

Понятие оптимизации

Сейчас стало общепринятым утверждение о том, что анализ, синтез и управление сложными системами практически любого происхождения, в том числе, естественно, и сетей связи, возможны лишь на основе описания этих систем в форме математических моделей [1]. Характерным для этих моделей является наличие в их составе двух групп элементов, увязанных в единую систему, в которой они тем или иным образом взаимодействуют. Значения одних элементов являются альтернативными, могут избираться в определенной области; значения других - производны от первых. Говорят о параметрах управления и целевой функции.

Оптимизация системы состоит в поиске такого набора параметров управления, при котором целевая функция достигает экстремума. Этим достигается **новое качественное состояние системы** — экстремальность результата ее функционирования по заданному критерию качества. Надо отметить, что новое качество постигается нашим сознанием в этом процессе на более высокой ступени познания не как первичная логическая категория, обеспечивающая лишь весьма элементарные различия и оценку качества (типа «больше - меньше»), а как значительно более развитая категория количественно определенной оценки качества. Оптимальное (наилучшее из всех возможных) состояние управляемой системы и представляет собой такую количественно определенную оценку качественного состояния системы. Оптимизация решений в

сложных системах придает им новое качество, количественно выражающееся в существенном повышении эффективности их функционирования.

Между тем, в текущей научной и близнаучной практике почти постоянно встречается неточное, расплывчатое понимание терминов «оптимум» и «оптимальный». Эти термины получают широкое распространение в самом причудливом многообразии различных оттенков оценки явлений внешнего мира, подменяя гораздо более четкие понятия и категории оценочного порядка, легко отличимые друг от друга. Такая расплывчатость, стертая трактовка понятия оптимума становится угрозой самому существованию этого понятия как научного.

Несколько разительных примеров:

- 1) видный ученый, генеральный конструктор одной из ведущих отраслей народного хозяйства, обещает (в печати) выполнить проект новой машины «оптимально и в нужный срок». В таком контексте термин «оптимально» не может означать ничего другого, как просто «хорошо»;
- 2) крупный хозяйственник, руководитель ведомства, говорит о развитии производства «наиболее оптимальным, кратчайшим путем, соответствующим интересам всего общества», не обращая внимание на то, что в этом определении совмещаются разные критерии качественной оценки, и что словосочетание «наиболее оптимальный» само по себе нелепо!
- 3) крупный археолог, ученый с большим кругозором и научным размахом, говоря об измерении расстояния с точностью до 10м, вместо того чтобы определить эту точность как приемлемую или

допустимую в данных условиях, предпочитает определять ее как оптимальную, что просто противоречит смыслу этого термина, но, видимо, кажется автору более «научным».

Примеры эти приведены не для курьеза, а только потому, что характеризуемое ими ложное «онаучивание» — занятие совсем не невинное. При широком распространении оно может обернуться против науки.

Задачи оптимизации, как правило, трактуются как задачи распределения ресурсов. Ресурсы — одно из ключевых понятий, относящихся к проблематике пособия. Прежде всего это ассигнования (инвестиции). Не следует думать, что этот компонент ресурсов имеет значение только при разработке программ создания новых или модернизации существующих сетей. Он определяющим образом влияет и на процесс обоснования решений при комплектации различного рода вторичных сетей из арендуемых компонентов (элементов) реально существующих первичных сетей. К ресурсам можно отнести также время, отпускаемое на создание системы, производственные мощности, спектр возможных частот передачи информации, полосу пропускания магистральных каналов и т. п. [2].

Комплекс проблем оптимизации сетей связи

Обширный комплекс задач оптимизации сетей включает, как минимум, триаду из выбора структуры, планов функционирования и программ создания (модернизации) сетей (рис. В1). Выбор структуры, обычно называемой в литературе топологией [3], сводится к обоснованию архитектуры и оптимальному распределению ресурсов между подсистемами, комплексами, элементами и средствами связи (рис. В2).

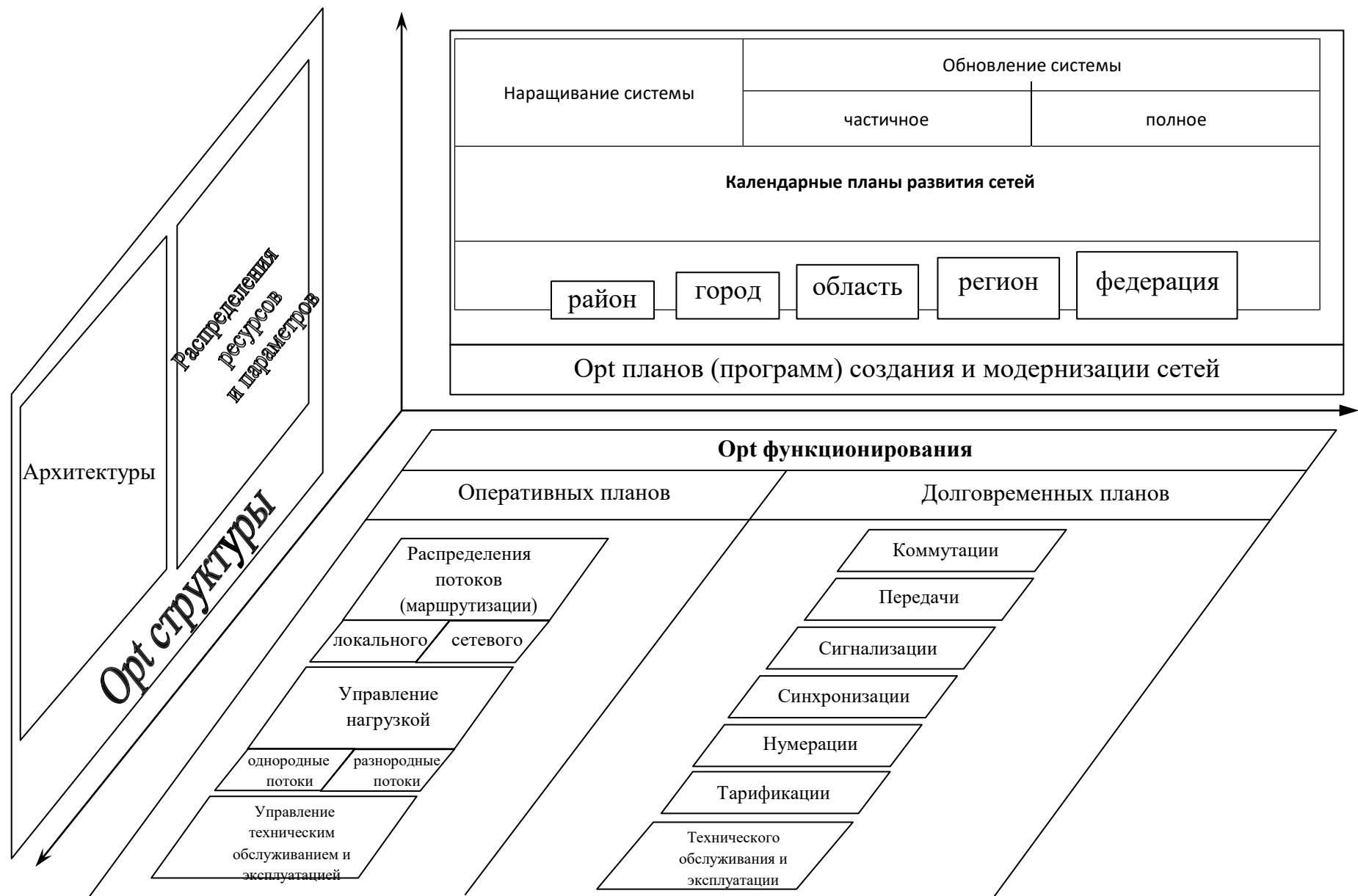


Рис. В1. Комплекс проблем оптимизации сетей связи

В свою очередь, архитектуру формализуют заданием количества уровней иерархии системы (оптимизация количества уровней — чисто комбинаторная задача) и решением задач оптимизации структуры на каждом из уровней. Структура уровня определяется числом и расположением (географическим) элементов и характером связей между ними. Оптимизационные задачи здесь подобны задачам Штейнера и коммивояжера [4] как по характеру математических моделей, так и трудностям получения оптимальных решений.

