с внешними сетями, обеспечивает правильную пересылку пакетов данных между городской сетью и внешними сетями, а также обеспечивает бесперебойную передачу различных услуг между различными сетевыми доменами.

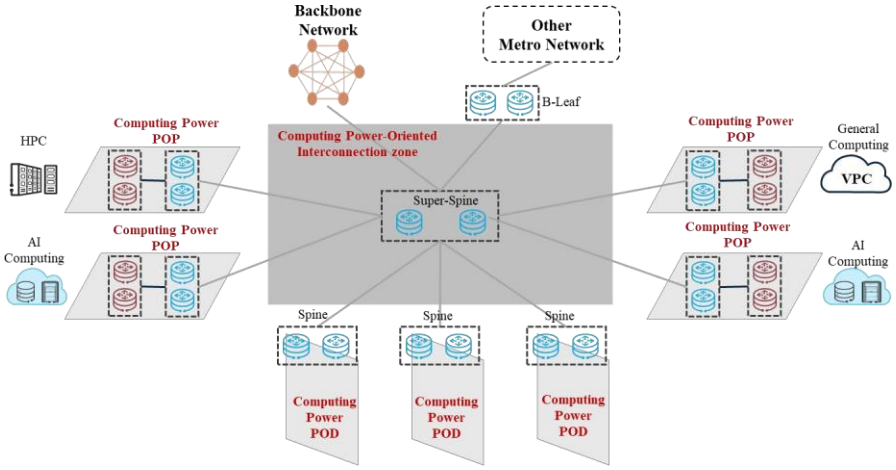

Рисунок 3-3-3: Зона интерконнекта, ориентированная на вычислительную мощность

Зона межсетевого взаимодействия, ориентированная на вычислительную мощность, создает интегрированную базу городских сетей для «высокоскоростного межсетевого взаимодействия и интеллектуального планирования». Поддержка инноваций в сфере услуг за счет дифференцированных услуг, повышение эффективности использования вычислительных ресурсов за счет интеллектуального планирования, устранение барьеров для междоменного сотрудничества, обеспечение циркуляции вычислительной мощности и конвергентных приложений новой сетевой инфраструктуры.

Глава IV Ключевые технологии городских сетей в эпоху ИИ

4.1 Интегрированные вычисления и сеть, конвергентная сеть передачи данных

4.1.1 Универсальная сеть носителей услуг

В эпоху искусственного интеллекта городские сети сталкиваются с новыми вызовами, вызванными резким ростом вычислительной мощности, требований к совместной работе и передаче данных. Поэтому для одновременной поддержки нескольких фиксированных, мобильных, облачных и вычислительных сервисов крайне необходим единый стек протоколов, что снижает сложность сети. Значительно повышается эффективность развертывания сервисов, а также эксплуатации и обслуживания. Конвергентная архитектура на базе SRv6 и EVPN представляет собой идеальное решение для создания единой сети передачи данных. Она реализует логическую изоляцию и гибкое планирование сервисов в единой сети, избегая архитектурной избыточности, возникающей при использовании нескольких традиционных сетей, и значительно повышая эффективность использования сетевых ресурсов. Её основные преимущества:

- Унифицированный доступ пользователей для SRv6 поддерживает междоменные сквозные соединения на основе собственных протоколов IPv6. Корпоративные пользователи могут удовлетворять требованиям к нескольким сервисам, используя всего один доступ, что значительно снижает сложность доступа.
- Унифицированная поддержка сервисов для EVPN предоставляет гибкие VPN-сервисы уровня 2/3 и возможности маршрутизации от источника SRv6 для динамической адаптации к различным требованиям SLA, реализуя интеллектуальное планирование трафика и оптимизацию ресурсов.
- Удобное предоставление услуг для интеллектуальных технологий эксплуатации и обслуживания, таких как сеть автономного вождения,

обеспечивают автоматическую оркестровку услуг и предоставление услуг на поминутном уровне, значительно повышая гибкость сети. Кроме того, программируемость сети SRv6 закладывает основу для оптимизации сети на основе искусственного интеллекта, что дополнительно повышает эффективность использования сетевых ресурсов и интеллектуальность.

4.1.2 Интеллектуальное планирование вычислительных мощностей

В сценариях повсеместного использования вычислительных мощностей городские сети сталкиваются с основной проблемой динамического согласования спроса и предложения на вычислительные мощности. Поэтому необходимо решать такие ключевые проблемы, как распределение ресурсов, разнообразие требований и оперативность выполнения задач. Для городских сетей необходимо создавать интеллектуальный механизм планирования вычислительных мощностей. Этот механизм реализует динамическое ценообразование и распределение задач путем оперативного отслеживания состояния спроса и предложения и оптимизации алгоритмов, обеспечивая эффективное использование вычислительных ресурсов и удовлетворяя основные требования пользователей к низкой задержке, высокой надежности и низкой стоимости.

Основной целью интеллектуального планирования вычислительной мощности является достижение динамического согласования спроса и предложения и повышение эффективности совместного использования ресурсов вычислительной мощности. На основе географического положения, типа ресурсов и нагрузки поставщика в режиме реального времени, а также требований SLA к услугам и характеристик задач заказчика создается глобальная система отслеживания вычислительной мощности и унифицированная система измерения. В этом процессе SRv6 использует гибкие и программируемые функции для глубокой привязки планирования вычислительных мощностей и требований к услугам посредством оптимизации сетевого пути.

Интеллектуальное планирование вычислительных мощностей создает замкнутую систему, характеризующуюся динамическим распознаванием ресурсов, оптимизацией пути SRv6 и интеллектуальным принятием решений, чтобы реализовать точное планирование гетерогенных ресурсов между доменами и обеспечить низкую задержку и высокую эластичность вычислительных мощностей для таких вычислительных сценариев, как общие вычисления, интеллектуальные вычисления и супервычисления.

4.2 Эластичность, гибкость, адаптивность и эффективность

4.2.1 Планирование на основе задач

Технология планирования на основе задач облегчает передачу нереального времени задач (таких как задачи резервного копирования данных) в непиковые часы и улучшает использование простаивающих сетевых ресурсов. Эта технология основана на интеллектуальном механизме замкнутого цикла «осознание требований-прогнозирование ресурсов-динамическое выполнение», которая повышает эффективность использования ресурсов и улучшает пользовательский опыт. Во-первых, операционная система получает требования пользователей к передаче данных через стандартизированные интерфейсы и выполняет многомерную оценку осуществимости на основе исторических данных о пропускной способности, а затем возвращает пользователю подтвержденное окно времени передачи. Во-вторых, с помощью технологии сегментирования сети динамически распределяются ресурсы физических портов, и пользователям предоставляются выделенные каналы передачи данных. Наконец, качество передачи данных контролируется в режиме реального времени во время выполнения задачи, а после ее завершения ресурсы пропускной способности автоматически освобождаются.

