РАЗДЕЛ 5 РАСПРЕДЕЛЕНИЕ РЕСУРСОВ СЕТИ ПРИ ПРЕДОСТАВЛЕНИИ УСЛУГ ДОПОЛНЕННОЙ РЕАЛЬНОСТИ

Услуги дополненной реальности являются очередным шагом развития услуг подвижной связи. Сочетание свойств мобильности терминала, его вычислительных возможностей, способов взаимодействия с окружающей средой (распознавание видео, звуковых и тактильных образов, вычисление координат и ориентации в пространстве), а также современной сети связи позволяют реализовать качественно новый уровень услуг, обладающих высокой степенью интерактивности [370]. В частности, это услуги дополненной реальности. Сегодня уже широко известны и популярны такие услуги как интерактивные карты городов и населенных пунктов, звездного неба, различного рода путеводители, приложения для заказа товаров и услуг. Однако, если представить какой объем трафика порождают в сети все эти приложения, то становится очевидным, что необходимо менять существующие структуры сети и механизмы распределения трафика в сети. При предоставлении услуг дополненной реальности часто можно наблюдать тесное взаимодействие с устройствами Интернета вещей, которые также генерируют довольно внушительный объем трафика. Ясно, что структура услуги, при которой обращение идет к некоторому серверу в сети, неприменима при таком количестве устройств. В данной главе предлагается иерархическая структура размещения данных в системе обслуживания дополненной реальности и исследуются преимущества использования данной структуры. Достаточно большое число приложений дополненной реальности основано на распознавании объектов в окружении пользователя, т.е. в поле его зрения. Таким образом, зная географические координаты пользователя, можно заранее подгружать на его терминал информацию об окружающих объектах, используя в том числе свободные терминалы других пользователей или другие устройства, находящиеся в области восприятия пользователя. Тем самым

снижается время реакции системы на изменения окружения пользователя, что существенно увеличивает своевременность отображения информации об объектах.

Далее в главе предлагается новая структура системы предоставления услуг дополненной реальности на основе модифицированной многоуровневой системы граничных вычислений, использующей технологию взаимодействия *D2D*. Для предложенной новой структуры разработан метод выгрузки трафика приложений дополненной реальности, который позволяет обеспечить энергетически эффективное функционирование сети при работе таких приложений как потоковое видео с обзором в 360^0, web-приложения, многопользовательские игры. Для предложенного метода проведено моделирование и оценка эффективности для различных сценариев.

Следует отметить, что вопрос идентификации объектов дополненной реальности и Интернета вещей, также влияет на качество предоставления услуг и структуру организации услуги. Преимущество и удобство для пользователей большинства приложений дополненной реальности заключается в возможности визуальной идентификации объектов. Т.е. пользователь в очках дополненной реальности смотрит на какой-нибудь объект и ему предоставляется информация об этом объекте или меняется его состояние, таким образом пользователь не производит никаких действий для получения информации и управления объектом. В данной главе предложена система идентификации устройств Интернета Вещей с использованием технологии дополненной реальности и облачных сервисов. Для разработанной системы создана модельная сеть, на базе которое проведена оценка работы системы и соответствие показателям качества восприятия.

4.1 Структура реализации услуги

Услуги дополненной реальности позволяют пользователю своевременно получать необходимую информацию. При этом ее выбор выполняется

автоматически на основании данных о его состоянии, например, положении в пространстве (географические координаты), на карте и плане территории (на основе данных геолокации), о нахождении транспортного средства и др. Как было показано в [330, 371], реализация услуги требует организации обмена данными с сервером услуги и/или непосредственно с устройствами, находящимися в зоне связи абонентского терминала, при использовании технологий *D2D* [189].

При этом время между запросом и доставкой данных не должно превышать некоторой величины, при которой пользователь еще не ощущает снижения качества услуги. Это время определяется временем: формирования запроса (зависит от реализации услуги), доставки запроса от терминала до сервера услуги, обработки запроса, доставки данных от сервера услуги до терминала и представления информации пользователю. Их условно можно разделить на три группы: время, определяемое обработкой данных терминалом пользователя, время доставки данных по сети связи и время обработки данных сервером. В общем случае эти составляющие взаимно зависимы.

Существенную роль играет процесс формирования запроса данных. Запрос формируется при изменении окружения пользователя (или состояния пользователя) о чем можно судить по изменению некоторых параметров. Такими параметрами могут быть данные датчиков, например, географические координаты, положение терминала в пространстве, ускорение, а также результаты анализа изображения или звука, получаемого от камер и микрофонов терминала. Например, если запрос данных формируется по результатам распознавания образа (видео, снятого камерой терминала), то функции распознавания образов могут быть реализованы или в приложении терминала, или на сервере услуги. В первом случае при низкой вычислительной производительности терминала время будет расходоваться на выполнение функций распознавания терминалом, во втором – на передачу видео через сеть связи и время его обработки сервером.

Очевидно, что выбор первого или второго варианта зависит от производительности терминала, пропускной способности (ПС) сети связи, производительности и загрузки сервера, т.е. имеет место задача выбора оптимального варианта реализации услуги. Описанную здесь модель можно расширять, введя дополнительные параметры, например, зависимость времени обработки запроса сервером услуги от объема данных (размера базы данных) и интенсивности запросов. В таком случае имеет смысл кластеризация данных и организация локальных серверов услуги.

