УДК 681.324

РОЛИК А.И.

ДЕКОМПОЗИЦИОННО-КОМПЕНСАЦИОННЫЙ ПОДХОД К УПРАВЛЕНИЮ УРОВНЕМ УСЛУГ В КОРПОРАТИВНЫХ ИТ-ИНФРАСТРУКТУРАХ

Предложен декомпозиционно-компенсационный подход к управлению уровнем ИТ-услуг в корпоративных ИТ-инфраструктурах, предполагающий декомпозицию задач управления уровнем услуг и компенсацию негативного влияния различных факторов выделением дополнительных ресурсов критичным приложениям. Подход основан на интегрированном взаимодействии трех иерархических процессов — согласования уровня услуг, планирования ресурсов и управления уровнем услуг с учетом иерархии ИТ-инфраструктуры.

Decomposition-compensation method of service level management of corporate IT infrastructures was proposed. The method based on the interaction of three integrated hierarchical processes — matching the level of services, resource planning and service level management.

Введение

Бизнес рассматривает информационные технологии (ИТ) в качестве средства повышения своей производительности и улучшения конкурентоспособности. Эффективность выполнения бизнес-процессов существенно зависит от качества ИТ-услуг. Увеличивающееся количество ИТ-услуг, необходимых для автоматизации бизнес-технологий, усложнение приложений и возрастающее количество компонентов ИТ-инфраструктуры приводят к снижению эффективности работы ИТ-подразделений и повышению затрат на поддержание штатного режима функционирования ИТ-инфраструктуры. Обслуживание ИТ-услуг, внесенных в каталог, осуществляет ИТ-подразделение. Предоставление ИТ-услуг регламентируется пакетом соглашений об уровне сервиса (SLA), заключенных между бизнес-подразделениями и ИТ-подразделением. В SLA определяются значения ключевых показателей эффективности (KPI) и качества (KQI) — ограниченный набор объективно измеряемых параметров, позволяющий оценить качество ИТ-услуг [1]. Для поддержания значений KPI и KQI на уровне, зафиксированном в SLA, администраторы обеспечивают непрерывное функционирование ИТ-инфраструктуры, осуществляют обслуживание и ремонт с использованием методов автоматического, автоматизированного и ручного управления

Для повышения производительности ИТ-инфраструктуры и автоматизации процессов обслуживания создаются ИТ управления ИТ-инфраструктурой (УИТИ), на основе которых разрабатываются системы управления (СУ) ИТ-инфраструктурой [2, 3]. Увеличение бизнес-спроса на ИТ-услуги, многообразие ИТ, а также постоянное совершенствование бизнес-процессов и вызванная этим необходимость разработки и внедрения новых ИТ приводят к чрезмерному усложнению СУ, являющейся продуктом системной интеграции многообразных подходов и порой несовместимых решений от различных производителей. Возрастающая сложность управления ИТ-инфраструктурой, сопровождающаяся увеличением затрат на операционную деятельность, вызывает необходимость поиска новых подходов к управлению ИТ-инфраструктурой.

В настоящее время развитие ИТ находится под сильным влиянием таких факторов, как консолидация и виртуализация ресурсов, прогресс в области облачных вычислений, конвергенция телекоммуникационных технологий и услуг. В информационной индустрии под воздействием этих факторов осуществляется активный процесс глобализации информационных и телекоммуникационных технологий, формируется новая ИТ-среда, предоставляющая перспективные средства ведения бизнеса. Революционные преобразования в области ИТ способствуют прогрессу бизнес-технологий, однако эффективность применения ИТ сдерживает отставание в области технологий управления ИТ-инфраструктурой.

Высокая стоимость владения полнофункциональной ИТ-инфраструктурой корпоративного уровня и существенная зависимость успешности бизнеса от качества ИТ-услуг делают актуальной научно-прикладную проблему создания информационной технологии управления кор-

поративной ИТ-инфраструктурой, нацеленной на поддержание согласованного уровня ИТ-услуг при рациональном использовании ресурсов в условиях их виртуализации, кластеризации и консолидации при существенной динамике запросов пользователей.

Постановка проблемы в общем виде

Существенное влияние ИТ на достижение целей бизнеса не только подчеркивает значимость ИТ-сервисов, но и акцентирует необходимость управления ими. Лидирующее положение в области управления ИТ-сервисами принадлежит ITSM [4, 5] — признанному во всем мире и повсеместно реализуемому подходу к управлению ИТ, о чем свидетельствует появление международного стандарта ISO/IEC 20000 [6, 7]. Стандарт ISO/IEC 20000, являясь первым международным стандартом управления ИТ-услугами, содержит: требования к руководству, документированию, компетенции, осведомленности и подготовке персонала; требования к мониторингу, измерению, оценке и улучшению процессов; принципы формирования плана управления услугами и порядка применения цикла Деминга к управлению ИТ-услугами. Стандарт внес изменения и дополнения в процессную модель, перейдя от набора процессов к созданию целостной системы управления ИТ-услугами. Выделяется 13 процессов, разделенных на пять групп, и две области управления верхнего уровня.

Требования к управлению ИТ-услугами и к системе управления определены в ISO/IEC 20000-1, а рекомендации по организации деятельности по управлению ИТ-услугами — в ISO/IEC 20000-2. В соответствии со стандартом СУ, включающая политики и структурированный подход, должна реализовывать внедрение и эффективное управление всеми ИТ-услугами. При этом рассматривается только процессное управление, а вопросы оперативного управления ИТ-инфраструктурой, состояние и функционирование которой оказывает самое непосредственное влияние на уровень услуг, не рассматриваются.

Среди процессов ITSM и ISO/IEC 20000 отсутствуют процессы и область оперативного управления уровнем ИТ-услуг и ИТ-инфраструктурой. Это объясняется тем, что управление ИТ-услугами — сравнительно новая и активно развивающаяся область управления. Возможность и необходимость управления ИТ-услугами уже признана, о чем свидетельствует появление стандарта, в то время как рассмотрение ИТ-инфраструктуры в качестве объекта управления при менеджменте ИТ-услуг до сих пор в полной мере не осознано.