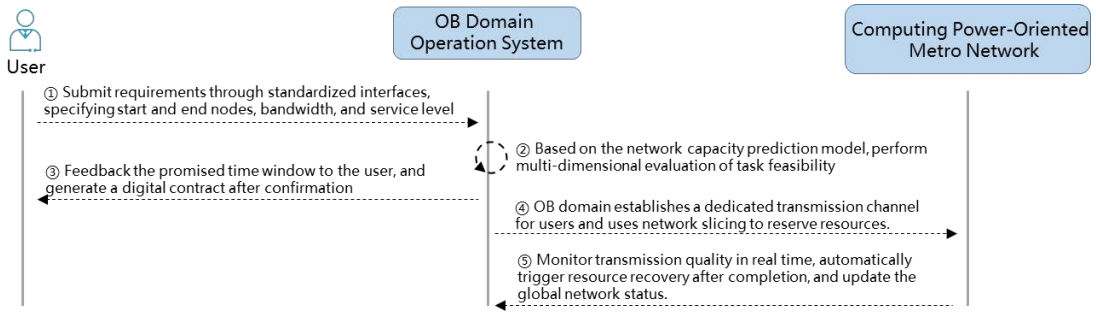

Рисунок 4-1: Процесс планирования на основе задач

Технология планирования на основе задач реализует сквозные автоматизированные процессы обслуживания, значительно оптимизирует время отклика от подачи запроса до готовности ресурсов и значительно улучшает использование ресурсов в масштабах всей сети. Эта технология устанавливает точный механизм гарантии эффективности по времени и опирается на предварительный расчет маршрута и алгоритмы динамической оптимизации, чтобы обеспечить детерминированное соблюдение эффективности времени передачи. Цифровая сеть-близнец используется для моделирования сложных задач, создания интеллектуальных возможностей планирования ресурсов и реализации предотвращения конфликтов и глобальной оптимальной координации в сценариях многозадачной параллельной работы. В результате достигается динамическое и точное согласование между предложением сетевых ресурсов и колебаниями спроса пользователей в течение секунд.

4.2.2 Эластичная пропускная способность

В типичных сценариях применения, таких как хранение больших объемов данных, предприятия сталкиваются с проблемами настройки пропускной способности, вызванными периодическими пиковыми нагрузками при передаче данных. Длительное использование выделенных линий с высокой пропускной способностью приводит к нерациональному использованию ресурсов в периоды простоя, в то время как выделенные линии с низкой пропускной способностью приводят к простоям вычислительных ресурсов из-за длительной задержки

передачи данных. Технология эластичной пропускной способности позволяет динамически расширять ресурсы пропускной способности по требованию и опирается на гибкость обслуживания домена управления и контроля, эффективно решая дилемму «высокая пропускная способность не может быть использована, а низкая пропускная способность не может быть использована». Технология эластичной пропускной способности реализует динамическое планирование ресурсов пропускной способности, выстраивая глубокое взаимодействие между сетью и операционной системой. Эта технология получает инструкции по настройке пропускной способности от пользователей на основе стандартных интерфейсов обслуживания, создает возможности автоматической оркестровки услуг на основе операционной системы и поддерживает синхронную настройку скоростей портов, политик QoS и маршрутизации на уровне минут.

Весь процесс формирует замкнутый контур управления распознаванием требований, разработкой политик и реорганизацией ресурсов, реализуя гибкое масштабирование пропускной способности отдельной выделенной линии в диапазоне от 100 Мбит/с до 10 Гбит/с/100 Гбит/с. Эта технология не только предоставляет предприятиям возможности оперативного масштабирования в режиме реального времени с точностью до минуты, эффективно справляется с мгновенными потребностями в сетевых ресурсах, вызванными пиковыми нагрузками, но и поддерживает точную тарификацию в зависимости от продолжительности и использования, значительно повышая качество сетевых услуг.

4.2.3 Каналы связи с высокой пропускной способностью

Быстрое развитие интеллектуальных вычислительных сервисов предъявляет более высокие требования к пропускной способности каналов связи. Чтобы обеспечить высокоскоростную загрузку корпоративных данных уровня терабайт/петабайт за минуты или часы, городские вычислительные сети ускоряют модернизацию каналов связи. Периферийные узлы используют

высокоскоростной доступ 100G, а на уровне агрегации развернуты каналы связи с высокой пропускной способностью 400G для передачи агрегированного трафика. Кроме того, чтобы справиться с недостаточной вычислительной мощностью одного интеллектуального вычислительного центра, необходимо интегрировать ресурсы нескольких интеллектуальных вычислительных центров для поддержки обучения больших моделей. В этом контексте высокоскоростные каналы связи 400G широко используются в сетях параметров плоскости центров обработки данных, а городские вычислительные сети необходимо модернизировать до архитектуры 400G для повышения эффективности использования полосы пропускания, возможностей динамического планирования и создания сетевой инфраструктуры с высокой пропускной способностью.

Увеличение скорости одного порта является ключевой технологией для эффективной и экономичной передачи сверхбольшого трафика. Это стало ключевым направлением развития интеллектуального вычислительного Интернета. В настоящее время технология портов 400GE для городских сетей достигла зрелости. Массовое развертывание каналов связи 400G позволяет эффективно снизить стоимость передачи одного бита в интеллектуальных вычислительных сетях, заложить основу для будущего развития технологии 800G и непрерывно оптимизировать стоимость передачи одного бита.

4.2.4 Балансировка нагрузки на уровне сети

Большая модель ИИ реализует распределенное обучение на основе агрегированного обмена данными. Трафик характеризуется высокой степенью синхронизации, большим объемом трафика и периодической передачей данных. В этом режиме обслуживания каждый путь в сети с одинаковой стоимостью одновременно передает большое количество потоков данных. В результате традиционная технология балансировки нагрузки на основе хеширования не может обеспечить полную балансировку между путями. Балансировка нагрузки

на уровне сети используется для решения проблемы потери пакетов, вызванной перегрузкой в однородной сети без сбоев в сценарии совместного обучения с использованием кросс-AIDC. В этом сценарии без сбоев сетевое устройство не имеет таких неисправностей, как повреждение оптического модуля и периодическое отключение соединения. В однородном сценарии пропускная способность и задержка сетевого устройства симметричны и синхронизированы. Эта технология эффективно повышает эффективность передачи данных в интеллектуальных сценариях совместного обучения за счет оптимизации механизма распределения трафика.