Оптимальное распределение ресурсов в комплексе задач оптимизации структуры реализуется многоуровневой (на схеме рис. В2 для простоты — трехуровневой) совокупностью моделей, информационно связанных между собой. В соответствии с принципом минимаксного построения моделей [5] главной характеристикой модели системы определенного уровня является отражение сложности именно данного уровня. Нельзя «тащить» на более высокий уровень сложности предыдущих уровней, поэтому модели предыдущего уровня передают на следующие уровни лишь обобщенную информацию.

Так, от моделей низших уровней, с помощью которых оптимизируются отдельные комплексы (например, концентраторы, мультиплексоры, узлы коммутации и т. п.), на следующий уровень передаются лишь оптимизированные значения вероятностно-временных и стоимостных характеристик комплексов.

В моделях среднего уровня на основе этих данных оптимизируются сети различного назначения. Одиночным результатом решения подобных задач является минимальное значение какого-нибудь ресурсного показателя (например стоимости)

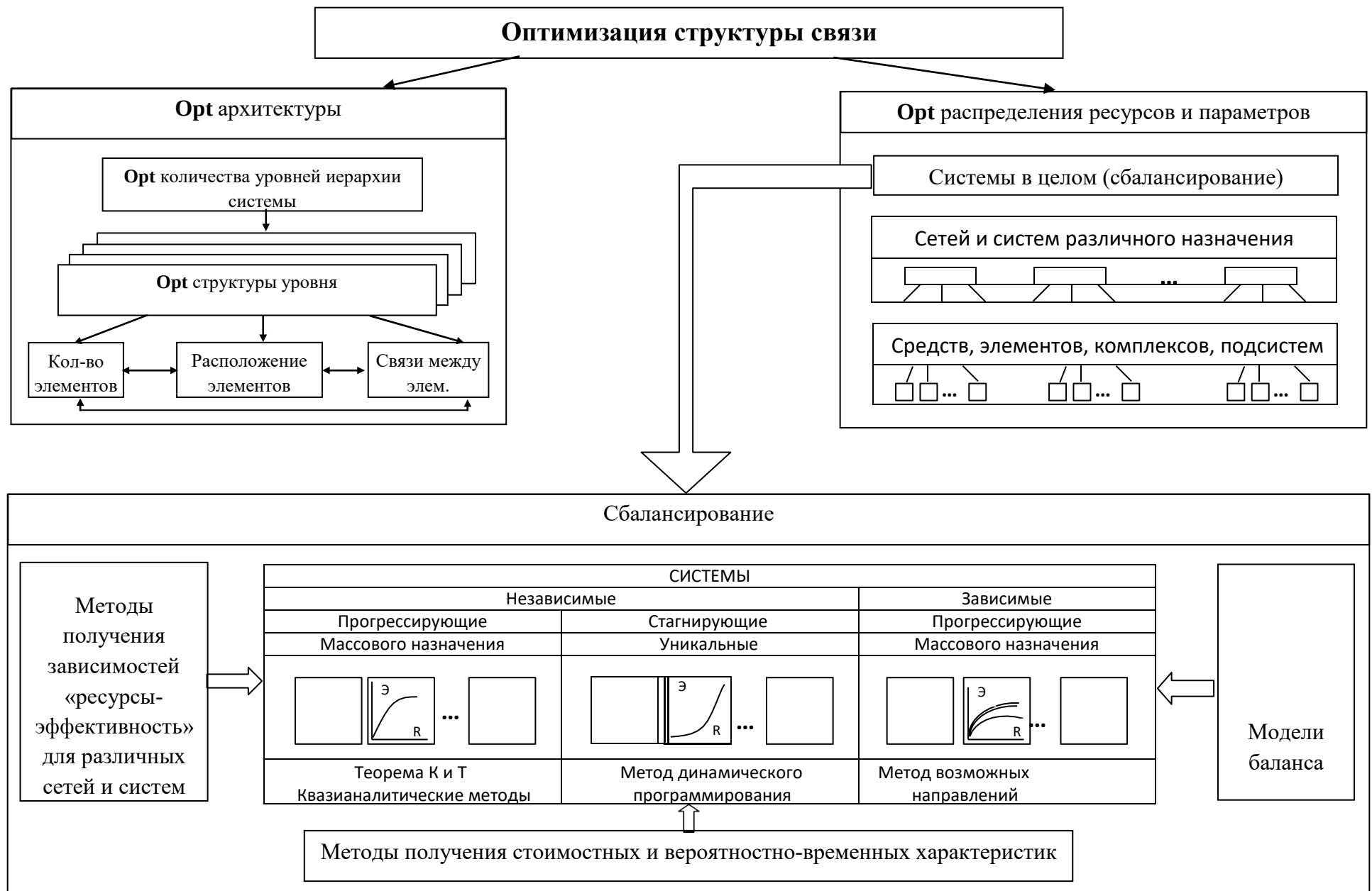


Рис. В2. Комплекс задач оптимизации структуры сетей связи

системы при заданном значении ее эффективности. Варьируя задаваемые значения эффективности и решая для каждого из этих значений соответствующую оптимизационную задачу, получают зависимость «ресурсы - эффективность» (РЭФ), обобщенно характеризующую рассматриваемую подсистему.

На следующем, самом высоком уровне моделирования нет ни комплексов, ни конкретных характеристик подсистем. Есть только зависимости РЭФ для подсистем, входящих в общую сеть связи, на основании которых и осуществляется сбалансирование, т. е. распределение ресурсов между подсистемами.

Модели баланса (сбалансирования) существенно зависят от характера подсистем. Прежде всего различают системы массового и уникального значения (соответственно СМН и СУН). Различие состоит в том, что для СУН недостаточно, как для СМН, **в среднем** хорошо выполнять задачу, когда недовыполнение заданий (требований, нормативов) одними компонентами системы может быть компенсировано перевыполнением заданий другими компонентами. Необходимо выполнение заданий всеми компонентами. Невыполнение задания хотя бы одним компонентом делает бесполезными любые успехи других компонентов.

Принципиальным является то, что СМН и СУН обладают качественно отличающимися зависимостями РЭФ, благодаря чему СМН относят к прогрессирующему, а СУН — стагнирующему типу. У прогрессирующих систем зависимости РЭФ являются вогнутыми, т. е. даже минимальное вложение ресурсов приводит к быстрому росту эффективности системы. Величина прироста является максимальной при малых значениях ресурса и постепенно убывает при возрастании ресурса. Здесь оказываются эффективными квазианалитические

методы получения оптимальных решений, основанные на применении теоремы Куна и Таккера [4].

В отличие от этого стагнирующие системы характеризуются невогнутыми, S-образными функциями РЭФ, содержащими обширный участок крайне медленного роста. Применение методов, опирающихся на теорему Куна и Таккера, здесь неэффективно, в частности из-за многоэкстремальности. Оптимальные решения могут быть получены методом динамического программирования.

Наконец, системы делят на зависимые и независимые. У независимых систем функции РЭФ зависят каждая только от одной переменной. Другими словами, эффективность данной системы зависит только от объема ресурса, выделенного на ее развитие. У зависимых же систем предоставление ресурса какой-нибудь системе увеличивает не только ее эффективность, но и — в определенной мере — эффективность каждой из остальных систем. Каждая из функций РЭФ здесь является функцией многих переменных.

Решение может быть получено методом возможных направлений [4]. Основная трудность, однако, заключается не в отыскании подходящего метода решения оптимизационной задачи, а в построении функций РЭФ, что требует одновременного анализа различных подсистем связи.

Блок проблем оптимизации функционирования содержит два класса задач. Одни характерны для этапа проектирования сетей и предназначены для выработки различного рода долгосрочных планов. Другие возникают непосредственно в процессе эксплуатации сетей, носят динамический характер и предназначены для выработки оперативных планов. Центральными являются задачи распределения потоков и управления нагрузкой.

Важнейшим в общем комплексе проблем является блок обоснования планов (программ) создания и модернизации систем. Учет динамики производства, т. е. последовательности и сроков создания элементов различных типов, приводит к задачам календарного планирования, трудности решения которых известны [6]. Общие методы точного решения таких задач отсутствуют, однако, в каждом конкретном случае, как показывает опыт, могут быть предложены алгоритмы получения приближенных решений, в которых используются различные эвристические правила.