В то же время рассмотрение ИТ-инфраструктуры как объекта управления для поддержания качества предоставления ИТ-услуг на приемлемом уровне при изменении состояния компонентов ИТ-инфраструктуры и с учетом динамики запросов пользователей не только возможно, но и необходимо.

Таким образом, вопросы процессного управления уровнем услуг хорошо проработаны и стандартизованы, а вопросам управления уровнем услуг посредством оперативного управления ИТ-инфраструктурой до сих пор не уделялось достаточного внимания. Поэтому в данной статье, учитывая важность для бизнеса получения без перебоев ИТ-услуг со стабильно высоким качеством, внимание уделено именно этому аспекту управления ИТ-инфраструктурой.

Целью данной работы является разработка такого подхода к управлению ИТ-инфраструктурой, чтобы качество предоставляемых ею ИТ-услуг соответствовало заданному, согласованному с бизнес-подразделением, уровню, при рациональном использовании ресурсов.

Постановка задачи управления ИТ-инфраструктурой при управлении уровнем услуг

Для такого объекта управления, как ИТ-инфраструктура, очень сложно определить и формализовать единую задачу управления. Поэтому производится декомпозиция общей задачи управления ИТ-инфраструктурой [8], заключающейся в выборе такого допустимого управления, которое максимизирует значение эффективности управления $(\mathrm{K}(U) \to \max\limits_{U \in \mathcal{U}})$, на отдельные задачи с последующей их формализацией.

ИТ-инфраструктура предназначена для предоставления пользователям ИТ-сервисов. Эффективность управления в этом случае может быть оценена по качеству Q предоставляемых сервисов и затратам на управление. При оперативном управлении ИТ-инфраструктурой задача управления качеством сервисов состоит в поддержании заданного уровня качества сервисов с минимальным количеством используе-

мых для этого ресурсов. Тогда максимальная эффективность управления достигается выбором такого управления, при котором фактический уровень сервисов соответствует согласованному с бизнес-подразделением Q_{cor} и достигается с минимальными затратами:

$$\text{Затраты}(Q(U) \to Q_{\text{cor}}) \to \min_{U \in \mathcal{U}}. \quad (5)$$

Качество Q сервисов определяется качеством $Q_j, j = \overline{1, N}$ всех ИТ-услуг:

$$Q = f(Q_1, ..., Q_N), \quad (6)$$

следовательно, управляющие воздействия должны поддерживать заданный уровень качества каждой услуги с использованием для этих целей минимального количества ресурсов:

$$\text{Затраты}(Q_j(U) \to Q_j^{\text{cor}}) \to \min_{U \in \mathcal{U}}, \quad \forall j = \overline{1, N}, \quad (7)$$

где $Q_j^{\text{cor}}, j = \overline{1, N}$ — согласованный уровень j-й услуги.

В отличие от процессного управления, перед которым стоит задача постоянного повышения качества услуг, оперативное управление нацелено на поддержание наиболее дешевым способом качества услуг на согласованном уровне, при этом управление должно быть таким, чтобы обеспечивалось

$$q_{k,j} - q_{k,j}^{*} \to 0, \forall j, k, \quad (8)$$

где $q_{k,j}$ и $q_{k,j}^{*}$ — соответственно, целевое и фактическое значения k-го показателя качества j-й услуги.

С точки зрения бизнеса критерием эффективности управления ИТ-инфраструктурой при обеспечении качества j-й услуги может быть выбор такого управления $U \in \mathcal{U}$, при котором достигается минимальное фактическое время обработки i-го запроса к приложению $A_j, j = \overline{1, N}$, N — количество приложений

$$\min_{\forall i, j}(T_{\Phi_{i,j}} = (t_{R_{i,j}} - t_{A_{i,j}})), \quad (9)$$

где $t_{A_{i,j}}$ — время поступления i-го запроса от пользователя j-й услуги, $t_{R_{i,j}}$ — время поступления ответа пользователю на i-й запрос к приложению A_j.

Что касается качества предоставления услуг, то ИТ УИТИ должна осуществлять управление, руководствуясь не критерием (9), поскольку невозможно достичь времени обработки запросов, равного 0, а стремиться минимизировать разницу между заданным $T_{З_j}$ и фактическим T_{Φ_j} временем выполнения запросов к j-му приложению, измеренным на стороне пользователя, используя наименьшее количество ресурсов

$$\min_{\forall j}(T_{З_j} - T_{\Phi_j}) = \min_{\forall j}(\Delta T_j), \text{ при } T_{\Phi_j} > T_{З_j}. \quad (10)$$

В случае, когда $T_{\Phi_j} - T_{З_j} > 0$, причиной чему может быть увеличение количества запросов пользователей, выполнение критерия (9) или (10) возможно путем выделения приложению A_j дополнительных информационно-вычислительных ресурсов и/или приоритетного прохождения данных пользователей приложения A_j по телекоммуникационной сети.

Подход к оперативному управлению уровнем ИТ-услуг

Суть предлагаемого подхода к оперативному управлению уровнем услуг. Основная цель бизнеса — получение максимальной прибыли. Максимальная прибыль за счет ИТ достигается тогда, когда бизнесу предоставляется множество $S = \{s_i\}$, $i = \overline{1, K}$ необходимых ИТ-услуг с максимальным качеством Q и минимальными затратами C.

Управление уровнем услуг в корпоративных ИТ-инфраструктурах предлагается осуществлять интегрированным взаимодействием трех процессов: согласования уровня услуг, планирования ресурсов и управления уровнем услуг (рис. 1).

Процесс согласования уровня услуг запускается по инициативе бизнес-менеджеров и заканчивается формированием или обновлением элементов множества S и матрицы $Q = \|q_{ki}\|$, элемент q_{ki}, $k = \overline{1, M_i}$, $i = \overline{1, K}$ которой соответствует согласованному значению k-го показателя качества i-й услуги. Бизнес выделяет услугам S количество r ресурсов. Полученное в результате ресурсное обеспечение в виде системы

$$< Q, r > \quad (11)$$

является основой для решения задач на нижерасположенном уровне.