Балансировка нагрузки на уровне сети реализует бесконфликтное и сбалансированное планирование между путями посредством унифицированного планирования трафика в масштабах всей сети. В этом механизме сетевое устройство сначала собирает информацию о трафике сервиса в режиме реального времени и передаёт её сетевому контроллеру. Сетевой контроллер запускает глобальный алгоритм выбора маршрута на основе состояния топологии и характеристик трафика и интеллектуально выделяет оптимальный путь передачи для каждого потока. Наконец, контроллер передаёт решение о выборе пути сетевому устройству для его корректировки. Этот динамический механизм планирования трафика, основанный на глобальной перспективе, реализует эффективное и равномерное распределение нагрузки, достигает эффективности передачи потока более 95% и эффективно обеспечивает эффективную и стабильную работу процесса обучения.

4.3 Точное управление и динамическая конвергенция

4.3.1 Интеллектуальная идентификация и планирование больших (слонов) потоков трафика

В эпоху искусственного интеллекта транспортные характеристики городских сетей претерпевают заметные изменения. Традиционный режим обслуживания, основанный на массивных малых и микропотоках, постепенно эволюционирует

в новые формы обслуживания, такие как обучение искусственному интеллекту и распределенные вычисления, которые характеризуются высокой пропускной способностью и длительными потоками данных. Такие услуги с интенсивным трафиком подвержены перегрузке сети и приводят к общему ухудшению пропускной способности. Следовательно, интеллектуальная система идентификации и планирования трафика необходима для улучшения использования сетевых ресурсов и обеспечения эффективной передачи ключевых услуг искусственного интеллекта и общей производительности сети.

Интеллектуальная технология идентификации и планирования большого трафика (elephant flow) создает замкнутую систему оптимизации «восприятие-принятие решения-выполнение» (perception-decision making-execution) для максимального увеличения глобальной пропускной способности сети. Эта технология обнаруживает большие потоки в режиме реального времени посредством углубленного анализа характеристик трафика и в режиме реального времени передает контроллеру подробные данные, такие как характеристики потока и пропускная способность, с помощью технологии телеметрии. На основе программируемой функции SRv6 и сетевой ситуации в режиме реального времени (такой как состояние топологии и загрузка канала) контроллер устанавливает точную модель соответствия между требованиями к трафику и выделению ресурсов и динамически генерирует оптимальную политику планирования SRv6. Интеллектуально направляя потоки трафика по оптимальному маршруту, эта технология не только обеспечивает приближение пропускной способности служб ИИ к физическому пределу пропускной способности, но и значительно снижает вероятность перегрузки канала за счет точного планирования на уровне потока, создавая среду передачи с высокой пропускной способностью, малой задержкой и низкой перегрузкой, а также обеспечивая надежную гарантию для крупномасштабного обмена данными. Для службы RDMA она также может разделять поток трафика на основе информации во внутренних заголовках пакета, чтобы реализовать детальную идентификацию

и управление трафиком.

4.3.2 Точное управление потоком данных

С быстрым развитием интеллектуальных вычислительных сервисов, таких как совместное обучение с перекрестным AIDC и совместное обучение/вывод на границе облака и периферии, широкое применение протокола передачи RDMA предъявляет более высокие требования к механизму управления потоком городских сетей. В настоящее время механизм PFC широко используется в центрах обработки данных для обеспечения передачи без потерь. Однако грубое управление на уровне очереди портов подвержено проблемам блокировки заголовка и ложных повреждений. В отличие от этого, технология точного управления потоком на основе потока данных реализует точное управление противодавлением на уровне потока посредством тонкой идентификации на основе IP 5-кортежа. Эта технология не только эффективно устраняет неотъемлемые недостатки традиционного PFC, но и динамически оптимизирует политики управления потоком на основе состояния сети в реальном времени. Эта функция обеспечивает эффективную и стабильную передачу данных в сложных сценариях WAN с несколькими арендаторами и предоставляет лучшую среду для услуг RDMA.

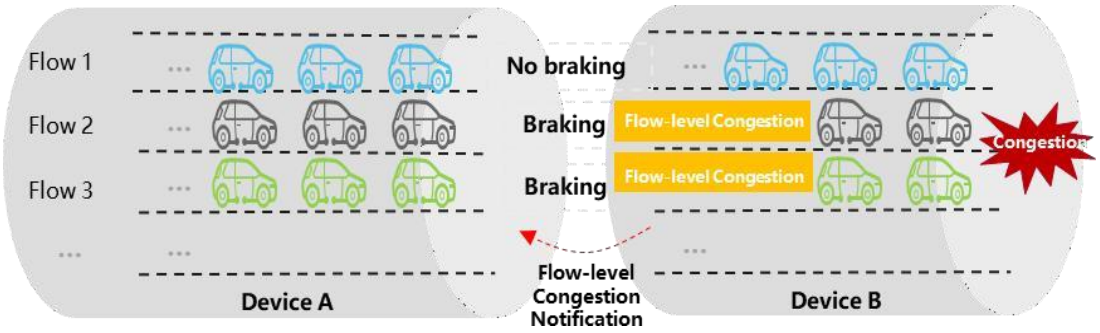

Рисунок 4-2: Технология точного регулирования расхода на уровне потока

Чтобы удовлетворить требованиям протокола RDMA к передаче данных без потерь, технология точного управления потоком на уровне потока создает систему тонкого управления. Эта технология преодолевает ограничения традиционного физического управления на уровне порта PFC. Она выделяет независимый буфер очереди для каждого потока службы RDMA и отслеживает уровень буфера в режиме реального времени, обеспечивая более точное управление трафиком. Когда происходит перегрузка определенного потока услуг, система изолирует и сохраняет пакет в специальном буфере. Когда глубина очереди превышает заданный порог, система отправляет сигналы обратного давления на уровне потока в устройство выше по потоку в режиме пошагового отслеживания. Эта технология ограничивает только скорость перегруженного потока, эффективно избегая проблем, вызванных традиционной технологией PFC, таких как перегрузка начала очереди. Практические испытания показывают, что предлагаемая технология позволяет контролировать коэффициент потери пакетов при передаче RDMA в глобальной сети на уровне ниже 0,001% и поддерживать стабильную пропускную способность на уровне выше 95%. Затем риск распространения перегрузки сети полностью устраняется благодаря механизму изоляции домена неисправностей между потоками услуг.