Более удобен в практической реализации итерационный процесс, на каждом шаге которого решаются две задачи. В первой оптимизируется состав системы (задача из блока оптимизации структуры) с учетом лишь интегральных ограничений на количества создаваемых элементов различных типов, динамика создания элементов не рассматривается. В результате формируется вектор-заявка для реализации промышленными предприятиями. Для второй задачи вектор-заявка является идеалом, который должен быть достигнут с учетом динамики производства элементов. Если существует хотя бы одно допустимое решение второй задачи, то можно считать задачу решенной в целом. Выбор между несколькими допустимыми решениями может осуществляться с помощью дополнительных критериев, оптимизирующих функционирование предприятий. При отсутствии допустимых решений (наиболее типичная ситуация) вектор-заявка не может быть реализована в полном объеме. Однако при этом может быть получено решение, лежащее наиболее близко к вектор-заявке; одновременно уточняются интегральные возможности промышленности, определенные в данном случае с учетом динамики ее функционирования. Это дает основание вернуться к первой задаче, после чего процесс повторяется. В

формулировках обеих задач существенными являются вопросы учета начального состава элементов системы. При этом возможны различные стратегии, начиная от наращивания системы до полного ее обновления.

Реализация

Формой реализации исследуемых подходов является разрабатываемый пакет прикладных программ (ППП) решения задач оптимизации сетей связи (рис. В3). Пакет содержит, прежде всего, блок общих моделей и методов, применимых не только к сетям связи, но и к другим системам. Здесь можно выделить методы случайного поиска (в основном, по алгоритмам работы [4]), градиентный метод применительно к решению задачи Штейнера, методы отыскания экстремумов унимодальных функций (метод чисел Фибоначчи и его асимптотическая аппроксимация -метод золотого сечения) и семейство методов получения точных и приближенных решений ряда задач математического программирования — также преимущественно по алгоритмам работы [4]. Раздельно рассматриваются задачи для однородных и разнородных ресурсов. Необходимость разделения связана с существенным усложнением задач распределения при переходе от однородных к разнородным ресурсам.

Так, для распределения однородных ресурсов с успехом, вплоть до получения аналитических или квазианалитических выражений, могут быть применены следствия из теоремы Куна и Таккера (для прогрессирующих систем, одно- и двусторонних ограничений), методы динамического программирования (для стагнирующих систем и только односторонних ограничений) и др.

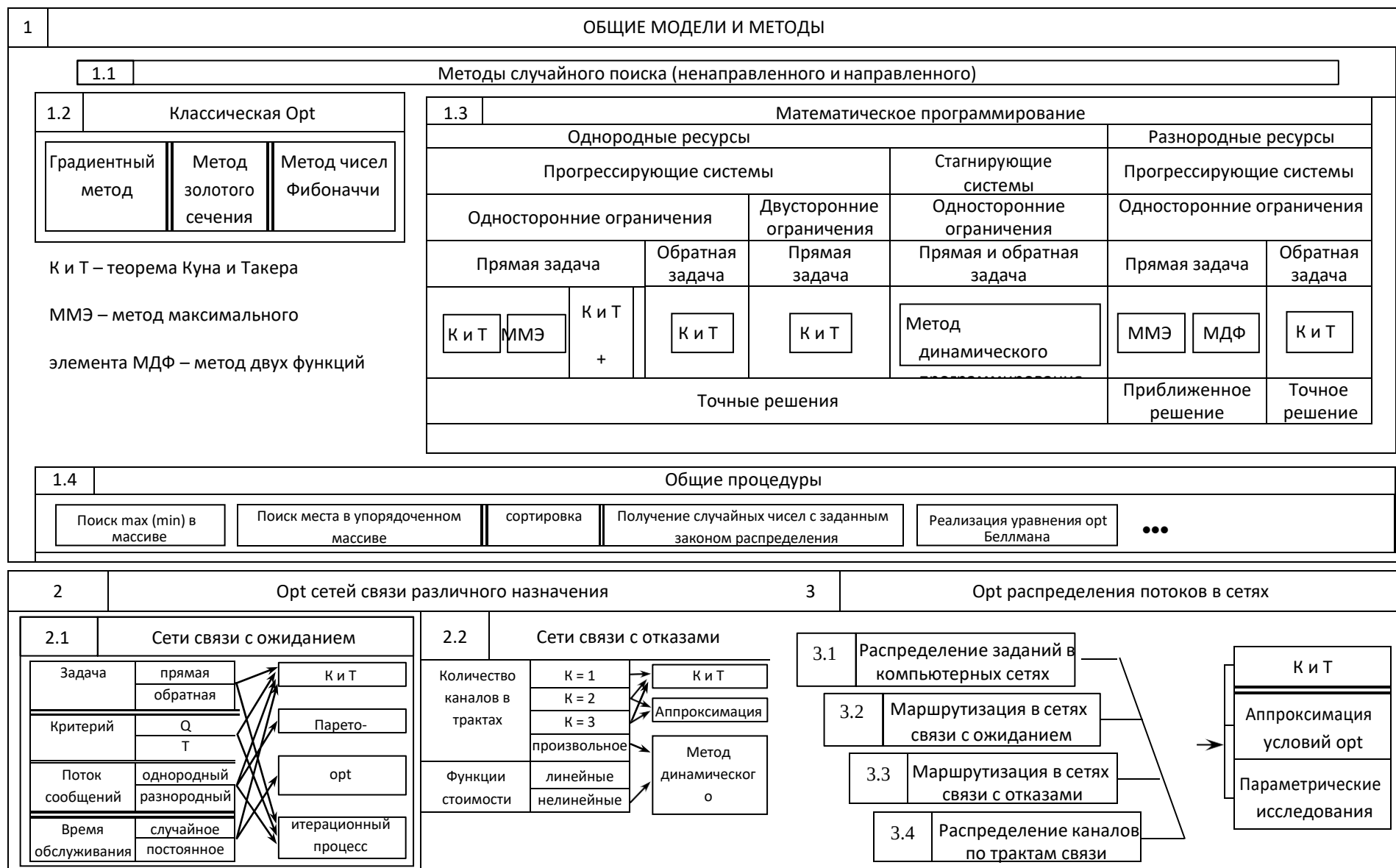


Рис. В3. Структура пакета прикладных программ решения задач оптимизации сетей связи

Задачи блока 1 и 2 относятся к выбору или распределению пропускных способностей для сетей связи с ожиданием и отказами, т. е. к отдельному фрагменту комплекса задач оптимизации структуры сетей.

Подобные задачи рассматривались в работе [8] и ряде других, связанных с нею, но в предложенных там алгоритмах имеются существенные некорректности. Они возникают в связи с тем, что, признавая на словах принадлежность задачи к классу задач математического программирования, авторы на деле решают ее как классическую задачу условной оптимизации. Вместо использования в полном объеме теоремы Куна и Таккера они ограничиваются классическим методом Лагранжа, открытым в прошлом веке и пригодным только для решения задач классической оптимизации.

Разработанные и используемые в блоке 2 пакета методы и алгоритмы свободны от указанных некорректностей. Их можно применять для решения по различным критериям прямых и обратных оптимизационных задач, ситуаций обслуживания однородных и разнородных потоков с экспоненциальным или постоянным временем обслуживания, при различном количестве каналов в трактах связи, линейных и нелинейных функциях стоимости.

Эффективными здесь оказываются методы, основанные на использовании теоремы Куна и Таккера, и метод динамического программирования. В отдельных случаях требуется предварительная аппроксимация условий оптимальности или применение итерационного процесса в случаях, когда условия оптимальности оказываются практически неразрешимыми относительно неизвестных пропускных способностей. Помимо скалярной, рассматриваются и задачи векторной оптимизации, которые возникают при исследовании

мультисервисных систем, обслуживающих разнородные потоки. Здесь используется метод получения Парето-оптимальных решений.