Процесс планирования заключается в выделении и закреплении за каждой услугой s_i, $i = \overline{1, K}$ части из ресурсов $R_1, ..., R_m$ ИТ-инфраструктуры, выделенных для поддержания услуг, при этом $r_1, ..., r_m$ — количество ресурсов $R_1, ..., R_m$, а $c_1, ..., c_m$ — стоимость единицы ресурса $R_1, ..., R_m$ соответственно.

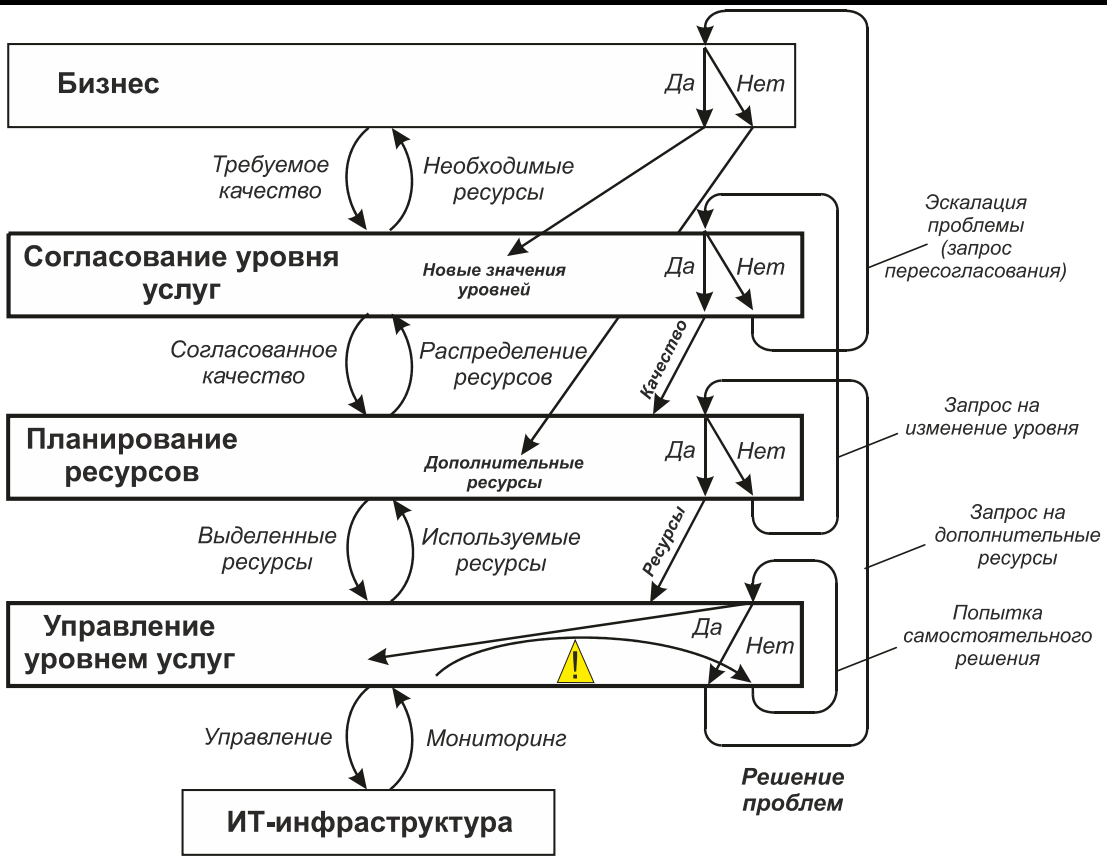


Рис. 1. Взаимодействие процессов при управлении уровнем услуг

Тогда количество всех ресурсов вычисляется как:

$$r = \sum_{j=1}^{m} r_j \,, \qquad (12)$$

а стоимость c ресурсов определяется следующим образом:

$$c = \sum_{j=1}^{m} r_j \cdot c_j \,. \qquad (13)$$

Использование услугами ресурсов задается матрицей $\mathrm{P} = \left\| \rho_{ij} \right\|$, где ρ_{ij} равно количеству выделенного услуге s_i ресурса R_j, $j = \overline{1,m}$ или 0, если ресурс не требуется.

Процесс управления уровнем услуг осуществляет управление ИТ-инфраструктурой так, чтобы фактические значения q_{ki}^*, $k = \overline{1,M_i}$, $i = \overline{1,K}$ показателей качества услуг соответствовали согласованным значениям из матрицы Q, т. е. чтобы выполнялось равенство

$$q_{ki} - q_{ki}^* = 0, \quad \forall k = \overline{1,M_i}, i = \overline{1,K} \,. \qquad (14)$$

Суть предлагаемого подхода к управлению уровнем услуг заключается в следующем.

При невыполнении условия (14) выявляются элементы матрицы фактических значений показателей качества $Q^* = \left\| q_{ki}^* \right\|$, для которых $q_{ki}^* < q_{ki}, \forall k = \overline{1,M_i}, i = \overline{1,K}$. ИТ УИТИ пытается решить проблему на нижнем уровне (см. рис. 1), изменяя значения параметров функционирования элементов ИТ-инфраструктуры либо перераспределяя ресурсы между приложениями так, чтобы увеличить значение q_{ki}^*. Если в результате восстановительных мероприятий удалось обеспечить выполнение равенства (14), то функционирование ИТ-инфраструктуры продолжается с новыми настройками. Если полномочий нижнего уровня недостаточно для достижения (14), то осуществляется эскалация проблемы на уровень планирования ресурсов.

В ходе процесса планирования ресурсов осуществляются попытки решить проблему выделением дополнительных ресурсов $R_1,...,R_m$ услуге s_i, для которой выполняется условие $q_{ki}^* < q_{ki}$. Если дополнительные ресурсы выделяются, то формируется матрица $\mathrm{P}' = \left\| \rho'_{ij} \right\|$ с новыми значениями элементов, причем $\rho'_{ij} > \rho_{ij}$, $j = \overline{1,m}$ или 0, если j-й ресурс не требуется. Если дополнительные ресурсы отсутствуют, то на уровне планирования ресурсов предпринимаются попытки произвести перераспределение ресурсов между услугами, отда-

вая ресурсы более важным услугам за счет менее важных. Если проблема решается, то значения матрицы $P' = \|\rho'_{ij}\|$ с новым планом закрепления ресурсов поступают на нижний уровень. При невозможности разрешения проблемы на уровне планирования ресурсов производится эскалация проблемы на вышерасположенный уровень.