Рассматривая перспективную сеть *5G*, технологию связи *D2D*, а также применение *SDN* (*Software-Defined Network*, программно-конфигурируемые сети), можно представить структуру реализации услуги, приведенную на рисунке 4.1.

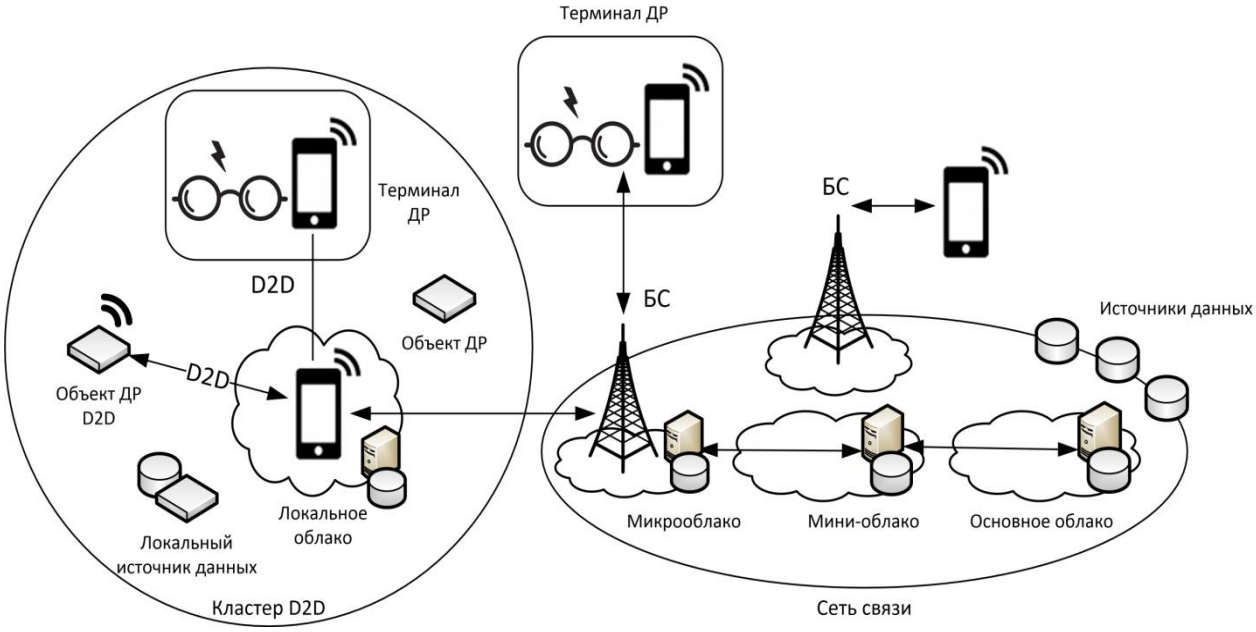

Рисунок 4.1 - Возможная структура реализации услуги ДР

Полагаем, что сеть связи построена с использованием архитектуры *SDN*, в которой присутствуют центры обработки данных (ЦОД) различных уровней [130], что дает возможность локализовать трафик и данные «ближе» к пользователям. На схеме эти ЦОД изображены как облака микро, мини и основного уровня. В реальной сети таких уровней может быть столько,

сколько будет необходимо для наилучшей реализации услуги. Базовая станция сети взаимодействует непосредственно с терминалом ДР или с мобильным терминалом, выполняющим роль локального облака, взаимодействующего с терминалом ДР с помощью технологий *D2D*, что повышает эффективность использования радиочастотного спектра [189]. Здесь под облаком понимается некий объем вычислительных ресурсов и ресурсов памяти, который может быть применен для организации сервера и базы данных (БД) услуги.

Как будет показано ниже, предоставление услуги может быть реализовано на нескольких уровнях таких серверов и БД, что позволяет за счет локализации данных и трафика снизить требования к ПС сети и повысить показатели качества представления услуги. Ниже задача кластеризации данных и локализации их обработки рассматривается как задача распределения ресурсов.

4.1.1 Модель услуги

Для построения модели услуги необходимо связать показатели (параметры), характеризующие качество ее предоставления, с параметрами системы связи. В качестве основного показателя выберем время реакции на изменение окружения пользователя - τ. Будем полагать, что это время включает все составляющие: время распознавания изменения и предварительной обработки приложением мобильного терминала t_r, время передачи данных (запроса) на сервер услуги через сеть связи t_q, время обработки запроса сервером услуги t_s, время доставки данных через сеть связи t_a и время представления информации пользователю приложением мобильного терминала t_d. Модель представления услуги дополненной реальности представлена на рисунке 4.2.