1. СИСТЕМЫ СВЯЗИ С ОЖИДАНИЕМ

1.1. Постановка задачи. Математическая модель системы

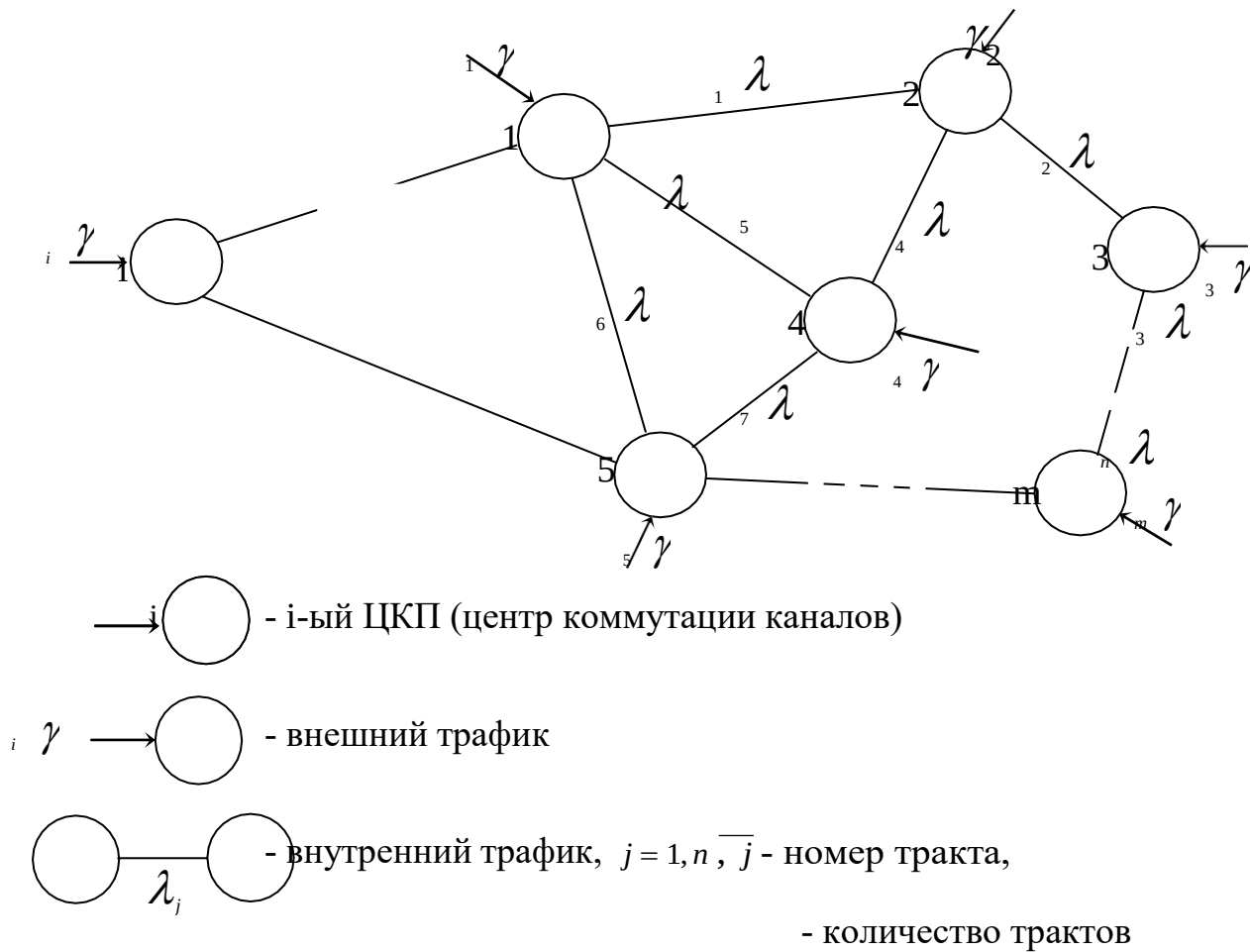


Рис. 1.1. Принципиальная схема оптимизируемой сети связи

Рассмотрим сеть связи с m центрами коммутации пакетов (ЦКП) ($i = 1, m$) и n участками передачи ($i = 1, n$) между ЦКП (n трактами связи). На каждый ЦКП от пользователей поступает поток сообщений (кадров, пакетов) с интенсивностью $\gamma_i, i = 1, m$ (рис. 1.1). Суммарный внешний трафик $\gamma_{\Sigma} = \sum_{i=1}^m \gamma_i$ в соответствии с принятой процедурой маршрутизации распределяется по трактам связи, порождая внутренний трафик с интенсивностями $\lambda_j, j = 1, n$. Имеет место естественное соотношение

$$\gamma_{\Sigma} = \sum_{i=1}^m \gamma_i = \sum_{j=1}^n \lambda_j = \lambda_{\Sigma} \quad (1.1)$$

где λ_{Σ} – суммарный внутренний трафик.

Из (1.1) вытекает такое неравенство:

$$\pi = \lambda_{\Sigma} \gamma_{\Sigma} \geq 1 \quad (1.2)$$

где π – среднее число транзитов в сети.

Каждый тракт представляет собой систему обслуживания и должен характеризоваться производительностью (пропускной способностью). Имеется некоторый ресурс S_0 , распоряжаясь которым можно создавать тракты различной производительности μ_j .

Может быть сформулирована следующая практически важная задача: при заданных структуре и параметрах сети, характеристиках внешнего и внутреннего трафиков и значении выделенного ресурса средств определить такие значения пропускных способностей трактов, при которых обеспечивается максимальная в том или ином смысле эффективность функционирования сети. Будем называть такие оптимизационные задачи *прямыми*, отличая их от *обратных*, где заданное значение эффективности сети обеспечивается минимальным объемом расходуемых ресурсов.

Эквивалентом сформулированной вербальной постановки служит следующая математическая модель (рис. 1.2).

Введем следующие допущения [1]:

- потоки в трактах являются пуассоновскими, т. е. вероятность того, что интервал между пакетами (сообщениями) меньше t , равна

$$A_j(t) = 1 - \exp(-\lambda_j t) \quad (1.3)$$

где λ_j – интенсивность потока (1/с);

I. Факторы и параметры

1.1. Интенсивность	$\begin{bmatrix} \text{потока} \\ \text{пакетов} \end{bmatrix}$	$\begin{bmatrix} \lambda_j \\ \mu_j \end{bmatrix}$
--------------------	---	--

1.2. Надежность	$\begin{bmatrix} \text{коэффициент} \\ \text{среднее время} \end{bmatrix}$	$\begin{bmatrix} \text{готовности} \\ \text{простоя} \\ \text{исправности} \\ \text{простоя} \end{bmatrix}$	$\begin{bmatrix} k_{rj} \\ k_{\Pi j} \\ T_{\Pi j} \\ T_{\Pi j} \end{bmatrix}$	тракта
-----------------	--	---	---	--------

1.3. Старение информации	$\begin{bmatrix} \text{среднее время} \end{bmatrix}$	$\begin{bmatrix} T_{cj} \\ \nu_j = 1 / T_{cj} \end{bmatrix}$
--------------------------	--	--

1.4. Стоимостные характеристики создания и эксплуатации трактов		
1.4.1. Затраты	начальные оперативные	S_{cj} S_j
1.4.2. Выделенный ресурс S_0		

1.5. Минимально возможная интенсивность обслуживания (пропускная способность) трактов μ_{0j}
--

II. Модельные соотношения

2.1. Характеристики тракта	$\begin{bmatrix} \text{вероятность своевременной} \\ \text{доставки пакета с учетом} \\ \text{факторов 1.1, 1.2, 1.3} \\ \text{стоимость с учетом факторов 1.4.1} \end{bmatrix}$	$\begin{bmatrix} Q_j(\mu_j) \\ S_j(\mu_j) \end{bmatrix}$
$Q_j(\mu_j) = \frac{\mu_j k_{aj} - \lambda_j}{\mu_j k_{aj} - \lambda_j + (\nu_j + \mu_j k_{aj} f_j)}, S_j(\mu_j) = s_{0j} + s_j \mu_j$		
2.2. Характеристики сети	$\begin{bmatrix} \text{среднесетевая вероятность} \\ \text{своевременной доставки} \\ \text{стоимость} \end{bmatrix}$	$\begin{bmatrix} Q = \sum_{j=1}^n \frac{\lambda_j}{\lambda_{\Sigma}} Q_j(\mu_j) \\ S = \sum_{j=1}^n S_j(\mu_j) \end{bmatrix}$

III. Оптимизационная задача

$Q = \sum_{j=1}^n \frac{\lambda_j}{\lambda_{\Sigma}} Q_j(\mu_j) \Rightarrow \max_{\{\mu_j\}}$
$S = \sum_{j=1}^n S_j(\mu_j) \leq S_0$
$\mu_j \geq \mu_{0j}, j = 1, 2, \dots, n$

Рис. 1.2. Математическая модель и формулировка задачи

- время обслуживания пакетов в тракте – случайная величина с экспоненциальной функцией распределения (ФР), т. е. вероятность того, что время обслуживания пакета будет меньше t , равна

$$B_j(t) = 1 - \exp(-\mu_j t) \quad (1.4)$$

где $\mu_j = R_j / l_j$ – интенсивность обслуживания пакетов (1/с);

R_j – эффективная скорость передачи (бит/с),

l_j – средняя длина пакета (бит);

- каналы в трактах связи обладают конечной (неабсолютной) надежностью, вследствие чего периоды нормального функционирования из-за возникающих отказов чередуются с периодами простоя, в течение которых отыскивается и устраняется неисправность.