Процесс согласования уровня услуг по инициативе процесса планирования осуществляет пересмотр сначала значения q_{ki}, для которого $q^*_{ki} < q_{ki}$, а затем, возможно, и значений всех элементов q_{ki}, $k = \overline{1, M_i}$, $i = \overline{1, K}$ матрицы качества услуг Q в сторону уменьшения. Если удается сформировать матрицу $Q' = \|q'_{ki}\|$ с новыми значениями показателей качества услуг, то она передается на уровень ниже, где производится высвобождение ресурсов и выделение их услугам, для которых выполняется условие $q^*_{ki} < q_{ki}$, $k = \overline{1, M_i}$, $i = \overline{1, K}$. Если процесс согласования уровня услуг не имеет полномочий для формирования матрицы $Q' = \|q'_{ki}\|$, то производится эскалация проблемы на уровень бизнеса, который вынужден либо сгенерировать матрицу $Q' = \|q'_{ki}\|$ с новыми значениями, либо увеличить общий объем ресурсов, что приводит к увеличению значений $r_1,...,r_m$, либо довольствоваться фактическим уровнем услуг.

Рассмотрим подробнее процессы, реализуемые при согласовании уровня услуг, планировании ресурсов и управлении уровнем услуг.

Процесс согласования уровня услуг в корпоративных ИТ-инфраструктурах. Требования максимизации ($\max Q$) качества услуг и минимизации связанных с этим затрат ($\min C$) находятся в естественном противоречии, что приводит к необходимости установления экономически обоснованного уровня услуг с учетом как возможностей компании, так и достигнутого отраслью уровня и ожиданий клиентов.

Если $\hat{D}$ — доходы бизнеса от ИТ, а $\hat{C}$ — затраты на ИТ-инфраструктуру, то бизнес совместно с процессом согласования уровня услуг пытается добиться

$$\max(\hat{D} - \hat{C}). \qquad (15)$$

Стремясь к постоянному повышению качества, бизнес- и ИТ-подразделения должны в рамках процессного управления ITSM выполнять аналитическую оценку зависимостей качества сервиса от стоимости ресурсов $\hat{Q} = f_1(c)$, потерь бизнеса $\hat{L}$ от некачественного сервиса с учетом рисков M_0 и неопределенностей N_0 (рис. 2), после чего с учетом уровня качества, достигнутого отраслью, определяются допустимые значения уровня услуг $Q = \|q_{ki}\|$, которые еще не приводят к потерям при выполнении бизнес-операций. Таким образом, система

$$< \hat{D}, \hat{C}, \hat{L}, \hat{Q}, M_o, N_0 > \qquad (16)$$

определяет задачу по формированию двойки (11).

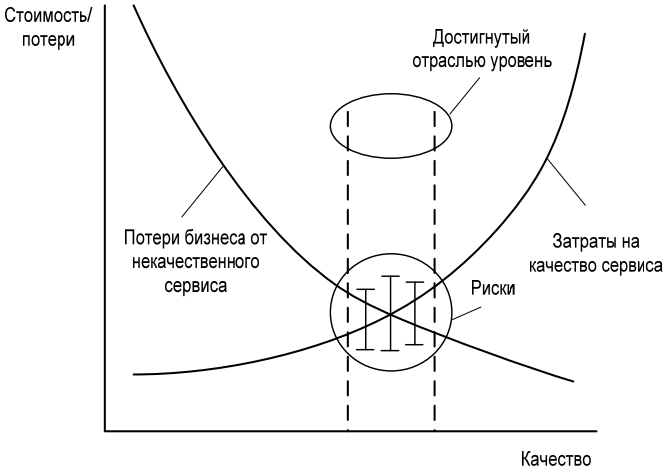


Рис. 2. Нахождение оптимального соотношения уровня качества сервиса и стоимости достижения этого качества

Если полученные с учетом рисков значения уровня услуг превышают достигнутый отраслью уровень, то есть повод для оптимизма и развития бизнеса. Иное говорит о проблемах и является основой для сворачивания бизнеса в соответствующих направлениях.

Бизнесу предоставляются варианты реализации услуги со стоимостными, техническими, организационными и временными показателями (рис. 3). Бизнес определяет оптимальное отношение доходность/(качество ИТ-услуг). Для этого оцениваются потери в зависимости от падения уровня услуг и стоимость предоставления услуг высокого качества, после чего находится точка минимально допустимого качества. Затем бизнес одобряет один из вариантов или корректирует требования к новой ИТ-услуге, исходя из значимости сервиса, бенчмаркинга, опыта и пр. Согласованные уровни услуг фиксируются в SLA или каталоге услуг.

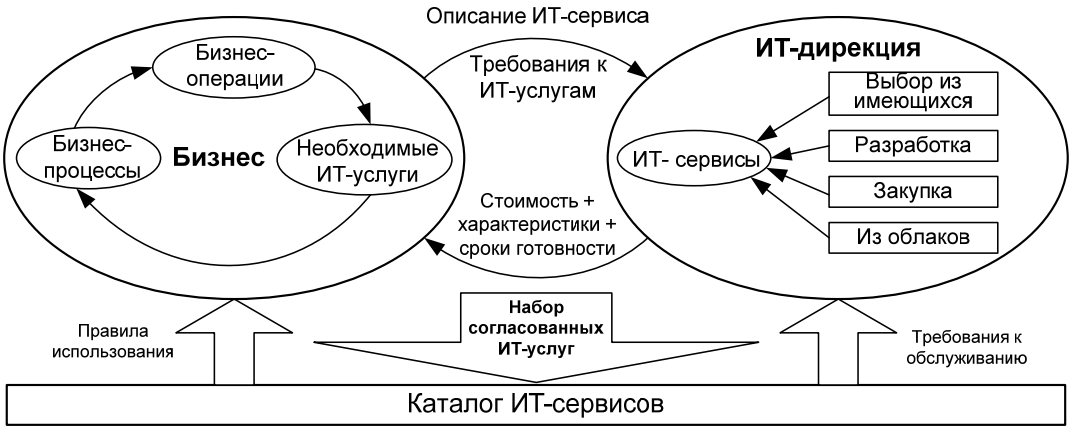


Рис. 3. Процесс формирования каталога ИТ-сервисов

План внедрения новых услуг ИТ-дирекция готовит с учетом финансового и ресурсного обеспечения их предоставления и управления. Эту задачу целесообразно решать с применением методов эффективного использования ресурсов ИТ-инфраструктуры, предложенных в [9—11].