4.3.3 Сеть с высоким коэффициентом конвергенции

В сценарии совместного обучения с перекрестными AIDC городским сетям необходимо осуществлять крупномасштабную синхронизацию данных на плоскости параметров между несколькими центрами обработки данных. Например, скорость передачи данных 200 Гбит/с для одной сетевой карты приводит к пиковому трафику на плоскости параметров, достигающему 2000 Тбит/с. При использовании неконвергентного сетевого решения стоимость построения сети будет высокой. Поэтому сетевая технология с высоким коэффициентом конвергенции обеспечивает эффективную конвергенцию трафика совместного обучения между центрами обработки данных за счет углубленной совместной оптимизации набора алгоритмов связи и сетевой

архитектуры, что позволяет значительно снизить затраты на развертывание сетевой инфраструктуры между несколькими AIDC.

Эта технология инновационно использует совместный механизм «алгоритмическое снижение пиковых нагрузок, ограничение пиковых нагрузок кэша и ускорение планирования», который может поддерживать более 95% сквозной вычислительной эффективности в сетевой архитектуре с высоким коэффициентом конвергенции, например 32:1 и 64:1. Создана новая сетевая парадигма, адаптированная к совместному обучению между AIDC. Ее суть заключается в реконструкции агрегированного процесса связи с помощью иерархического алгоритма агрегирования градиентов, что позволяет эффективно сократить количество вычислительных карт для связи между центрами обработки данных и реализовать начальную конвергенцию сетевой пропускной способности. Кроме того, в городских сетях развернуты интеллектуальные маршрутизаторы с буферами большой емкости. Двойной механизм «буферизация пакетов + планирование очередей» используется для разделения задач обучения на контролируемые потоки микропакетов. Кэш используется для поглощения воздействия трафика, а планирование приоритетов для обеспечения своевременной передачи сигналов управления GPU. Избегается простаивание вычислительных ресурсов, что значительно снижает требования к пропускной способности сети центров обработки данных и обеспечивает эффективность обучения.

4.3.4 Детерминированная сеть обслуживания

В последние годы, с бурным ростом интеллектуальных вычислительных услуг, сценарии применения услуг инференса (inference) становятся все более обширными. Переход услуг инференса от однорежимного к многорежимному режиму и постоянная эволюция в направлении взаимодействия в реальном времени предъявляют более строгие требования к сети. С одной стороны, сеть должна характеризоваться детерминированной низкой задержкой, чтобы

обеспечить представление результатов инференции в реальном времени и без задержек, избегая зависания кадров и задержек. С другой стороны, детерминированные услуги пропускной способности также незаменимы. Они могут обеспечить стабильную пропускную способность сети и бесперебойную передачу данных при передаче больших объемов данных без перегрузок. Таким образом, сеть должна предоставлять детерминированные сетевые услуги.

Детерминированная сетевая технология позволяет разделить сетевые ресурсы на различные логические сети и предоставить независимые логические сети для различных услуг с целью реализации дифференцированных услуг. Благодаря технологиям SRv6 и Flex-E он может более гибко планировать пути передачи данных и оптимизировать распределение сетевого трафика с помощью SRv6 с сетевым контроллером, реализуется вычисление пути с низкой задержкой, что эффективно обеспечивает задержку обслуживания. Сетевой контроллер собирает информацию о топологии сети и состоянии каналов связи в режиме реального времени. На основе собранной информации и функции программируемого маршрута SRv6 сетевой контроллер вычисляет оптимальный маршрут, который соответствует требованиям к задержке для услуг. Когда данные службы поступают в сеть, путь SRv6 перенаправляется по запланированному пути SRv6, чтобы избежать перегруженных узлов и каналов связи, что сокращает задержку передачи. Кроме того, благодаря механизму резервирования пропускной способности Flex-E выделяются выделенные ресурсы пропускной способности для служб, что гарантирует удовлетворение требований к пропускной способности конкретных служб даже при перегрузке или занятости сети, предотвращая ухудшение производительности служб из-за конфликтов пропускной способности и обеспечивая пропускную способность и качество обслуживания даже при высокой нагрузке на сеть.

4.4 Интеллектуальные функции эксплуатации и технического обслуживания, безопасность и надежность

4.4.1 Интеллектуальные возможности эксплуатации и технического обслуживания

С быстрым развитием таких технологий, как 5G, Интернет вещей (IoT) и пограничные вычисления, городские сети сталкиваются с такими проблемами, как всплеск трафика, диверсификация услуг и строгие требования к качеству обслуживания. Традиционный режим эксплуатации и технического обслуживания, основанный на ручных правилах и статических политиках, не может удовлетворить требованиям к сети в эпоху искусственного интеллекта, которые включают в себя работу в режиме реального времени, надежность и гибкость. Поэтому для городских сетей срочно требуется интеллектуальная система эксплуатации и технического обслуживания. TMF определяет автономность сети как шесть уровней (L0–L5). Суть интеллектуальной системы эксплуатации и технического обслуживания для городских сетей в эпоху искусственного интеллекта заключается в создании нового сетевого «мозга» искусственного интеллекта, который обладает функциями самоощущения, самоанализа, самопринятия решений и саморазвертывания, помогает уровню автономности сети эволюционировать от условной автономности L3 до расширенной автономности L4 и, наконец, достигает цели полной интеллектуальной автономности L5.

Интеллектуальная система эксплуатации и технического обслуживания городских сетей построена на основе нескольких ключевых технологий. Во-первых, распределенные датчики и встроенные чипы искусственного интеллекта развернуты для реализации многомерного осознания состояния сети в реальном времени. Во-вторых, интеллектуальный аналитический движок, построенный на основе глубокого обучения, обрабатывает огромные объемы данных O&M в режиме реального времени. Наконец, контроллер SDN и система автоматической

оркестрации используются для доставки и корректировки политик за считанные секунды. Ниже приведены ключевые интеллектуальные возможности эксплуатации и технического обслуживания:

- Высокоточное моделирование: создается сеть цифрового зеркалирования в режиме реального времени для реализации многоуровневого визуализированного моделирования физической топологии, маршрутов, туннелей, VPN и потоков. Система автоматически синхронизирует текущие конфигурации сети в реальном времени, маршруты BGP и характеристики трафика, создает эталонную модель сети зеркалирования и создает систему предварительной оценки для изменений конфигурации на основе технологии цифровых двойников. При изменении конфигурации сети система автоматически генерирует новую сеть зеркалирования, сравнивает и анализирует состояние топологии, распределяет трафик и эффективность сходимости маршрутов до и после изменения, а также предоставляет отчет об оценке воздействия для эффективного выявления потенциальных высокорисковых ошибок конфигурации. Кроме того, благодаря технологии динамического моделирования трафика система может моделировать политики маршрутизации и изменения трафика за миллисекунды, точно прогнозировать динамику ключевых показателей производительности, таких как флуктуация задержки и пороговое значение потери пакетов, предоставляя данные для принятия решений по оптимизации сети.
- ИИ-диагностика: Многомерная модель самодиагностики неисправностей создается на основе извлечения признаков неисправностей второго уровня на стороне устройства, вывода графа знаний и анализа закономерностей временных рядов. Система использует технологию цепочки больших моделей для реализации интеллектуального агрегирования аварийных сигналов и

прогнозирования тенденций состояния, а также поддерживает динамическое обоснование первопричин неисправностей и выявление потенциальных рисков. Механизм онлайн-внедрения знаний позволяет системе выполнять направленную диагностику неизвестных неисправностей и генерировать рекомендации по их устранению в замкнутом цикле, формируя комплексное интеллектуальное решение для эксплуатации и технического обслуживания.

– Самовосстанавливающаяся сеть: Комплексный механизм устранения неисправностей в замкнутом цикле создается на основе базы данных знаний по эксплуатации и техническому обслуживанию и возможностей динамической оркестровки больших моделей, реализуя автоматическую обработку на протяжении всего процесса «обнаружение-диагностика-принятие решения-исполнение». В случае неаппаратных неисправностей система автоматически реализует политики восстановления, такие как переключение на резервный путь. В случае аппаратных неисправностей система формирует точные заказы на техническое обслуживание на основе моделирования цифрового двойника.

4.4.2 Изоляция сетевого фрагмента на уровне арендатора

Городские сети должны предоставлять традиционные услуги и интеллектуальные вычислительные сервисы унифицированным образом и соответствовать дифференцированным требованиям SLA в различных сценариях обслуживания. Механизм двойной изоляции логических и физических ресурсов эффективно предотвращает вытеснение ресурсов и обеспечивает детерминированную доступность ключевых показателей, таких как пропускная способность и задержка, для сервисов обучения и вывода. Являясь новым сетевым решением на основе IPv6, технология изоляции сетевого сегмента на уровне арендатора в полной мере использует программируемость SRv6 и преимущества

адресного пространства IPv6 для предоставления нескольким арендаторам сетевых сегментов, которые совместно используют физические ресурсы, но логически изолированы. Основной механизм этой технологии заключается в следующем. Узел-источник инкапсулирует уникальный идентификатор сегмента в соответствии с требованиями арендатора, а узлы на пути реализуют идентификацию сегмента путем анализа пакета и применения предопределенной политики пересылки.

Технология сегментации сети на уровне арендаторов обладает тремя основными преимуществами: во-первых, идентификаторы сегментов используются для представления детальных ресурсов, гарантируя, что такие показатели, как пропускная способность и задержка между сегментами, не будут мешать друг другу. Во-вторых, программируемость сети SRv6 поддерживает гибкую оркестровку сервисов, отвечая требованиям быстрого развертывания сервисов. В-третьих, она обеспечивает высоконадежное решение для сегментации сети, позволяющее добиться оптимального использования ресурсов и точного обеспечения сервисов в многоарендной среде.

4.4.3 Обеспечение сквозной безопасности

В эпоху искусственного интеллекта для городских вычислительных сетей требуется многоуровневая система глубокой защиты, чтобы противостоять утечкам данных и рискам горизонтального проникновения в многопользовательских средах. Её суть заключается в реализации сквозной изоляции данных арендаторов и шифрованной передачи данных, особенно в сценариях планирования вычислительной мощности и междоменного обмена данными, для обеспечения конфиденциальности и целостности данных. На основе VPN SRv6 и технологии сетевой нарезки можно построить трёхуровневый механизм изоляции «устройство доступа – сетевой нарезка – VPN» для эффективного блокирования угроз безопасности путём разделения физического уровня, уровня протокола и уровня сервисов по всем измерениям. Кроме того, для

достижения нулевого уровня перекрестного проникновения данных арендаторов используются групповые политики безопасности и управление списками доверия междоменного трафика.

Архитектура безопасности использует механизмы двойной защиты: изоляцию слайсов и VPN-шифрование, что позволяет повысить уровень безопасности с пассивной защиты до активного иммунитета, достигая цели безопасности: «данные не фрагментируются, риски не пересекаются, и трафик открытого текста не остаётся». В будущем передовые технологии, такие как квантовое шифрование (включая постквантовую криптографию и квантовое распределение ключей) и доверенная среда выполнения, ещё больше повысят возможности защиты сети. Конвергенция этих технологий будет способствовать развитию интеллектуальных вычислительных сетей в сторону архитектуры нулевого доверия, характеризующейся «активным иммунитетом, динамической осведомлённостью и надёжностью всей цепочки», обеспечивая надёжную основу безопасности для инновационных сервисов в эпоху ИИ.

4.4.4 Экологичные и низкоуглеродные сети

Как ключевая инфраструктура, поддерживающая развитие искусственного интеллекта и цифровой экономики, городские сети сталкиваются с проблемами высокого энергопотребления и низкой эффективности. По мере экспоненциального роста требований к вычислительной мощности и продолжающегося роста нагрузки на пропускную способность сети, такие проблемы, как высокое энергопотребление на один бит и резкое увеличение затрат на тепловыделение существующей платформы 100G, становятся всё более заметными. В результате одновременно увеличиваются эксплуатационные расходы и выбросы углерода. Кроме того, многослойная архитектура сети приводит к избыточности устройств и потерям при преобразовании протоколов, что ещё больше усугубляет проблему энергоэффективности. В рамках стратегии «двойного углеродного следа» крайне важно использовать технологические

инновации для революционного повышения энергоэффективности сети и создания городских сетей с высокой энергоэффективностью, большой пропускной способностью и функциями интеллектуального планирования.