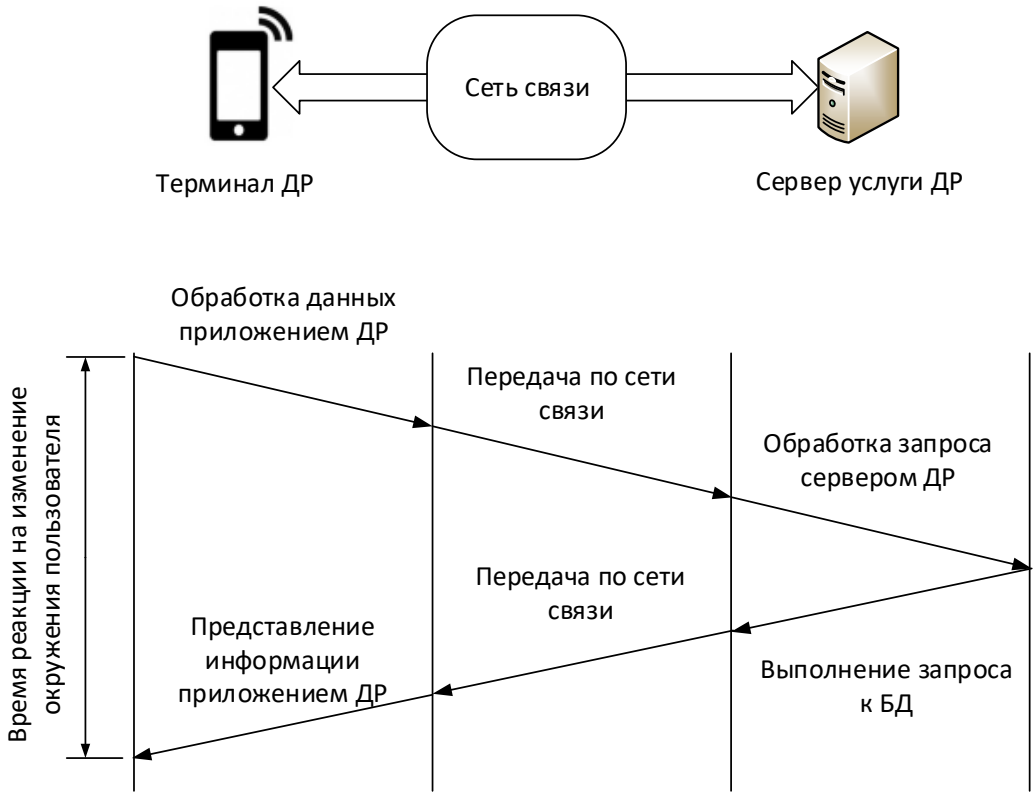

Рисунок 4.2 - Модель предоставления услуги ДР

Общее время можно представить арифметической суммой всех составляющих. Будем полагать, что каждая из них является случайной величиной. Тогда, сделав допущение об их независимости, среднее значение для времени реакции будет определяться как

$$\bar{\tau} = \bar{t}_r + \bar{t}_q + \bar{t}_s + \bar{t}_a + \bar{t}_d \,. \qquad (4.1)$$

Рассмотрим отдельно каждую из составляющих. Время распознавания изменения окружения пользователя t_r, в свою очередь, включает все составляющие, связанные с обнаружением этого изменения и сбором информации, необходимой для формирования запроса, направляемого на сервер услуги. Обнаружение изменения может быть реализовано путем анализа данных от различных датчиков и устройств (датчики магнитного поля земли, освещенности, ускорения, приемника сигналов глобальных систем позиционирования, сенсорного экрана и др.), а также видеокамер и микрофонов. Анализ может включать, как относительно простые задачи сравнения нескольких численных значений, так и ресурсоемкие задачи

распознавания образов. Поэтому численное значение t_r зависит от вида услуги, способа ее реализации и вычислительных ресурсов мобильного терминала.

Таким образом, ресурсы мобильного терминала влияют на качество услуги через величину t_r. Будем полагать, что существует некая функциональная зависимость между этим временем и вычислительными ресурсами мобильного терминала

$$\bar{t}_r = f_r(O), \qquad (4.2)$$

где O – параметр, характеризующий производительность мобильного терминала, например, количество выполняемых в секунду операций или команд, тактовая частота процессора, объем памяти или некоторый комплексный показатель.

Время передачи запроса на сервер услуги t_q определяется объемом передаваемых данных и ПС маршрута между мобильным терминалом и сервером услуги C. Численная оценка этого времени при допущении, что время тратится только на передачу данных (ПД), т.е. без учета потерь на ожидание передачи в узлах маршрута может быть получена как

$$\bar{t}_q = f_q(C) = \frac{\bar{v}_q}{C}, \qquad (3)$$

где C –ПС маршрута (бит/с); $\bar{v}_q$ – средний объем данных, передаваемых в запросе (бит).

Средний объем данных в запросе $\bar{v}_q$ зависит от вида услуги и способа ее реализации. Например, если для идентификации изменения окружения требуется анализ изображений, получаемых от видеокамеры устройства, то этот анализ может быть выполнен как средствами приложения мобильного терминала, так и сервером. В первом случае запрос будет содержать относительно мало данных, которые являются лишь идентификаторами объектов в БД услуги, информацию о которых требуется предоставить. Во втором случае необходимо передать все данные изображения (или нескольких изображений), анализ которых необходимо провести средствами сервера

услуги. Возможны и промежуточные варианты, когда на сервер будет отправляться лишь часть видеоданных.