Предполагается, что длительности исправной работы и восстановления – суть случайные величины с экспоненциальными ФР

$$C_j = 1 - \exp(-c_j t), \quad c_j = 1 / T_{Иj}, \quad (1.5)$$

$$D_j = 1 - \exp(-d_j t), \quad d_j = 1 / T_{Пj}, \quad (1.6)$$

где $T_{Иj}$ ($T_{Пj}$) – среднее время исправной работы (простоя) j -го тракта связи (с);

- качество информации со временем ухудшается, информация «стареет». Предполагается, что время старения – случайная величина с экспоненциальной ФР

$$Z_j(t) = 1 - \exp[-v_j t] \quad v_j = 1/T_{cj} \quad (1.7)$$

v_j – интенсивность старения информации (1/с),

T_{cj} – среднее время старения информации.

Можно показать [2], что при принятых допущениях справедливы следующие соотношения:

- вероятность своевременной* доставки сообщения в j -ом тракте связи равна

$$Q_{\mu_j} = \frac{\mu_{\varepsilon j} - \lambda_j}{\mu_{\varepsilon j} - \lambda_j + v_{\varepsilon j}} \quad (1.8)$$

«Своевременной» считается доставка сообщения за время не превышающее времени старения информации

- среднее время доставки сообщения в j -м тракте связи равно

$$T_j \mu_j = \frac{\mu_{\varepsilon j} T_{\Pi j} k_{\Pi j} + 1}{\mu_{\varepsilon j} - \lambda_j} \quad (1.9)$$

Здесь $\mu_{\varepsilon j}$ – эквивалентная пропускная способность тракта связи

$$\mu_{\varepsilon j} = \mu_j k_{\Gamma j} \quad (1.10)$$

$$k_{\Gamma j} = \frac{T_{\Pi j}}{T_{\Pi j} + T_{\Pi j}} \text{ — коэффициент готовности } j\text{-го тракта связи;} \quad (1.11)$$

$$k_{\Pi j} = \frac{T_{\Pi j}}{T_{\Pi j} + T_{\Pi j}} \text{ — коэффициент простоя } j\text{-го тракта связи;} \quad (1.12)$$

$$v_{\varepsilon j} = v_j + \mu_{\varepsilon j} f_j \text{ — эквивалентная интенсивность старения информации;} \quad (1.13)$$

$$f_j = \frac{k_{nj}}{k_{ej} + 1} (v_j T_{nj}) - \text{коэффициент ненадежности } j\text{-го тракта связи.} \quad (1.14)$$

В последнее время предпринимаются шаги по коррекции соотношений (1.8) и (1.9) в связи с более полным и точным учетом эффекта старения информации. Это приводит к тому, что значения Q и T , получаемые в (1.8) и (1.9), становятся соответственно нижней и верхней оценками соответствующих вероятностных характеристик. В последующем необходимо учитывать это обстоятельство при формулировке и решении оптимизационных задач.

Показатели Q_j и T_j характеризуют каждый тракт связи по отдельности. Критерий эффективности, характеризующий сеть в целом, формируется по-разному для систем *массового* и *уникального* назначения [3]. Эффективность систем массового назначения характеризуется *средним* ожидаемым результатом. При этом частные эффекты Q_j μ_j могут быть любыми, в том числе и равными нулю. Эффективность же систем уникального назначения характеризуется гораздо более жестким критерием, когда неудача хотя бы в одном из трактов перечеркивает любые достижения в других.

В соответствии с этим в качестве критерия эффективности для систем массового назначения можно выбрать математическое ожидание числа своевременно доставленных сообщений или среднее по сети время доставки сообщений и т. п. Если T – заданное время функционирования системы, то $\lambda_j T$ – математическое ожидание количества сообщений, которые за это время пройдут через j -и тракт, а

$$K_j(\mu_j) = \lambda_j T Q_j(\mu_j) \quad (1.15)$$

– математическое ожидание числа своевременно доставленных сообщений (без учета их транзита).

Математическое ожидание числа своевременно доставленных сообщений по всей сети равно

$$K_c \mu = T \sum_{j=1}^n \lambda_j Q_j(\mu_j) \quad (1.16)$$

Аналогичной характеристикой является среднесетевая вероятность своевременной доставки, в которой вероятности Q_j «взвешиваются» с помощью значений

$$p'_j = \lambda_j / \lambda_\Sigma, j = 1, n \quad (1.17)$$

Каждое из значений p'_j представляет собой по смыслу долю общего внутреннего трафика следующего по тракту. Таким образом приходим к следующему критерию:

$$Q_c \mu = \sum_{j=1}^n p'_j Q_j(\mu_j) \quad (1.18)$$

Видно, что критерии K_c и Q_c совпадают с точностью до постоянного множителя, не влияющего на оптимизацию. В дальнейшем эффективность системы массового назначения характеризуется критерием

$$Q_c \mu = \sum_{j=1}^n \lambda_j Q_j(\mu_j) / \lambda_\Sigma, \quad (1.19)$$

допускающим любую из двух предложенных трактовок.

Аналогичной является такая характеристика, как среднесетевое значение математического ожидания времени доставки сообщений

$$T_c \mu = \sum_{j=1}^n \lambda_j T_j(\mu_j) / \lambda_\Sigma \quad (1.20)$$

Уникальность системы предъявляет повышенные требования к качеству ее функционирования. Среднесетевые показатели, в которых недостаточная или даже нулевая эффективность в отдельных трактах может компенсироваться избыточной эффективностью в других, здесь неприемлемы. Показатель эффективности в данном случае должен быть чувствителен к факту несвоевременной доставки хотя бы одного сообщения хотя бы в одном из трактов. Трактую величину T как время, в течение которого не должно быть ни одного из таких казусов, приходим к такому показателю:

$$W' = \prod_{j=1}^n [Q_j(\mu_j)]^{\mu_j T} \quad (1.21)$$

или, что эквивалентно (с точки зрения оптимизации),

$$W = \frac{1}{T} \ln W' = \sum_{j=1}^n \lambda_j \ln Q_j(\mu_j) \quad (1.22)$$

Введем стоимостные характеристики. При расчете стоимости j -го тракта S_j необходимо учитывать начальные затраты S_{0j} связанные с проектированием и капитальными вложениями на создание тракта, и затраты на создание μ_j единиц пропускной способности. Как правило, второй компонент суммарных затрат представляется линейной функцией $s_j \mu_j$ (S_j – затраты на единицу пропускной способности). Таким образом, приходим к такому выражению для $S_j(\mu_j)$:

$$S_j(\mu_j) = \begin{cases} S_{0j} + s_j \mu_j, & \text{если } \mu_j > 0 \\ 0, & \text{если } \mu_j = 0 \end{cases} \quad (1.23)$$

Функции $S_j(\mu_j)$ терпят в нуле разрыв первого рода. Использование таких функций при оптимизации приводит к значительным трудностям. Положение, однако, облегчается в связи с тем, что, как будет ясно из дальнейшего, величины μ_j не могут обращаться в нуль. К тому же введением актуальных бесконечно

малых $\varepsilon_j = \lim_{\mu_j \rightarrow 0} \mu_j > 0$ можно всегда добиться справедливости данного допущения. Поэтому в (1.23) всегда «работает» верхняя ветвь, и величины s_{0j} не оказывают влияния на оптимальные решения. Нужно только учитывать, что если S_0 – объем выделенных ресурсов при решении прямой задачи распределения (максимизация Q при ограничении на S), то фактически будет распределяться меньшая величина ресурса, равная

$$S_0 - \sum_{j=1}^n s_{0j} = S_0 - S_0 = S_{10} \quad (1.24)$$

При решении обратной задачи (минимизация S при ограничениях на Q) после получения оптимального решения $x_{\text{опт}}$ полученный минимальный расход ресурсов $S_{\min}(x_{\text{опт}})$ необходимо увеличить на ту же величину S_0 :

$$S_{\min} = S_{\min}(x_{\text{опт}}) + S_0 \quad (1.25)$$

Далее рассматривается прямая задача распределения для системы массового назначения.