Согласование уровня услуг производится в условиях избытка или дефицита ресурсов (см. ниже). При дефиците ресурсов запускается итеративная процедура согласования уровня услуг, когда на основании (12) процесс планирования определяет значения q_{ki}, $k=\overline{1,M_i}$, $i=\overline{1,K}$, затем полученные показатели в виде матрицы Q предъявляются бизнесу, и если значения элементов матрицы Q не удовлетворяют бизнес, то производится перераспределение ресурсов между услугами и запускается новый цикл определения значений q_{ki}, $k=\overline{1,M_i}$, когда для заданного плана закрепления ресурсов $R_1,...,R_m$ за услугами s_i, $i=\overline{1,K}$ определяется ожидаемое качество Q.

При решении задач согласования уровня услуг могут быть использованы методы математического программирования, исследования операций, теории принятия решений, методы теории игр, методы решения многокритериальных задач, принятия решений в условиях неопределенности и риска, методы искусственного интеллекта, балансовые модели и др.

Процесс планирования ресурсов. На уровне планирования ресурсов решаются задачи в интересах процесса согласования уровня услуг, производится определение необходимого количества ресурсов, распределение и закрепление ресурсов за услугами s_i, $i=\overline{1,K}$.

Услуги s_i, $i=\overline{1,K}$ поддерживаются приложениями из множества $A=\{A_l\}$, $l=\overline{1,I}$, где I — количество приложений, причем каждая услуга s_i, $i=\overline{1,K}$ поддерживается одним или несколькими приложениями, в то время, как каждое приложение A_l, $l=\overline{1,I}$ поддерживает одну или несколько услуг.

Введем понятие степени поддержки услуги. Под поддержкой услуги s_i будем понимать взаимодействие нескольких приложений из множества A, обеспеченных ресурсами, направленное на достижение единого результата — работоспособность услуги s_i. Пусть ресурсы R_j, $j=\overline{1,m}$ используются для поддержки услуг s_i, $i=\overline{1,K}$. Сначала введем в рассмотрение булевую переменную x_i, $i=\overline{1,K}$, определяющую степень поддержки i-й услуги и принимающую значения:

$$x_i=\begin{cases}1, \text{если } i\text{-я услуга поддерживается в полном объеме;}\\ 0, \text{если } i\text{-я услуга не поддерживается.}\end{cases} \tag{17}$$

Пусть пользователи услуги s_i формируют в среднем a_{li} запросов к приложению A_l за единицу времени. Тогда количество клиентских запросов приложения A_l определяется следующим образом:

$$a_l=\sum_{i=1}^{K} a_{li}\cdot x_i\,, \tag{18}$$

а количество запросов к приложениям A_l, $l=\overline{1,I}$ представим вектором $\hat{a}=\{a_l, l=\overline{1,I}\}$.

Потребности приложения A_l в ресурсах типа j определяются выражением:

$$r_{jl} = b_{jl}\, a_l + d_{jl}, \qquad (19)$$

где b_{jl} — среднее количество ресурсов типа j, используемых приложением A_l для обработки одного клиентского запроса; d_{jl} — количество ресурсов типа j, используемых приложением A_l независимо от количества клиентских запросов.

Тогда общее количество ресурсов r, необходимое для поддержания всех услуг из S , определяется с помощью выражения:

$$r = \sum_{j=1}^{m}\sum_{l=1}^{I} r_{jl} = \sum_{j=1}^{m}\sum_{l=1}^{I}(b_{jl}a_l + d_{jl}). \qquad (20)$$

Задачи планирования существенно зависят от ограничений на бюджет ИТ-подразделения. При отсутствии финансовых ограничений, когда критерий $\min C$ не учитывается, планируется и проектируется ИТ-инфраструктура, ресурсы r которой будут достаточными для поддержания приложений $\{A_l\}$ с требуемыми значениями показателей качества услуг q_{ki}, $k = \overline{1, M_i}$, $i = \overline{1, K}$ при максимальных допустимых значениях вектора $\hat{a}$. Кроме того, наиболее важные ресурсы дублируются или резервируются $(r_r > 0)$, а также учитывается возможное запредельное увеличение количества запросов a_{li}.

При ограниченном бюджете на создание и развитие ИТ-инфраструктуры дефицит ресурсов может закладываться уже на стадии проектирования или возникать в процессе эксплуатации, и сложившаяся ИТ-инфраструктура в принципе не может обеспечить выполнение условия (14). Несмотря на это, ИТ-инфраструктура, спроектированная с дефицитом ресурсов, способна эффективно предоставлять услуги, если априорно обеспечивается поддержка наиболее важных приложений. Для этого ИТ УИТИ перераспределяет ресурсы в процессе работы в соответствии с определенным регламентом использования ресурсов. Решению подобных задач посвящена работа [12].

Таким образом, задачи планирования и управления ресурсами необходимо решать в условиях как избытка, так и дефицита ресурсов. В каждом конкретном случае необходимо учесть некоторые специфические особенности распределения ресурсов.

После расчета потребности ресурса по выражению (20) и сравнения ее с имеющимися ресурсами получим задачу управления уровнем сервисов при дефиците ресурсов, если нештатные ситуации в ИТ-инфраструктуре, увеличение интенсивности клиентских запросов $\hat{a}$ и другие факторы приводят к невыполнению равенства (14).

В этом случае для принятия решений при распределении ресурсов необходимо учитывать дополнительную информацию.

Введем понятие важности w_i услуги s_i, $i = \overline{1, K}$, которое будем использовать при решении задачи согласования уровня услуг в условиях дефицита ресурсов.