Время обработки запроса сервером t_s является наиболее сложной характеристикой, поскольку зависит от многих параметров: времени анализа данных запроса τ_s, поступающих данных; интенсивности запросов от мобильных терминалов пользователей λ_s; производительности сервера μ_s, которая, в свою очередь, зависит от размера БД n_s:

$$\bar{t}_s = f_s\left(\tau_s, \lambda_s, \mu_s\left(n_s\right)\right). \tag{4.4}$$

Сервер обслуживает запросы, поступающие от множества пользователей, время обслуживания которых определяется временем обработки запроса и временем ожидания. Сервер может быть описан моделью СМО, в которой время обслуживания запроса определяется размером БД услуги и производительностью сервера. Техническая реализация сервера может быть различна, поэтому имеет смысл рассматривать его как СМО с ожиданием общего вида с одним обслуживающим устройством G/G/1. При допущении, что поток входящих запросов можно представить моделью простейшего потока (M/G/1), средняя задержка может быть описана формулой Полячека-Хинчина [363, 372]. С учетом этого

$$\bar{t}_s = \frac{\rho_s}{\mu_s\left(v_s\right)2\left(1-\rho_s\right)}\left(1+V_s^2\right)+\frac{1}{\mu_s\left(v_s\right)}, \tag{4.5}$$

где $\rho_s = \dfrac{\lambda_s}{\mu_s\left(n_s\right)}$ – нагрузка на сервер;

$V = \sigma_s \mu_s\left(v_s\right)$ – коэффициент вариации времени обслуживания;

σ_s^2 – среднеквадратическое отклонение времени обслуживания.

Принятие модели простейшего потока весьма полезно, так как дает возможность получить аналитические выражения для зависимостей, особенно при неизвестных свойствах реального потока. Производительность сервера, зависящая от размера БД $\mu_s\left(n_s\right)$, также представляет собой некую зависимость и определяется способом реализации БД. В частности, наиболее

распространенные модели описывают эту зависимость как $ln(n_s)$ или $n\ ln(n_s)$ операций [373], где n_s – количество записей в БД.

Выберем в качестве примера логарифмическую зависимость. С учетом того, что время обслуживания включает предварительную обработку запроса, среднее время выполнения которой τ_s, получим

$$\mu_s(v_s) = \frac{1}{\eta \ln(v_s) + \tau_s}, \tag{4.6}$$

где η – время выполнения из расчета на одну запись.

Время передачи ответа сервера t_a, как и время передачи запроса, определяется объемом передаваемых данных и ПС маршрута между сервером услуги и мобильным терминалом *C*. Численная оценка этого времени при аналогичных допущениях может быть записана как

$$\bar{t}_a = f_a(C) = \frac{\bar{v}_a}{C}, \tag{4.7}$$

где *C* – ПС маршрута (бит/с);

$\bar{v}_a$ – средний объем данных, передаваемых в ответе сервера (бит). Средний объем данных в ответе сервера $\bar{v}_a$ зависит от вида услуги и способа ее реализации. Этими данными может передаваться текст, растровые или векторные изображения, звук, численные значения.

Время представления сообщения t_d включает все составляющие, связанные с обработкой и представлением принятых приложением мобильного терминала данных. В общем случае сообщение может быть представлено визуально: в виде текста, пиктограммы, видео или иного изображения; звука – речи или мелодии; тактильно – вибрации. Будем полагать, что существует некая функциональная зависимость между временем и вычислительными ресурсами мобильного терминала:

$$\bar{t}_d = f_d(O). \tag{4.8}$$

Как видно из выбранных выше моделей, время реакции существенно зависит от таких параметров как производительность мобильного терминала,

ПС сети связи и времени обработки запроса сервером, определяемое его производительностью и загрузкой.

Ниже предлагается метод выбора структуры и параметров оборудования для обеспечения требований ко времени реакции для услуги ДР.

4.2 Метод выбора структуры сети и параметров оборудования

Учитывая приведенные выше модели, можно сказать, что обеспечение приемлемого времени реакции представляет собой задачу выбора объема ресурсов (пропускной способности, производительности и памяти), а также их распределения между элементами системы обслуживания. Это задача с несколькими переменными, количество которых и определено самими моделями. Если рассуждать с позиции построения метода организации услуги, то не все переменные могут быть доступны для изменения. Например, если полагать, что производительность мобильного терминала и характеристики оборудования сервера можно учесть, а изменить в рамках данной задачи нельзя, то выражение для времени реакции может быть записано как

$$\bar{\tau} = \bar{t}_m + \frac{\bar{v}_q + \bar{v}_a}{C} + \frac{\rho_s}{\mu_s(n_s)2(1-\rho_s)}\left(1+V_s^2\right) + \frac{1}{\mu_s(v_s)}, \qquad (4.9)$$

где $\bar{t}_m = \bar{t}_r + \bar{t}_a$ – суммарная задержка, вносимая приложением мобильного терминала при обработке входных данных и отображении информации. Эта величина зависит от производительности терминала и особенностей приложения. Будем полагать, что она постоянна, т.е.

$$\bar{\tau} = \bar{t}_m + \frac{\bar{v}_q + \bar{v}_a}{C} + \frac{\lambda_s}{\mu_s(n_s)2(\mu_s(n_s)-\lambda_s)}\left(1+V_s^2\right). \qquad (4.10)$$

На рисунке 4.3 приведена зависимость, полученная с помощью (4.10) для разного количества записей в БД. Из рисунка видно, что время реакции увеличивается согласно закону 1/(*a*-*x*). Постоянная составляющая времени обусловлена временем ПД запроса на сервер и временем его обработки

сервером. Согласно выбранной модели (4.6), время обслуживания также зависит от количества записей в БД.