1.2 . Обоснование метода решения оптимизационной задачи

1.2.1. Некорректность известных методов

Преобразуем выражение (1.18) для критерия эффективности Q , введя следующие обозначения:

$$a_{0j} = \frac{\lambda_j}{k_{rj}} \quad — \quad (1.26)$$

— эффективная интенсивность потока сообщений j -м тракте,

$$b_j = 1 + f_j, b_{0j} = \frac{\lambda_j - v_j}{b_j k_{2j}}, p_j = \frac{p'}{b_j}, c_{0j} = p_j (a_{0j} - b_{0j}) \quad (1.27)$$

Тогда справедливо

$$Q(\mu) = \sum_{j=1}^n p'_j Q_j(\mu) = \sum_{j=1}^n p'_j \frac{\mu_j - a_{0j}}{b_j(\mu_j - b_{0j})} = Q_{\max} - \sum_{j=1}^n \frac{c_{0j}}{\mu_j - b_{0j}} \quad (1.28)$$

где $Q_{\max} = \sum_{j=1}^n p'_j$.

Рассмотрим прямую задачу распределения.

Задача 1.1

$$Q(\mu) \Rightarrow \max\{\mu_j\} \quad (1.29)$$

$$S_1(\mu) = \sum_{j=1}^n s_j \mu_j \leq s_{10}. \quad (1.30)$$

Весьма часто [4, 5] ее решают, не накладывая никаких ограничений на величины μ_j . Как обычно, формируется функция Лагранжа

$$\Phi(\mu, \psi) = Q(\mu) + \psi(S_{10} - \sum_{j=1}^n s_j \mu_j), \quad (1.31)$$

но, вместо применения в полном объеме основной теоремы математического программирования [6], просто приравняются нулю ее производные по всем переменным

$$\frac{c_{0j}}{\mu_j - b_{0j}} - \psi s_j = 0, j = 1, 2, \dots, n. \quad (1.32)$$

Из (1.32) с учетом условия (1.30) получаются следующие выражения для «оптимального» решения:

$$\mu_j = b_{0j} + \frac{S_{10} - \sum_{j=1}^n s_{\eta j} b_{0j}}{\sum_{j=1}^n s_{\eta j} c_{0j}} \quad (1.33)$$

Ясно, что «произвол» в обращении с основной теоремой математического программирования может приводить к ошибкам. Как видно из (1.33), компоненты найденного, якобы оптимального, решения при произвольной величине ресурса в частности при

$$S_{10} < \sum_{j=1}^n s_{\eta j} b_{0j} \quad (1.34)$$

(а это возможно, поскольку априорных ограничений на величину **не** накладывалось), могут быть меньше величин b_{0j} и даже отрицательными. Впрочем, уже при $\mu_j < b_{0j}$ вероятность своевременной доставки сообщений $Q_j > 1$ (рис. 1.3), что лишено всякого смысла. Но даже при

$$S_{10} > \sum_{j=1}^n s_{\eta j} b_{0j} \quad (1.35)$$

в решении возможны несуразности.

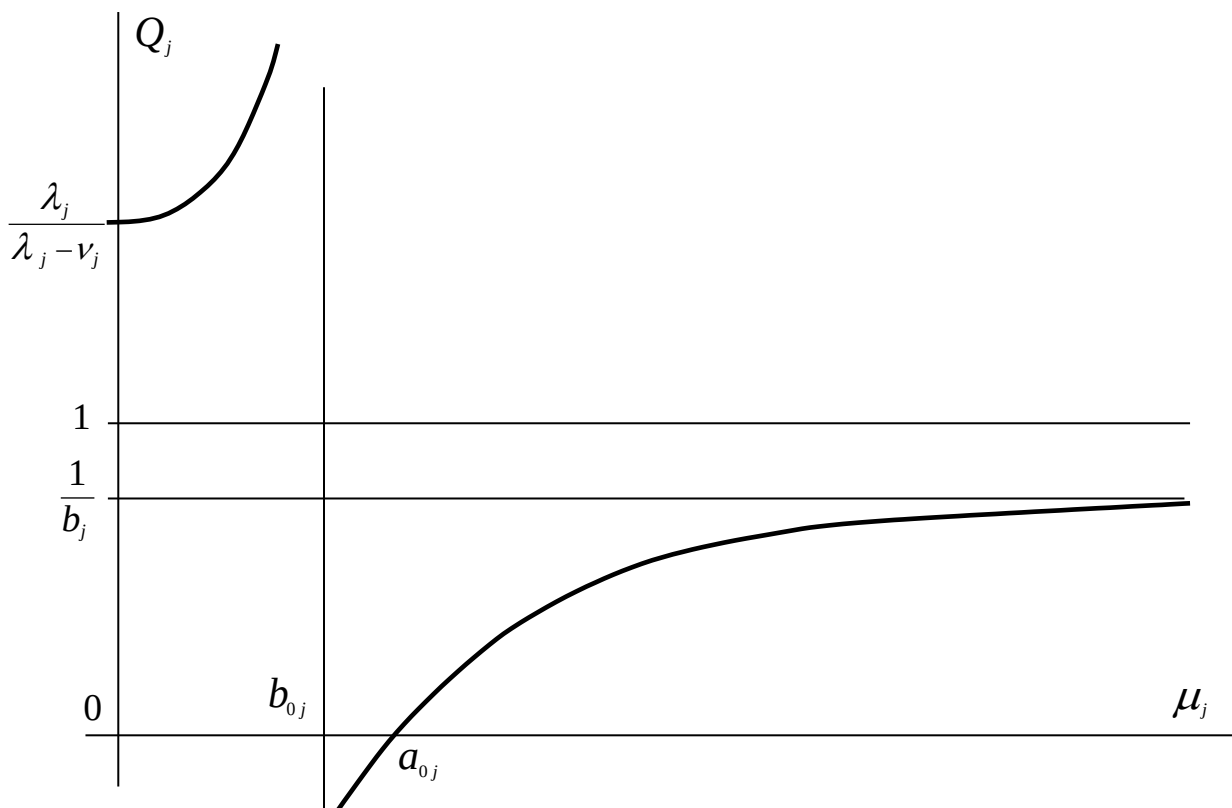


Рис 1.3. Зависимость вероятности своевременной доставки сообщений Q_j от пропускной способности μ_j

В частности, при $\mu_j \in b_{0j}, a_{0j}$ значения вероятностей Q_j отрицательны, что, конечно, тоже абсурдно. Только в области $\mu_j \geq a_{0j}$ значения имеют смысл, поэтому в формулировку задачи обязательно нужно вводить условия

$$\mu_j \geq a_{0j} = +\frac{\lambda_j}{k_{ej}}, j = 1, n \quad (1.36)$$

1.2.2. Метод получения оптимального решения

С учетом условий (1.36) и с введением новых переменных $x_j = \mu_j - a_{0j}, j = 1, n$ задача приобретает такую формулировку.