Экологичная и низкоуглеродная трансформация городских сетей сосредоточена на трёх технических направлениях с точки зрения модернизации сверхскоростной платформы, сети 400G/800G могут значительно снизить энергопотребление на единицу и поддерживать передачу данных без потерь на большие расстояния, отвечая требованиям сценариев с высокой пропускной способностью, таких как обучение крупных моделей ИИ. С точки зрения интеллектуальной системы энергосбережения, основанные на ИИ алгоритмы прогнозирования нагрузки в реальном времени и многофакторного принятия решений реализуют динамическую оптимизацию и регулировку мощности устройств, политик теплоотвода и состояния оптических модулей. С точки зрения реконструкции архитектуры, SRv6 и EVPN используются для упрощения сетевых уровней, продвижения плоской архитектуры и использования SDN для реализации точного планирования ресурсов. Благодаря совместным инновациям в области сверхскоростных платформ, интеллектуального управления и упрощенной архитектуры мы создадим новую вычислительную сеть с высокой энергоэффективностью и низким уровнем выбросов, обеспечивая поддержку «зелёной» инфраструктуры для развития высококачественной цифровой экономики.

Глава V Типичные сценарии развертывания

5.1 Сценарий 1: Передача больших выборочных данных в AIDC

Характеристики сценария: все три этапа разработки большой модели ИИ — предварительная подготовка, постобучение и тонкая настройка требуют передачи больших объёмов выборочных данных в AIDC. На этапе предобучения объёмы данных достигают петабайтного масштаба. Хотя объём выборочных данных на пользователя на этапах постобучения и тонкой настройки относительно меньше (обычно на уровне ГБ/ТБ), совокупный объём данных значительно возрастает по мере увеличения числа пользователей. Следовательно, городские сети должны соответствовать требованиям сверхвысокой пропускной способности сценариев доставки обучающих данных и обладать возможностями слайсинга на уровне арендаторов для обеспечения безопасной изоляции между различными арендаторами.

Решение:

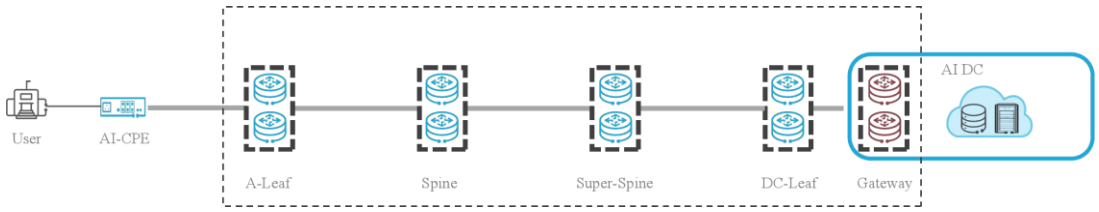

Рисунок 5-1: Передача больших выборочных данных в решение AIDC

Для поддержки сценариев доставки больших выборок данных городской сети требуются ключевые технические возможности, такие как изоляция на уровне арендаторов (tenant-level slice isolation) и балансировка нагрузки на уровне сети:

- Изоляция на уровне арендаторов (tenant-level slice isolation). Изолирует трафик от обучающих данных от обычного сервисного трафика с помощью технологии иерархического среза (slicing), эффективно предотвращая конкуренцию за ресурсы между арендаторами;
- Балансировка нагрузки на уровне сети. Реализует бесконфликтное сбалансированное планирование по всем сетевым маршрутам

посредством унифицированного планирования трафика, значительно повышая эффективность использования сетевых ресурсов.

5.2 Сценарий 2: Обучение модели с разделением хранения и вычислений

Характеристики сценария: такие отрасли, как финансы и здравоохранение, предъявляют чрезвычайно высокие требования к безопасности конфиденциальных данных. При аренде сторонних AIDC для обучения больших моделей требуется, чтобы конфиденциальные данные не хранились на сторонних AIDC. Поэтому в сценарии удаленного обучения узлы хранения данных и AIDC развертываются в глобальных сетях. Данные выборки извлекаются по запросу для обучения и сразу же удаляются после вычислений, что эффективно отвечает требованиям безопасности данных клиентов, работающих с конфиденциальными данными. Городские сети должны соответствовать требованиям RDMA к передаче без потерь в рамках этого сценария удаленного обучения и обладать такими возможностями, как сегментация сети на уровне арендатора и шифрование данных, чтобы гарантировать сохранность данных выборки во время передачи.

Решение:

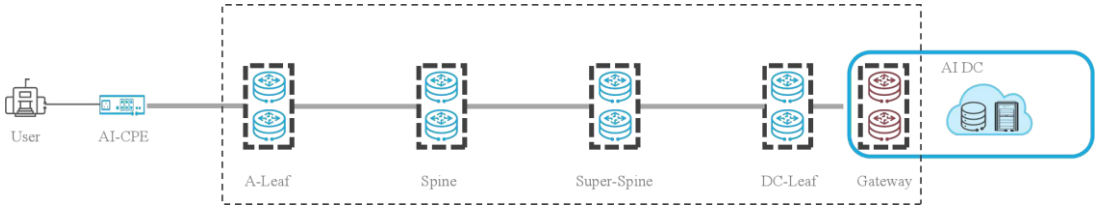

Рисунок 5-2: Обучение модели с использованием решения, дезагрегированного по хранению и вычислениям

Для поддержки сценария удаленного обучения городских сетей необходимо обладать ключевыми техническими возможностями, такими как изоляция срезов (slicing) на уровне арендаторов и передача данных без потерь по широкополосному RDMA:

- изоляция срезов (slicing) на уровне арендаторов поддерживает изоляцию трафика удаленного обучения от трафика обычных

сервисов, предотвращая перегрузку трафика при контроле перегрузки, влияющую на другие сервисы;

- RDMA без потерь по широкополосному RDMA благодаря точному управлению потоком на уровне потока предотвращается потеря пакетов во время синхронизации выборки с обучением, что гарантирует сохранение вычислительной эффективности во время удаленного обучения;
- балансировка нагрузки Elephant Flow на основе информации транспортного уровня разделяет трафик и распределяет нагрузку между несколькими подпотоками одного Elephant Flow по разным маршрутам слайсов, обеспечивая высокую пропускную способность;
- шифрование данных поддерживает сквозную шифрованную передачу данных, гарантируя безопасность данных выборки во время передачи.

5.3 Сценарий 3: Совместное обучение модели на нескольких AIDC

Характеристики сценария: во время совместного обучения большой модели на нескольких географически разнесённых AIDC промежуточные данные, генерируемые на каждой итерации обучения (параметры оптимизатора, градиенты и т. д.), должны быть синхронизированы между всеми AIDC перед переходом к следующей итерации. Этот цикл повторяется до завершения обучения. Синхронизация данных на уровне параметров основана на RDMA, который очень чувствителен к потере пакетов, при этом объёмы одновременных данных достигают терабайт. Следовательно, городские сети должны обеспечивать сверхвысокую пропускную способность и возможность передачи данных без потерь, одновременно используя высококонвергентные сети для баланса затрат на полосу пропускания с эффективностью вычислений при обучении.