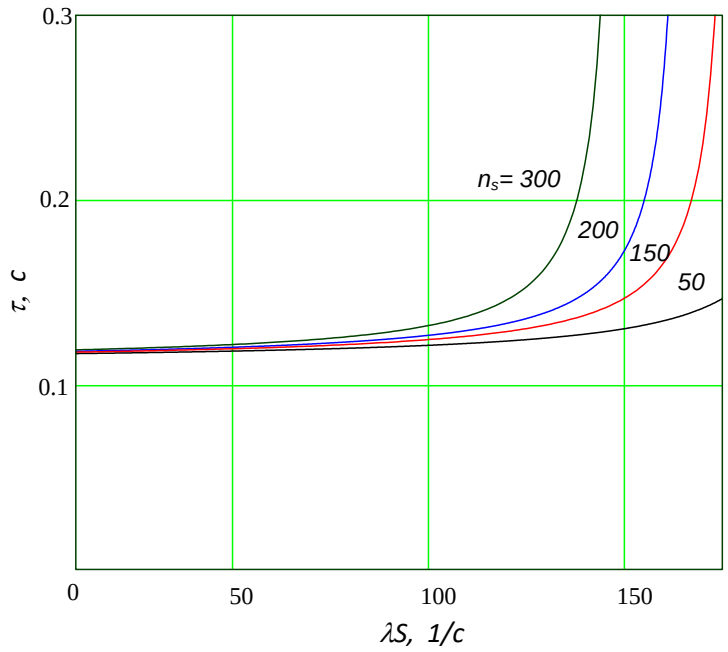

Рисунок 4.3 - Зависимость времени реакции от интенсивности запросов и размера БД

Таким образом, для организации услуги необходимо обеспечить требуемое время реакции. Для этого, исходя из представленных моделей (4.9) и (4.10), нужно выбрать структуру сети предоставления услуги с учетом трафика, объема данных БД и времени доставки данных.

4.2.1 Обработка данных об окружении

Как было отмечено выше, формирование запроса производится на основе результатов распознавания об изменении окружения. Распознавание может быть основано на различных данных, полученных как с датчиков, так и с видеокамер и микрофонов. В последнем случае задача распознавания может иметь значительную вычислительную сложность, следовательно, и затраты

времени будут весьма существенными. Поэтому имеет смысл выбрать средство решения этой задачи: мобильный терминал или сервер услуги.

Если изображение обрабатывается на терминале полностью, то на сервер передается запрос, содержащий лишь относительно малый объем данных, необходимый для идентификации объекта в БД сервера. В случае, когда изображение полностью поступает на сервер для обработки, требуется передача относительно большого объема данных, определяемого разрешающей способностью камеры и форматом представления данных. Возможны и промежуточные решения, например, когда приложение терминала не производит полной идентификации объекта, но выделяет объект (полезных данных) в изображении. В этом случае на сервер поступает лишь выделенная часть (части) изображения для дальнейшей обработки.

Таким образом, перенос работы по распознаванию объектов в терминал пользователя экономит время передачи данных по каналу и ресурс ПС. Перенос этой работы на сервер позволяет экономить на времени обработки изображения терминалом, однако приводит к росту задержки на передачу и расход ПС канала. С точки зрения использования сетевых ресурсов первый вариант выгоднее, однако, при реализации услуги следует учитывать реальную производительность мобильного терминала, реальную ПС канала и требования к времени реакции.

Сказанное выше опишем следующей моделью. Составляющая задержки, обусловленная обработкой изображения в терминале, сервере и временем ПД на сервер, определяется как

$$\tau_P = f_r(O_r) + \frac{\bar{v}}{C} + f_S(O_S), \qquad (4.11)$$

где O_r и O_s – время обработки на мобильном терминале и сервере из расчета на один бит, соответственно;

C – скорость ПД (бит/с);

$\bar{v}$ – объем обрабатываемых (передаваемых) данных.

Будем полагать, что время обработки в мобильном терминале и сервере линейно зависят от размера обрабатываемого блока данных (изображения или его части)

$$f(O) = \bar{v}O, \tag{4.12}$$

где O – время обработки из расчета на 1 байт изображения, характерное для мобильного терминала или сервера (для мобильного терминала и сервера эти значения могут существенно отличаться).

Тогда

$$\tau_P = \eta_r \frac{\bar{v}}{O_r} + (1-\eta_r)\left(\frac{\bar{v}}{C} + \frac{\bar{v}}{O_S}\right), \tag{4.13}$$

где η_r – доля данных, обрабатываемых в мобильном терминале.