Задача согласования уровня услуг при дефиците ресурсов сводится к определению значений матрицы Q при заданном количестве ресурсов:

$$Q = \mathcal{F}_1(S\ , r, W_p, Z_s), \qquad (21)$$

где $W_p = \{w_i \mid i = \overline{1, K}\}$; $Z_s = \{z_i \mid i = \overline{1, K}\}$; z_i — планируемая степень поддержки i-й услуги. Здесь нормативное значение количества клиентских запросов учитывается в величине r, а значение $z_i, i = \overline{1, K}$, в отличие от (17), является непрерывной переменной, принимающей значения из отрезка [0,1].

Если же после расчета потребности в ресурсах и ее сравнения с имеющимися ресурсами, последних оказывается больше, то получим задачу управления уровнем услуг при избытке ресурсов. В частности, может быть выделен резервный объем r_r ресурсов, и количество ресурсов $\hat{r}$, выделяемое для поддержания всех услуг из S , определяемое по выражению

$$\hat{r} = r + r_r, \qquad (22)$$

возрастает. Тогда объем ресурсов r определяется, исходя из значений элементов матрицы Q, а величина r_r — исходя из вероятности возникновения нештатных ситуаций и предельного значения величины a_{li}.

В этом случае выделяются две задачи согласования уровня услуг. Задача первого рода аналогична задаче (21) и состоит в определении значений показателей качества при известном объеме выделенных ресурсов:

$$Q = \mathcal{F}_2(S\ , \hat{r}, r_r, C). \qquad (23)$$

Задача второго рода состоит в определении необходимого объема ресурсов для обеспечения заданных значений показателей качества:

$$r = \mathcal{F}_3(S\ , Q, C). \qquad (24)$$

При решении задач (21), (23) и (24) могут быть использованы методы теории массового

обслуживания [13], теория надежности [14], теории фракталов, аналитическое и имитационное моделирование, в частности, с применением методов теории массового обслуживания и искусственного интеллекта.

Процесс управления уровнем услуг. После согласования уровня услуг и планирования ресурсов процесс управления осуществляется так, чтобы выполнялись критерии:

$$\min(q_{ki} - q_{ki}^*),\ k = \overline{1, M_i}, i = \overline{1, K},\ \text{при}\ q_{ki}^* < q_{ki} \quad (25)$$

или

$$\min \hat{C},\ \text{при}\ q_{ki}^* > q_{ki},\quad k = \overline{1, M_i}, i = \overline{1, K}. \quad (26)$$

В последнем случае для экономии затрат производится сокращение выделенных приложениям $\{A_l\}$ ресурсов, а незадействованные ресурсы отключаются. Такие задачи решаются в [9—11].

Критерии (25) и (26) применяются только тогда, когда $\forall k = \overline{1, M_i}$ и $\forall i = \overline{1, K}$, а при сравнении значений q_{ki}^* и q_{ki} выполняется условие «только не больше» или «только не меньше». В противном случае ресурсы между приложениями из $\{A_l\}$ могут быть перераспределены так, чтобы приложениям, для которых $q_{ki}^* < q_{ki}$, были выделены ресурсы за счет приложений, для которых $q_{ki}^* > q_{ki}$, $\forall k = \overline{1, M_i}$, $i = \overline{1, K}$.

Если средствами нижнего уровня не удается обеспечить равенство $q_{ki}^* = q_{ki}$ при $q_{ki}^* < q_{ki}$, то запускается итеративная процедура, в которой задействуются вышерасположенные уровни (см. рис. 1). В этом случае на вышерасположенных уровнях приложению $A_l^* \in A$, для которого $q_{ki}^* < q_{ki}$, дополнительно выделяется квант Δr ресурсов. После чего проверяется выполнение условия (14). Если по-прежнему $q_{ki}^* < q_{ki}$, то выделяется еще один квант ресурсов Δr. Процедура повторяется до тех пор, пока не начнет выполняться условие (14). При отсутствии в ИТ-инфраструктуре запаса ресурсов при управлении уровнем услуг возможны две ситуации:

1) пересматриваются значения матицы Q;

2) выделяются кванты ресурса Δr за счет приложений из множества $\{A_l\}$ с учетом важности W_p услуг.

Зависимость значений показателей качества q_{ki}, $\forall k = \overline{1, M_i}$, $i = \overline{1, K}$ от объема ресурсов r без потери общности можно представить в следующем виде:

$$q = f_{qr}(r), \quad (27)$$

где q — показатель качества услуги. Для увеличения значения q соответствующему приложению из множества $\{A_l\}$ необходимо выделить дополнительные ресурсы. Тогда

$$q' = f_{qr}(r + \Delta r). \quad (28)$$

Если $\Delta r > 0$, то $q' \geq q$, что позволяет сделать предположение о монотонном характере функции f_{qr}.

Аналогичным образом можно предположить, что функция

$$q = f_{qa}(\hat{a}) \quad (29)$$

также будет монотонной.

Тогда, если функции (27) и (29) монотонные, то функция

$$q = f_q(r, \hat{a}) \quad (30)$$

также будет монотонной [15].

Пусть управление $u^+ \in U$, где U — множество управляющих воздействий, заключается в выделении дополнительных ресурсов приложению $A_l^* \in A$, для которого фактическое качество $q_\text{ф}$ хуже целевого $q_\text{ц}$, $q_\text{ф} < q_\text{ц}$, а $u^- \in U$ — управление, изымающее ресурсы у приложения $A_l^* \in A$, если $q_\text{ф} > q_\text{ц}$.

Учитывая монотонный характер зависимости q_{ki}, $\forall k = \overline{1, M_i}$, $i = \overline{1, K}$ от r, докажем необходимые утверждения, предварительно сделав следующие выкладки.