Задача 1.1а

$$Q \ x = Q_{\max} - \sum_{j=1}^n \frac{p_j c_{0j}}{p_j x_j + c_{0j}} \Rightarrow \max \{ x \}_j \quad (1.37)$$

$$\sum_{j=1}^n s_j x_j \leq S_{30} = S_{10} - \sum_{j=1}^n s_j a_{0j} \quad (1.38)$$

$$x_j \leq 0, j = 1, n \quad (1.39)$$

Отличие от предыдущей формулировки (задачи 1.1) — в наличии условий (1.39). Отличие принципиальное, поскольку при выполнении этих условий всегда, для любого допустимого, а следовательно, и оптимального решения будет справедливо

$$Q_j \in 0,1, j = 1, n \quad (1.40)$$

как и должно быть. Кроме того, при решении сформулированной задачи нужно пользоваться полной, а не редуцированной

формулировкой основной теоремы математического программирования, т. е. исходить из следующих условий для функции Лагранжа Φ

$$\begin{aligned} \frac{\partial \Phi}{\partial x_j} &\leq 0, \text{ если } x_j = 0, \\ \frac{\partial \Phi}{\partial x_j} &= \overline{0}, \text{ если } x_j > 0. \end{aligned} \quad (1.41)$$

Отсюда получаем

$$\begin{aligned} \frac{\partial \Phi}{\partial x_j} &= \frac{p_j^2 c_{0j}}{(p_j x_j + c_{0j})^2} - \psi \leq 0, \text{ если } x_j = 0, \\ \frac{\partial \Phi}{\partial x_j} &= \frac{p_j^2 c_{0j}}{(p_j x_j + c_{0j})^2} - \psi = 0, \text{ если } x_j > 0. \end{aligned} \quad (1.42)$$

или

$$\begin{aligned} \frac{p_j^2 c_{0j}}{(p_j x_j + c_{0j})^2} &\leq \psi, \text{ если } x_j = 0, \\ \frac{p_j^2 c_{0j}}{(p_j x_j + c_{0j})^2} &= \psi, \text{ если } x_j > 0. \end{aligned} \quad (1.43)$$

Введем величины

$$\alpha_j = \frac{p_j^2 c_{0j}}{(p_j x_j + c_{0j})^2} \quad (1.44)$$

Тогда из условий оптимальности (1.43) получаем

$$\begin{aligned} \alpha_j &\leq \psi, j \in Z_1 = \{j \mid x_j = 0\}, \\ \alpha_j &> \psi, j \in Z_2 = \{j \mid x_j > 0\}. \end{aligned} \quad (1.45)$$

$$Z_1 + Z_2 = Z = (1, 2, \dots, n)$$

Условия (1.45) являются ключом к решению задачи. Ясно, что параметр ψ (множитель Лагранжа) разделяет два подмножества множества значений α_j . Для одного из них любое значение α_j строго больше ψ , а соответствующие значения строго больше нуля; для другого значения α_j не больше ψ , а значения x_j равны нулю. Параметр

ψ , служит, таким образом, границей между этими подмножествами. По сути, исходную задачу можно формулировать как задачу отыскания такого деления множества $\{\alpha_j\}$, при котором выполняются условия (1.45). А это, в свою очередь, подсказывает такой путь решения: упорядочить величины в вариационный ряд по невозрастанию (убыванию), перенумеровать каналы (тракты) в соответствии с расположением α_j в этом ряду

$$a_1 \geq a_2 \geq \dots \geq a_n \quad (1.46)$$

(в соотношениях (1.46) и везде далее, если специально не оговорено другое, используются новые номера каналов), и искать затем интервал значения содержащий параметр

$$a_r > \psi \geq a_{r+1} \quad (1.47)$$

При этом, как следует из (1.43) и (1.45),

$$x_j = \frac{c_{0j}}{p_j} \left(\frac{\overline{a_j}}{\psi} - 1 \right), j = 1, r, \quad (1.48)$$

$$0, j = r + 1, n$$

Умножая обе части (1.48) на s_j и суммируя по всем j с учетом условия (1.38) получаем выражение для определения ψ

$$\frac{1}{\psi} = \frac{\sum_{j=1}^r \frac{p_j}{s_j c_{0j}} \overline{a_j}}{\sum_{j=1}^r \frac{p_j}{s_j c_{0j}}} \quad (1.49)$$

Соответственно для μ_j получаем

$$\mu_j = b_{0j} = \frac{c_{0j}}{a_{0j}} \frac{p_j}{\overline{a_j \psi}}, j = 1, r, \quad (1.50)$$

$$a_{0j}, j = r + 1, n$$

Для отыскания решения, таким образом, осталось указать правило нахождения индекса r . Обратим внимание в связи с этим на то, что, согласно соотношениям (1.47), α_r и α_{r+1} являются соответственно верхней и нижней оценками параметра ψ . Это означает, что, подставив в (1.48) вместо ψ значения α_r и α_{r+1} , мы получим соответственно нижнюю и верхнюю оценки x_j . Суммируя произведения этих оценок на s_j , равно как и $s_j x_j$, по всем j получаем фундаментальное соотношение для отыскания значения индекса r :

$$H_r < S_{30} \leq H_{r+1}, \quad (1.51)$$

где

$$H_j = \sum_{j=1}^r \frac{p_j}{\alpha_j} \frac{\alpha_j}{\alpha_{r-1}} - \frac{1}{1}, H_n = 0, r = 1, n. \quad (1.52)$$

Величины H_r в целом характеризуют задачу и множество решений. Если выделенный ресурс S_{30} лежит в диапазоне $(H_r, H_{r+1}]$, то независимо от его конкретного значения ресурс распределяется по первым r трактам, а остальные $n - r$ трактов остаются на уровне минимально допустимых значений пропускной способности a_{0j} . Это означает, что по мере возрастания ресурса S_{30} последовательно развиваются первый, второй и последующие тракты, пока при

$$S_{30} > H_n = \sum_{j=1}^n \frac{p_j}{\alpha_j} \frac{\alpha_j}{\min \alpha_j} - 1, \quad (1.53)$$

не окажется $r = n$ и в распределение не будут включены все тракты, т. е. все x_j будут положительными и все μ_j будут больше a_{0j} .

Заметим теперь, что если для какого-нибудь тракта $x_j = 0$, т.е. $\mu_j = a_{0j}$, $\mu_{\Sigma j} = \lambda_j$, то вероятность своевременной доставки сообщения по этому тракту

$$Q_j \mu_j = \frac{\mu_j - a_{0j}}{b_j (\mu_j - b_{0j})}. \quad (1.54)$$

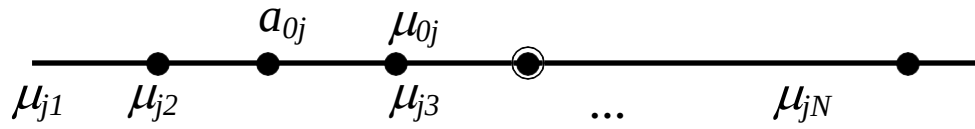
будет равна нулю. С точки зрения $Q \mu$ это является недопустимым. Но среднее время доставки сообщения при этом неограниченно возрастает

$$\lim_{\mu_j \rightarrow a_{0j}} T_j \mu_j = \infty \quad (1.55)$$

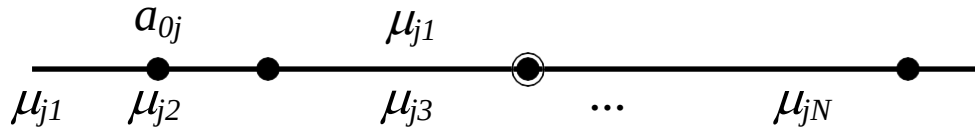
что, конечно, лишено смысла – такой тракт бесполезен, затраты на создание пропускной способности $\mu_j = a_{0j}$ бессмысленны, тракт практически не функционирует. В действительности должны выполняться строгие неравенства

$$\mu_j > a_{0j}, j = 1, n \quad (1.56)$$

Формализовать эти соотношения, придать им алгоритмически реализуемый вид помогают следующие дополнительные соображения. В реальных задачах пропускные способности не могут выбираться произвольно.



$$a) \mu_{j2} < a_{0j} < \mu_{j3}, \mu_{0j} = \mu_{j3}$$



$$б) \mu_{j2} = a_{0j}, \mu_{0j} = \mu_{j3}$$

Рис. 1.4. Схема выбора нижней границы μ_{0j} возможных значений пропускной способности μ_j .