Из (4.12) видно, что увеличение η_r приводит к уменьшению τ_P, если время обработки в мобильном терминале меньше, чем сумма времени передачи и времени обработки в сервере:

$$\frac{\bar{v}}{O_r} < \left(\frac{\bar{v}}{C} + \frac{\bar{v}}{O_S}\right). \tag{4.14}$$

Руководствуясь (4.14), система обслуживания может распределять функции обработки данных между сервером и терминалом, например, в зависимости от загрузки сервера.

4.2.2 Формирование и обновление данных

Данные, используемые для услуги, могут иметь различное происхождение и размещение в сети. Поставщиками данных могут быть разные люди, организации и технические системы, создающие или представляющие информационное обеспечение в областях, доступных пользователю услуги. Поиск необходимых данных производит сервер услуги с помощью соответствующего программного обеспечения, на основе данных запроса,

сформированных абонентским терминалом. Как было показано выше, таких поисковых систем (серверов) может быть несколько.

Каждая из них имеет собственную БД, в которой хранятся наиболее востребованные данные, что позволяет уменьшить затраты времени, необходимого для доставки данных. Например, система уровня клиента может быть размещена на терминале пользователя, а ее БД содержать данные о текущем окружении клиента. Это – данные об объектах, которые могут быть идентифицированы на основе показаний датчиков терминала пользователя. Например, данные об объектах, находящихся в непосредственной близости от пользователя, что оценивается по информации о географическом положении пользователя (терминала).

Функциональность поисковых систем услуги ДР аналогична функциональности существующих систем за исключением специфики предоставления услуги, определяемой набором признаков, по которым производится поиск, и способа представления результатов поиска. Сформулируем требования к поисковой системе.

1. Возможность поиска по таким признакам, как географические координаты, локальные координаты (внутри зданий и помещений), графические и текстовые идентификаторы объектов (*Bar*- и *QR*-коды, текстовые названия), неподвижные и подвижные изображения объектов поиска (поиск графических образов), речь и звуки (распознавание речи, музыки и др. звуков). Для реализации такой возможности указанные признаки должны быть достаточно формализованы, а также определены методы их формирования соответствующим приложением на мобильном терминале или сервере услуги.

2. Возможность классификации данных об объектах поиска по целевому назначению, прикладной области, типу услуги, географическому положению, принадлежности, виду представления, типу источника данных. Например, одному объекту может быть сопоставлено несколько блоков данных, предоставленных различными источниками. В таком случае необходим метод

выбора необходимого блока данных, который соответствует параметрам предоставляемой услуги.

3. Возможность выбора формы представления данных, например, текстовое сообщение, изображение или подвижное видео, пиктограмма, речевое или звуковое сообщение, язык сообщения и т.д.

Время хранения данных в локальных БД должно определяться исходя из информации об их востребованности и требований к качеству обслуживания (времени поиска).

Как было отмечено выше, основная функциональность услуги ДР – предоставление необходимой информации в соответствии с данными запроса. Очевидно, что в такой общей формулировке, эта задача аналогична той, которая решается поисковыми системами глобальной сети интернет. Однако в случае услуги ДР необходимо учесть ряд особенностей как запросов, так и предоставляемых данных. Сам принцип организации услуги определяет, например, такие особенности как корреляция между востребованностью информации и содержимым окружения пользователя. Из этого следует, что вероятность получения сервером запроса к определенному информационному блоку зависит от географического положения объекта, которому он сопоставлен и географического положения пользователя. Справедливо предположить, что чем они ближе друг к другу, тем выше эта вероятность. Фактически она определяет долю трафика, создаваемого пользователями на сервер (серверы услуги). Из приведенных рассуждений следует, что востребованность информации ДР различна и зависит от разных факторов, по крайней мере географических, что позволяет локализовать (кластеризовать) данные ДР как показано на рисунке 4.4. При этом физически кластер данных может размещаться в БД сервера, который, в свою очередь, находится в географической близости к потенциальным пользователям. Естественно, что в таком случае экономится ресурс сети связи.