Использование ресурсов приложениями зададим матрицей $\hat{\mathrm{P}} = \left\| \hat{\rho}_{lj} \right\|$, $l = \overline{1, I}$, $j = \overline{1, m}$, причем:

$$\hat{\rho}_{lj} = \begin{cases} n_{lj}\Delta r_j, \text{если } l\text{-е приложение использует } j\text{-й ресурс;} \\ 0, \text{если } l\text{-е приложение не использует } j\text{-й ресурс,} \end{cases} \quad (31)$$

где n_{lj} — количество квантов j-го ресурса, выделенного l-му приложению; Δr_j — размер кванта j-го ресурса. При этом должны выполняться ограничения:

$$\Delta r_j \sum_{l=1}^{I} n_{lj} \leq r_j, \quad \forall j = \overline{1, m}. \quad (32)$$

Тогда можно определить следующее отображение:

$$\hat{Q} = U \times \mathrm{P} \times \bar{\hat{a}} \to Q, \quad (33)$$

где $\bar{\hat{a}} = \{\hat{a}\}$ — множество векторов $\hat{a} \in \bar{\hat{a}}$.
В свою очередь :

$$\hat{\mathcal{U}} = Q^* \times \mathrm{P} \times \bar{\hat{a}} \to U. \quad (34)$$

Утверждение 1. Для заданных значений q_{ki}, $\forall k = \overline{1, M_i}, i = \overline{1, K}$ в случае, когда $q_{ki}^* < q_{ki}$, существует управление $u^+ \in U$, позволяющее обеспечить $q_{ki}^* = q_{ki}$ при $\min r$ на поддержание уровня услуг.

Доказательство следует из монотонности функций (27)—(30), конечности множеств Q^*, P и $\bar{\hat{a}}$ и сопоставимости целей процессов на рис. 1.

Управление $u^+ \in U$ находится итеративно.

Для доказательства следующего утверждения введем отображение:

$$\hat{\mathcal{F}} = Q \times \bar{\hat{a}} \to R. \quad (35)$$

Утверждение 2. При выполнении условия (22) и $q_{ki}^* < q_{ki}$, если известны значения $\hat{a}$, то управление $u^+ \in U$, позволяющее восстановить равенство $q_{ki}^* = q_{ki}$, $\forall k = \overline{1, M_i}, i = \overline{1, K}$, может быть найдено без использования итерационных процедур.

Доказательство. При фиксированных значениях вектора $\hat{a}$, исходя из (35), существуют зависимости $q_{\text{ц}} \to r$, $q_1 \to r_1$ и $q_2 \to r_2$. Тогда, по аналогии с $q_1 - q_2 \to r_1 - r_2 = \Delta r_{12}$, при $q_{ki}^* < q_{ki}$ для обеспечения выполнения равенства $q_{ki}^* = q_{ki}$ необходимо l-му приложению, поддерживающему i-ю услугу с учетом (31), выделить дополнительно $\Delta\hat{\rho}_{lj}, i = \overline{1, K}$, $l = \overline{1, I}$, $j = \overline{1, m}$, причем величина $\Delta\hat{\rho}_{lj}$ может быть определена на основе $\Delta q_{ki} = q_{ki} - q_{ki}^* \to \Delta\hat{\rho}_{lj}$. В общем случае выделение дополнительных ресурсов с учетом ограничения (32) без итераций может быть осуществлено только при выполнении условия (22), что и доказывает утверждение.

Следствие. Если $q_{ki}^* < q_{ki}$ и i-я услуга поддерживается приложением $A_l^* \in A$, а за счет приложений из множества A, для которых $q_{\text{ф}} > q_{\text{ц}}$, можно высвободить объем ресурсов $\Delta\hat{\rho}_{\text{в}}$, причем $\Delta\hat{\rho}_{\text{в}} \geq \Delta\hat{\rho}_{lj}$, где $\Delta\hat{\rho}_{lj}$ — дополнительный объем ресурсов, который нужно выделить l-му приложению для того, чтобы обеспечить выполнение $q_{ki}^* = q_{ki}$, то управление $u^+ \in U$ в условиях дефицита ресурсов позволяет восстановить уровень i-й услуги, поддерживаемой l-м приложением за один проход.

Реализацию нижнего уровня, на котором непосредственно осуществляется оперативное управление уровнем услуг, целесообразно осуществлять на основе координатора [15].

Исследование предложенного декомпозиционно-компенсационного подхода к управлению уровнем услуг не только подтвердило его работоспособность, но и выявило способность повышения эффективности использования ресурсов. Предлагаемый подход сравнивался с методом полного резервирования ресурсов и методом наращивания нод по эффективности использования ресурсов $q_{\text{Э}}$, определяемой, как:

$$q_{\text{Э}} = \frac{r_u}{r_{\text{в}}}, \quad (36)$$

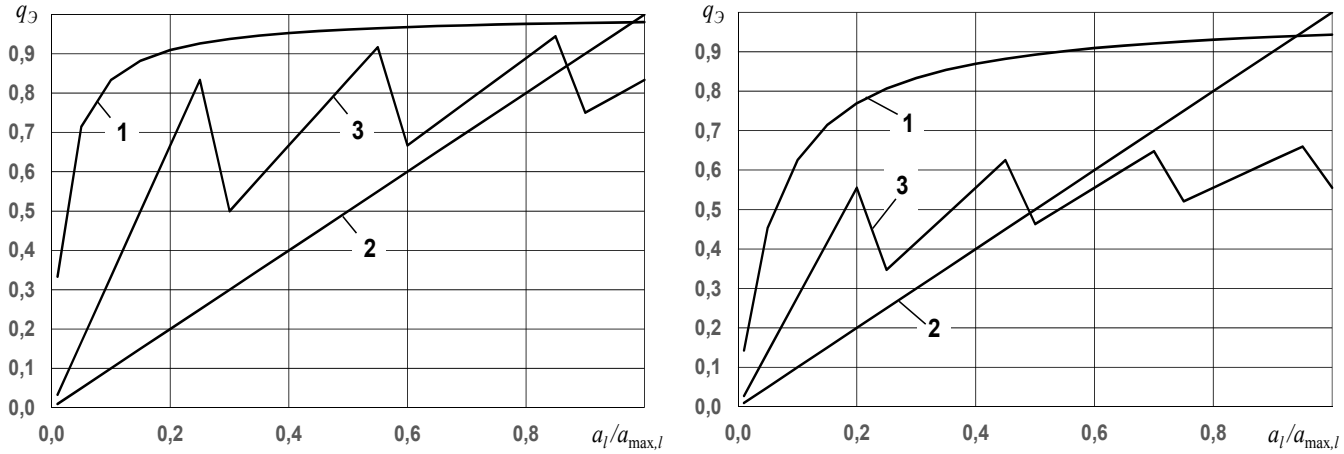
где r_u — количество фактически используемых ресурсов, а $r_{\text{в}}$ — количество зарезервированных и выделенных ресурсов.