Решение:

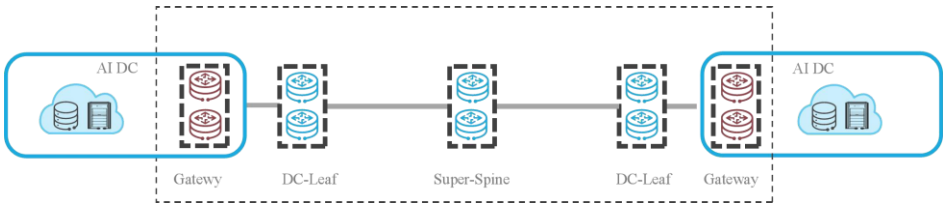

Рисунок 5-3: Совместное обучение модели в рамках нескольких решений AIDCs

Решения для городских сетей должны поддерживать следующие ключевые технологии:

- балансировка нагрузки на уровне сети обеспечивает бесконфликтное планирование межпутевого потока по всей сети благодаря унифицированному планированию трафика;
- RDMA без потерь в глобальных сетях: предотвращает потерю пакетов во время распределенного обучения благодаря точному управлению потоком на уровне каждого потока, гарантируя отсутствие снижения вычислительной эффективности;
- сетевые технологии с высокой степенью конвергенции: реализует эффективную конвергенцию трафика совместного обучения посредством коллективных алгоритмов связи и оптимизации сети, снижая затраты на развертывание сетевой инфраструктуры.

5.4 Сценарий 4: Совместное обучение/вывод модели в облаке и на периферии сети

Характеристики сценария: локальный подход к развертыванию интегрированных машин для обучения и вывода в корпоративном парке не позволяет удовлетворить быстро растущие потребности предприятий в тонкой настройке моделей и выводе. Таким образом, совместное обучение/вывод интегрированных машин и пулов вычислительных ресурсов в облаке и на периферии сети стало критически важным направлением для обеспечения гибкого масштабирования корпоративных вычислительных ресурсов, тем самым поддерживая развертывание крупных приложений с моделями. Совместная

работа в облаке и на периферии сети основана на разбиении модели на разделы, что требует от городской сети поддержки синхронизации данных межуровневой плоскости параметров. Это требует возможности передачи данных RDMA без потерь со сверхвысокой пропускной способностью.

Решение:

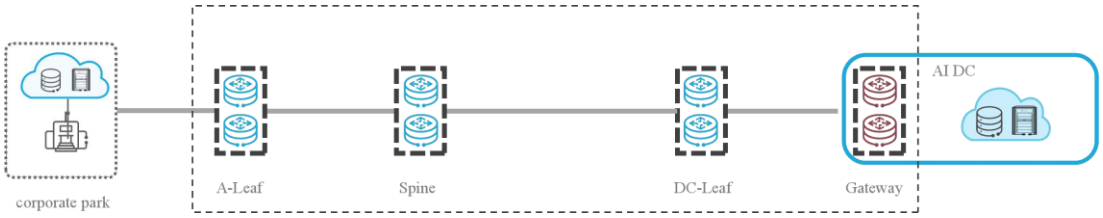

Рисунок 5-4：Решение для совместного обучения/вывода модели в облаке и на периферии

Решения для городских сетей должны поддерживать следующие ключевые технологии:

- балансировка нагрузки на уровне сети обеспечивает бесконфликтное сбалансированное планирование по всем маршрутам во всей сети благодаря унифицированному планированию трафика в масштабах всей сети;
- глобальная сеть RDMA без потерь (Wide-Area Lossless Networking) предотвращает потерю пакетов во время обучения модели благодаря точному управлению трафиком на уровне потока, гарантируя неизменно высокую вычислительную эффективность в ходе совместного обучения и вывода.

5.5 Сценарий 5: Доставка инференса

Характеристики сценария: предварительно обученные большие модели обычно имеют размер порядка гигабайта (ГБ). Во время развертывания их необходимо распределить между кластерами обучения и несколькими кластерами вывода. Городские сети должны обеспечивать сверхвысокую пропускную способность для обеспечения эффективности передачи данных во

время распределения, а также надежные механизмы безопасности для сохранения целостности модели. Более того, после развертывания моделей вывода на периферийных узлах они должны быстро реагировать на высококонкурентные запросы пользователей на вывод в режиме реального времени, что требует наличия детерминированных возможностей обслуживания в городских вычислительных сетях.

Решение:

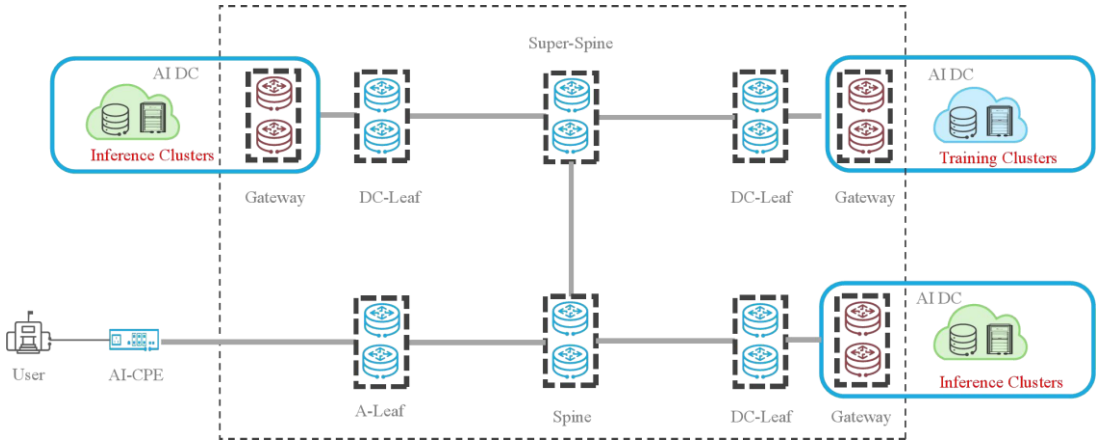

Рисунок 5-5: Решение для вывода результатов

Решения для городских сетей должны поддерживать следующие ключевые технологии:

- балансировка нагрузки на уровне сети благодаря оркестровке трафика в масштабах всей сети достигается сбалансированное планирование без конфликтов по всем маршрутам во время развертывания модели вывода;
- шифрование безопасности, комбинированные многоуровневые технологии шифрования обеспечивают безопасность данных при передаче;
- детерминированная низкая задержка обеспечивает оптимального пользовательского опыта при взаимодействии с приложениями вывода в режиме реального времени.