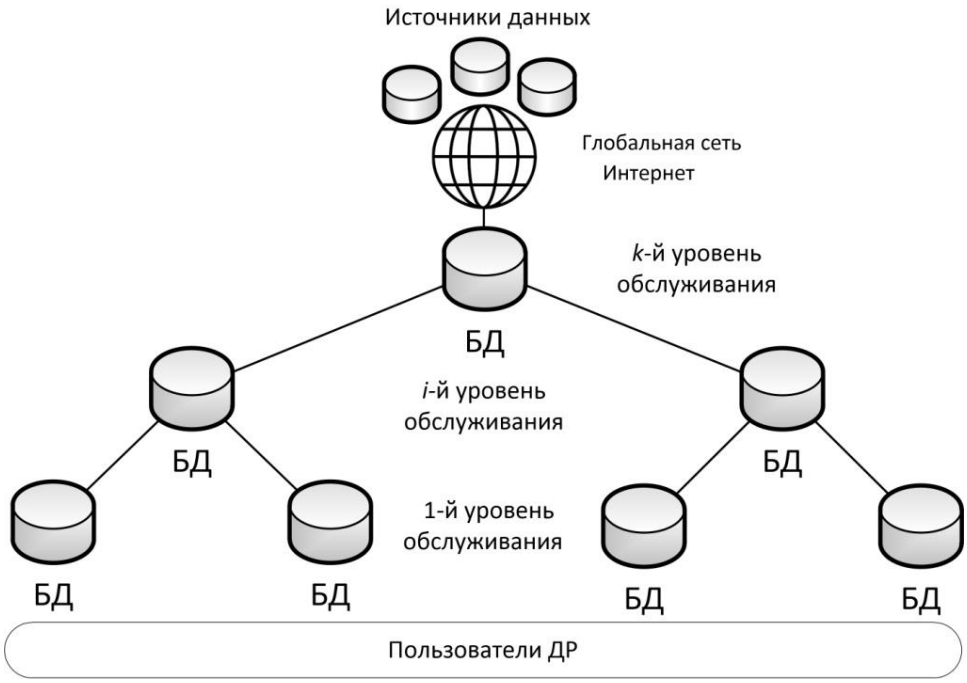

Рисунок 4.4 - Структура размещения данных в системе обслуживания ДР

Большая доля трафика пользователей будет замыкаться в рамках сети, обслуживающей определенную географическую область. Данный подход может быть реализован при организации нескольких уровней обслуживания запросов пользователей и распределении данных и трафика.

4.3 Иерархическая структура предоставления услуг дополненной реальности для распределения нагрузки и данных

Для уменьшения времени реакции путем снижения нагрузки на сервер услуги можно организовать иерархическую структуру, включающую несколько уровней обслуживания. Сервер каждого из уровней доступен разному количеству пользователей. Например, сервер первого (низшего) уровня может быть организован непосредственно в мобильном терминале и иметь единственного пользователя. Обращение к серверу более высокого уровня происходит в случае, когда требуемая информация не найдена в БД серверов низших уровней. БД сервера каждого из уровней содержит информацию об окружении каждого из пользователей, для которых этот

сервер доступен, а также информацию, востребованную пользователями, причем вероятность этого востребования равна *p*. Иерархическая структура предоставления услуги показана на рисунке 4.5.

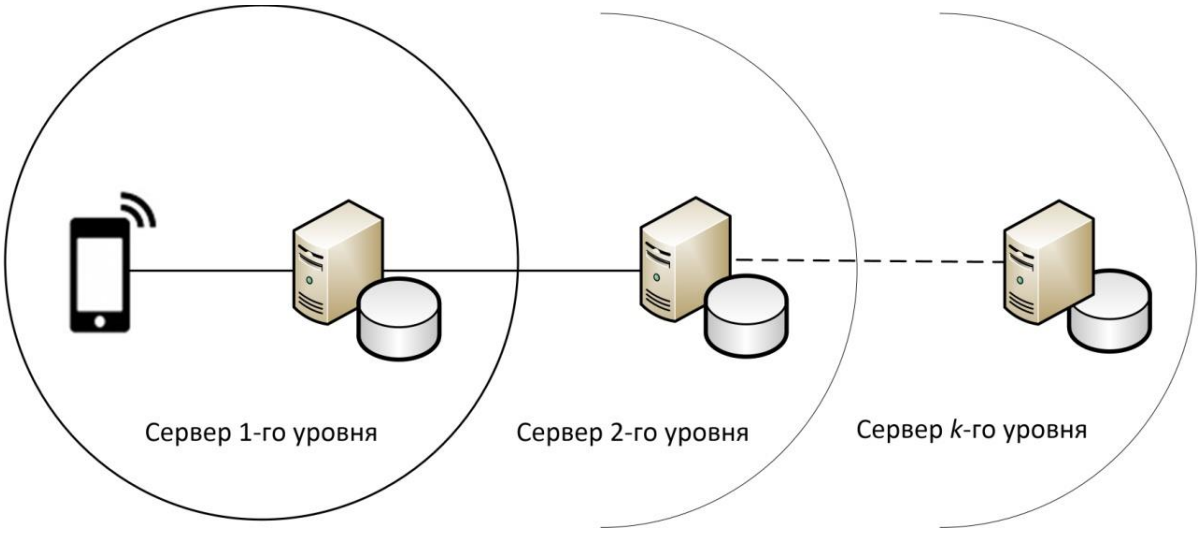

Рисунок 4.5 - Иерархическая структура предоставления услуги

Рассмотрим модель предоставления услуги. При обнаружении изменения окружения терминал пользователя передает данные (возможно уже сформированный запрос) на сервер услуги 1-го уровня. Сервер производит обработку данных и запроса. При успешной обработке и наличии данных сервер отправляет ответ в терминал пользователя. Если по каким-либо причинам запрос не выполнен сервером данного уровня, запрос передается на сервер следующего уровня и так далее. Причиной неуспеха может быть отсутствие необходимых данных на текущем уровне обслуживания. Построение иерархической модели предоставления услуги позволяет путем распределения трафика и данных между уровнями обеспечить требуемые показатели качества предоставления услуги.