Существует дискретный набор значений допустимых скоростей передачи R_k , $k = 1, N$ (N - количество типов каналов), и следовательно, дискретное множество допустимых значений пропускных способностей для каждого тракта

$$\mu_{jk} = \frac{R_k}{l_j}, k = 1, N \quad (1.57)$$

где l_j - средняя длина сообщения (кадра, пакета и т. п.) в j -м тракте.

Объединение условий (1.56) и (1.57) приводит к такому (алгоритмическому) ограничению снизу на значения μ_j . (рис. 1.4,а и б):

$$\mu_j \geq \mu_{0j}, j = 1, n \quad (1.58)$$

где

$$\mu_{0j} \geq \mu_{j,k(j)}, \mu_{j,k(j)-1} \leq a_{0j} < \mu_{j,k(j)}. \quad (1.59)$$

Сверху допустимые значения пропускных способностей μ_j ограничены предельно возможными величинами μ_{jN} , т.е.

$$\mu_j \leq \mu_{jN}, j = 1, n \quad (1.60)$$

Вводя новые переменные

$$x_j = \mu_j - \mu_{0j}, j = 1, n \quad (1.61)$$

а также обозначения

$$b_{1j} = \mu_{0j} - b_{0j}, a_j = \mu_{jN} - \mu_{0j}, S_{20} = S_{10} - \sum_{j=1}^n s_j \mu_j, \quad (1.62)$$

получаем окончательно интересующую нас оптимизационную задачу.

Задача 1.2

$$Q \ x = Q_{\max} - \sum_{j=1}^n \frac{c_{0j}}{x_j + b_{1j}} \Rightarrow \max \{ x_j \} \quad (1.63)$$

$$\sum_{j=1}^n s_j x_j \leq S_{20} \quad (1.64)$$

$$0 \leq x_j, \leq a_j, j = 1, n \quad (1.65)$$

Задача 1.2 существенно сложнее задачи 1.1 из-за двусторонних ограничений на неизвестные (1.65). Однако и здесь оказывается эффективным применение основной теоремы математического программирования. Функция Лагранжа в данном случае выглядит так:

$$\Phi \ \mu, \psi, \{\xi\} = Q \ x + \psi \ S_{20} - \sum_{j=1}^n s_j x_j + \sum_{j=1}^n \xi_j (a_j - x_j), \quad (1.66)$$

Для оптимальности решения необходимо существование числа ψ и вектора ξ таких, что выполняются следующие условия:

$$\begin{aligned} \frac{\partial \Phi}{\partial x_j} &\leq 0, \text{ если } x_j = 0, \quad \xi_j \geq 0, \text{ если } x_j = a_j, \\ \frac{\partial \Phi}{\partial x_j} &= 0, \text{ если } x_j > 0, \quad \xi_j = 0, \text{ если } x_j > a_j. \end{aligned} \quad (1.67)$$

Отсюда

$$\begin{aligned} \frac{c_{0j}}{s_j(x_j + b_{1j})^2} - \frac{\xi_j}{x_j} &\leq \psi, \ x_j = 0, \\ \frac{c_{0j}}{s_j(x_j + b_{1j})^2} - \frac{\xi_j}{x_j} &= \psi, \ x_j > 0, \end{aligned} \quad (1.68)$$

Следовательно, при $0 \leq x_j, \leq a_j$, когда $\xi_j = 0$, справедливо

$$\frac{c_{0j}}{s_j(x_j + b_{1j})^2} = \psi. \quad (1.69)$$

Из этого равенства вытекают два строгих неравенства

$$\frac{c_j}{s_j b_{1j}^2} = \beta_j > \psi \text{ и } \frac{c_j}{s_j (a_j + b_{1j})^2} = g_j < \psi. \quad (1.70)$$

Если $x_j = a_j$, то $\xi_j \geq 0$ и из (44) следует, что

$$\frac{c_{0j}}{s_j(x_j+b_{1j})^2} = \frac{c_{0j}}{s_j(a_j+b_{1j})^2} = g_j < \psi. \quad (1.71)$$

Наконец, при $x_j = 0$ справедливо $\xi_j = 0$ и из (1.68) находим:

$$\beta_j \leq \psi. \quad (1.72)$$

Объединяя (1.70), (1.71) и (1.72), получаем

$$\psi \begin{cases} \leq g_j, \text{ если } x_j = a_j \\ > g_j, \text{ если } 0 < x_j < a_j \\ \geq \beta_j, \text{ если } x_j = 0 \\ < \beta_j, \text{ если } 0 < x_j < a_j \end{cases} \quad (1.73)$$

Таким образом, ψ в данном случае является двойным сечением.

Во-первых, справедливо

$$\begin{aligned} g_j &\geq \psi > g_s, j \in Z_1 = \{j \mid x_j = a_j\} \\ s \in Z - z_1 &= \{1, 2, \dots, n\} - z_1 = z_2 + z_3 \\ z_2 + z_3 &= \{j \mid 0 < x_j < a_j\} \end{aligned} \quad (1.74)$$

Это означает, что ψ разделяет множество трактов на два подмножества. Если тракты перенумерованы в порядке невозрастания (убывания) величин g_j

$$g_1 \geq g_2 \geq \dots \geq g_l \geq \psi > g_{l+1} > \dots \geq g_n. \quad (1.75)$$

то для каналов с номерами $1, 2, \dots, l$ справедливо

$$x_j = a_j, j = 1, l \quad (1.76)$$

Для остальных $n - l$ трактов заведомо $x_j < a_j$ нужно решать задачу 1.2 без учета условий $x_j \leq a_j$, т. е. задачу типа 1.1а. Здесь

$$\begin{aligned} \beta_k &> \psi \geq \beta_r, k \in z_2 = \{j \mid 0 < x_j < a_j\}, \\ r \in z_3 &= \{j \mid x_j = 0\} \\ z_1 + z_2 + z_3 &= Z = \{1, 2, \dots, n\} \end{aligned} \quad (1.77)$$

Таким образом,

$$x_j = \begin{cases} a_j, j \in Z_1 = 1, 2, \dots, l \\ b_1 \frac{\beta_j}{\psi} - 1, j \in Z \text{ и } \psi < \beta_j \\ 0, j \in Z_1 \text{ и } \psi \geq \beta_j \end{cases} \quad (1.78)$$

Значения g_{l+1} и g_l могут служить соответственно нижней и верхней оценками для параметра ψ . Проводя рассуждения, аналогичные приведенным выше при исследовании задачи 1.1а, находим, что для отыскания значения индекса l необходимо выяснить, в какой интервал попадает величина ресурса S_{20} :

$$K_l \leq S_{20} \leq K_{l+1} \quad (1.79)$$

где

$$K_l = \sum_{j=1}^n s_j a_j + \sum_{\substack{j=l+1 \\ g_l < \beta_j}}^n s_j b_{1j} \frac{\beta_j}{g_l} - 1, j = 1, n \quad (1.80)$$

Если окажется, что $S_{20} < K_1$, то $l = 0$, т. е. условия $x_j \leq a_j$ априорно выполняются для всех трактов и можно решать задачу без их учета. В частности, это означает, что по приводимому ниже алгоритму можно решать и задачи с односторонними ограничениями, когда условия (1.65) выглядят так:

$$x_j \geq 0, j = 1, n \quad (1.81)$$

Для этого достаточно задать большие значения $a_j \rightarrow \infty$ что автоматически приведет к $K_l \rightarrow \infty$ и, следовательно, к $S_{20} < K_1$. Если $S_{20} \geq K_n$, то $l = n$, все $x_j = a_j$. Если же $K_l \leq S_{20} < K_{l+1}$, то для части объектов ($j = 1, l$) решение выписывается непосредственно ($x_j = a_j$), а для остальных решается задача 1.2 без учета условий $x_j \leq a_j$,

Из анализа задачи следует, что объем выделяемых ресурсов S_0 не может быть меньше $S_{0min} = \sum_{j=1}^r (s_{jj} \mu_{0j} + s_{0j})$ бы обеспечить хотя бы минимальные потребности сети. С другой стороны, бесполезно

выделять ресурсы в объеме, превышающем $S_{0max} = r_{=1}(s_j \mu_{jN} + \xi_{0j})$ поскольку предполагается, что пропускные способности каналов не могут превышать значений μ_{jN} .