5.6 Сценарий 6: Федеративное обучение

Характеристики сценария: в процессе федеративного обучения нескольких AIDC каждый участник обучает модели локально, используя данные из частного домена. Для агрегации параметров модели происходит обмен градиентами параметров модели. Городские сети должны обеспечивать стабильное подключение устройств, участвующих в федеративном обучении, обеспечивая при этом безопасность передачи данных.

Решение:

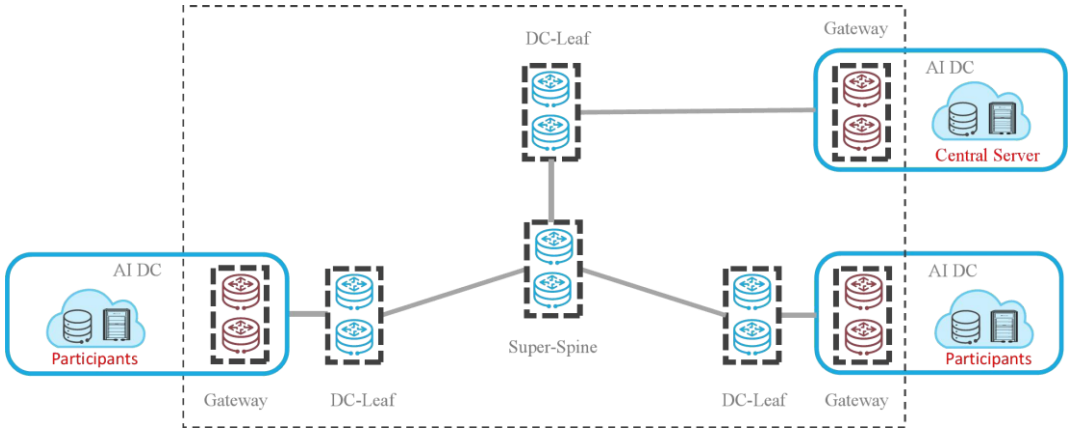

Рисунок 6-6: Решение для федеративного обучения Решения

Решения для городских сетей должны поддерживать следующие ключевые технологии:

- безопасное шифрование обеспечивается благодаря сочетанию технологий многоуровневого шифрования и гарантируется безопасность процесса передачи данных;
- глобальная сеть RDMA без потерь предотвращает потерю пакетов во время федеративного обучения благодаря точному управлению трафиком на уровне потока, гарантируя неизменность вычислительной эффективности на протяжении всего федеративного обучения.

5.7 Сценарий 7: Мультиагентная система/A2A

Характеристики сценария: Мультиагентная система (MAS) реализует взаимодействие в режиме реального времени, динамическое взаимодействие задач и безопасную связь между агентами через протокол A2A (Agent-to-Agent), предъявляя явные и строгие требования к сети. Это означает динамическое делегирование задач между агентами A2A, что требует низкой сетевой задержки для предотвращения блокировки цепочки задач, A2A должна выполнять длительные задачи (например, углубленный анализ, длящийся от нескольких часов до нескольких дней), требующие поддержания стабильных постоянных соединений; Изоляция разрешений (например, Агент А может вызывать только определенные интерфейсы Агента Б) требует сетевой поддержки для детального контроля доступа.

Решение:

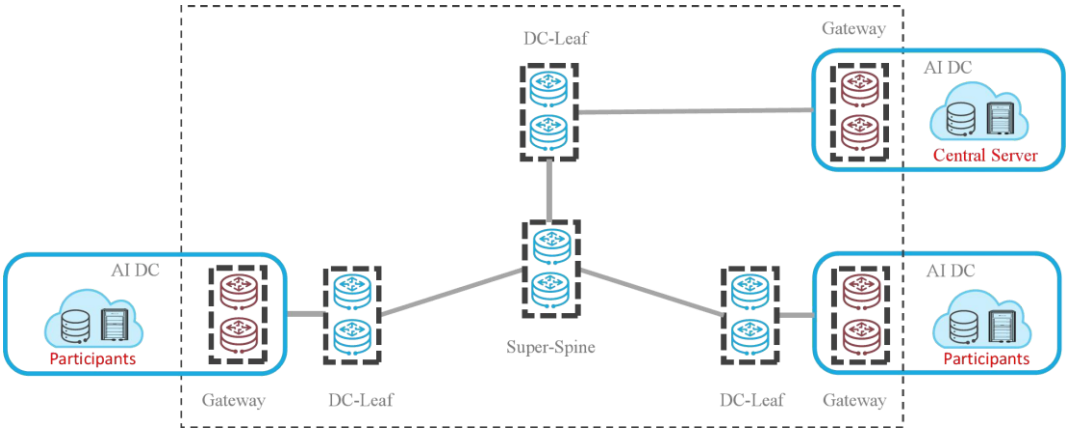

Рисунок 5-7: Мультиагентная система / решение A2A

Решения для городских сетей должны поддерживать следующие ключевые технологии:

- маршрут с малой задержкой для городской вычислительной сети обеспечивается гарантированная низкая задержка на уровне миллисекунд для делегирования задач между агентами;
- надежность сети для городской вычислительной сети обеспечивается за счет стабильных и надежных сетевых маршрутов путем длительной

передачи задач между агентами.

Глава VI Выводы и перспективы на будущее

В этом документе рассматриваются тенденции развития искусственного интеллекта и соответствующие требования к сервисам, проводится комплексное исследование сценариев применения, сетевых архитектур, ключевых технологий и решений по развертыванию городских сетей. Это активно способствует развитию традиционных городских сетей в сервисно-ориентированные вычислительные сети следующего поколения, тем самым способствуя технологическим инновациям и практическому развертыванию. Планирование и строительство городских сетей должны определяться как потребностями пользователей, так и достижениями в области технологий конвергенции вычислений и сетей. Благодаря исследованиям и анализу, представленным в этом документе, мы стремимся стимулировать более широкое участие отрасли и активизировать обсуждения. Мы рассчитываем на сотрудничество с партнерами по всей экосистеме для разработки городских сетей следующего поколения, характеризующихся всеобъемлющим покрытием, эластичной масштабируемостью, возможностью подключения к глобальным сетям без потерь, сверхвысокой надежностью и интеллектуальной автоматизацией.

媒体合作：contact@nida-alliance.org
工作机会：contact@nida-alliance.org
业务合作：contact@nida-alliance.org
官方网站：www.nida-alliance.org