Время обслуживания запроса в сети с несколькими серверами можно описать как

$$\bar{\tau} = \bar{t}_r + \sum_{j=1}^{k} p_j \sum_{i=1}^{j} \left(\bar{t}_q + \bar{t}_s\right) + \bar{t}_a + \bar{t}_d \,, \qquad (4.15)$$

где p_j – вероятность того, что запрашиваемые данные находятся в БД сервера *j*-го уровня;

k – количество уровней (см. рис. 4.5).

Или с учетом модели (4.5)

$$\bar{\tau} = \bar{t}_m + \sum_{j=1}^{k} p_j \sum_{i=1}^{j} \left(\frac{\bar{v}_q + \bar{v}_a}{C} + \frac{\rho_s}{\mu_s(n_s) 2(1-\rho_s)} \left(1 + V_s^2\right) + \frac{1}{\mu_s(v_s)} \right). \qquad (4.16)$$

При организации услуги в сети *SDN* ее функциональность может быть использована для динамического управления услугой путем изменения количества уровней обслуживания, т.е. увеличения или уменьшения k в выражениях (4.15) и (4.16).

Критерием принятия решения является время реакции $\bar{\tau}$, вернее его значение в сравнении с некоторой нормативной (целевой) величиной τ_0, значение которой наиболее приемлемо при реализации услуги. Очевидно, что целевым значением может быть выбран *0* (ноль), но очевидно, что такая цель недостижима. Уменьшение задержек на обработку и доставку может быть связано со значительными финансовыми затратами, поэтому приемлемость величины τ_0 целесообразно рассматривать как максимальное значение, при котором обеспечивается желаемое качество обслуживания (*QoS*) и качество восприятия услуги пользователями (*QoE*).

Таким образом, в рамках данной модели управление услугой заключается в поддержании возможной близости между реальной величиной времени реакции и ее целевым значением, т.е. в обеспечении $\min\left|\bar{\tau} - \tau_0\right|$. Тогда целевая функция данной задачи может быть записана как минимум разности между временем реакции и нормативом. Минимум разности можно выразить через минимум квадрата разности и применить метод наименьших квадратов. Тогда задачу можно сформулировать как задачу оптимизации с целевой функцией

$$\{k, p_i\} = \arg\min_{k, p_i} \left\{ \sum_{i=1}^{k} \left(\bar{\tau} - \tau_0\right)^2 \right\} \qquad (4.17)$$

и ограничениями

$$k \in \mathrm{N}, \quad k \le k_{\max}, \quad 0 \le p_i \le 1, \quad \bar{\tau} > 0, \quad \tau_0 > 0,$$

где $\bar{\tau}$ – определяется выражениями (4.15) или (4.16);

τ_0 – целевой значение времени реакции;

$k_{\max}$ – максимально допустимое количество уровней обслуживания. Стоит заметить, что выражения (4.15) и (4.16) являются лишь возможными моделями для описания временных параметров услуги. В качестве их могут выступать как аналитические, так и имитационные модели, позволяющие адекватно оценить интересующие параметры.

Приведенная выше задача (4.17) сформулирована как задача поиска оптимального количества уровней обслуживания k и значений p_i, которые определяют не только состав данных в БД, но и доли трафика на каждом из уровней обслуживания.

Фактически, состав данных в БД сервера уровня i может определяться согласно правилу: блок данных сохраняется в БД, если доля запросов к нему превышает величину p_i. Последнее фактически означает, что на данном уровне обслуживания замкнется трафик, создаваемый поступающими запросами.

Таким образом, реализация услуги дополненной реальности предполагает выполнение системой обслуживания функций по обработке, передаче, хранению, выборке данных и представлению информации пользователю. Выполнение каждой из этих функций требует затрат ресурсов времени, пропускной способности сети, производительности серверов и памяти.

Информационное обеспечение услуги (информация ДР) может формироваться различным образом, в том числе на основе результатов поиска информации, представленной в глобальной сети интернет. Поиск и хранение данных выполняется системой обслуживания, которая может иметь несколько уровней обработки. Их количество влияет на объем используемых ресурсов (пропускной способности сети, производительности серверов, памяти).

Основным показателем качества предоставления услуги является время реакции, т.е. время с момента изменения окружения пользователя до момента

представления пользователю необходимого сообщения. Это время зависит от распределения функций предоставления услуги по исполнительным элементам (терминал пользователя, серверы ДР, каналы сети ПД). Целевое значение этого времени не должно превышать время реакции пользователя на представляемое сообщение ДР.

Для обеспечения качества восприятия услуги ДР могут быть использованы ресурсы мобильного терминала (снижают время ПД), а также ресурсы *SDN*, позволяющие реализовать иерархическую модель предоставления услуги. Иерархическая модель дает возможность локализовать значительную долю данных и трафика, что позволяет экономить ресурсы пропускной способности сети связи.

Параметры иерархической модели предоставления услуги включают количество уровней иерархии и вероятность обращения к каждому из них. Вероятность обращения к некоторому уровню может быть использована как критерий для формирования БД.

Выбор количества уровней иерархии в модели предоставления услуги и вероятности обращений к каждому из них представляет собой задачу оптимизации, цель которой – обеспечение времени реакции, ближайшего к заданному значению. Решение этой задачи позволяет получить структурные параметры системы обслуживания на основе данных о трафике пользователей.