Наиболее распространенный на практике метод полного резервирования предполагает определение количества ресурсов $r_{\max,l}$ по выражению (19) для максимального количества $a_{\max,l}$ пользователей l-й услуги, оговоренному в SLA. Объем ресурсов $r_{\max,l}$ закрепляется за приложением A_l и в процессе работы не изменяется. Если фактическое количество клиентских запросов $a_l < a_{\max,l}$, то ресурсы используются неэффективно. При $a_l > a_{\max,l}$ качество услуг снижается, управление для повышения уровня услуг не осуществляется, и пользователи довольствуются фактическим качеством услуг.

Метод поддержания качества услуг путем наращивания нод при горизонтальном масштабировании отслеживает не уровень услуг, а процент утилизации выделенных ресурсов. При превышении степени задействованности отдельных ресурсов предварительно заданного порогового значения осуществляется увеличение объема выделенных приложению ресурсов на размер ноды. При этом производится увеличение объема всех видов ресурсов независимо от реальной потребности услуг в увеличении

объемов только некоторых из них. При этом незадействованные ресурсы не могут быть использованы другими приложениями.

Зависимость эффективности использования ресурсов $q_Э$ от соотношения $a_l / a_{\max,l}$ приведена на графиках рис. 4 для предлагаемого подхода (кривая 1), метода полного резервирования ресурсов (кривая 2) и метода наращивания нод (кривая 3) при разном количестве d_l ресурсов, используемых приложением A_l, независимо от количества клиентских запросов. Для проведения исследований величина кванта ресурсов устанавливалась на уровне 10% от размера ноды.



а) *б*)

Рис. 4. Зависимости $q_Э$ от $a_l / a_{\max,l}$ для случаев: а) пропорциональных и б) непропорциональных потребностей в ресурсах

Из анализа графиков на рис. 4 видно, что предлагаемый подход гораздо эффективнее использует ресурсы корпоративной ИТ-инфраструктуры, причем эффективность использования ресурсов увеличивается по мере уменьшения значения соотношения $a_l / a_{\max,l}$.

Выводы

Эффективное управление уровнем услуг в корпоративных ИТ-инфраструктурах возможно при условии применения предложенного декомпозиционно-компенсационного подхода, предполагающего декомпозицию задач управления уровнем услуг и компенсацию негативного влияния различных факторов выделением дополнительных ресурсов критичным приложениям. Подход основан на интегрированном взаимодействии трех иерархических процессов — согласования уровня услуг, планирования ресурсов и управления уровнем услуг с учетом иерархии ИТ-инфраструктуры. Это дает возможность создать иерархию решений по управлению и поддержанию согласованного уровня услуг с учетом существующих ресурсных ограничений и полномочий уровней по выбору управления, задействуя механизмы и возможности верхних уровней иерархии для выбора управления при невозможности реализации управления на нижних уровнях.

Список литературы

1. Брукс П. Метрики для управления ИТ-услугами: пер. с англ. / П. Брукс. – М.: Альпина Бизнес Букс, 2008. – 283 с.
2. Ролик А.И. Система управления корпоративной информационно-телекоммуникационной инфраструктурой на основе агентского подхода / А.И. Ролик, А.В. Волошин, Д.А. Галушко, П.Ф. Можаровский, А.А. Покотило // Вісник НТУУ «КПІ»: Інформатика, управління та обчислювальна техніка. – К.: «ВЕК+», 2010. – № 52. – С. 39–52.
3. Теленик С.Ф. Управління ресурсами центрів оброблення даних / С.Ф. Теленик, О.І. Ролік, К.О. Крижова // Сучасні проблеми прикладної математики та інформатики: XVI Всеукр. наук. конф., 8–9 жовт. 2009: матеріали. – Львів: Видавничій центр ЛНУ, 2009. – С. 203–205.

4. Hubbert E. TechRadar™ For I&O Professionals: IT Service Management Processes, Q1 2012 / E. Hubbert, J.P. Garbani, G. O'Donnell, S.Mann, J. Rakowski. – Forrester Research, Inc. – 2012. – Feb. 7. – 44 p.
5. IT Service Management: An Introduction// J.V. Bon, G. Kemmerling, D. Pondman, Publisher: Van Haren Publishing. – 2002. – 217 p.
6. Information technology. Service management. Part 1: Specification: ISO/IEC 20000-1:2005. – ISO/IEC, 2005. – 16 p.
7. Information technology. Service management. Part 2: Code of practice: ISO/IEC 20000-1:2005. – ISO/IEC, 2005. – 34 p.
8. Ролик А.И. Концепция управления корпоративной ИТ-инфраструктурой / А.И. Ролик // Вісник НТУУ «КПІ»: Інформатика, управління та обчислювальна техніка. – К.: «ВЕК+», 2012. – № 56. – С. 31–55.
9. Теленик С.Ф. Моделі управління віртуальними машинами при серверній віртуалізації / С.Ф. Теленик, О.І. Ролік, М.М. Букасов, А.Ю. Лабунський // Вісник НТУУ «КПІ»: Інформатика, управління та обчислювальна техніка. – К.: «ВЕК+», 2009. – № 51. – С. 147–152.
10. Теленик С.Ф. Управляемый генетический алгоритм в задачах распределения виртуальных машин в ЦОД / С.Ф. Теленик, А.И. Ролик, П.С. Савченко, М.Е. Боданюк // Вісник ЧДТУ. – 2011. – № 2. – С. 104–113.
11. Теленик С.Ф. Адаптивный генетический алгоритм для решения класса задач распределения ресурсов ЦОД / С.Ф. Теленик, А.И. Ролик, П.С. Савченко // Вісник НТУУ «КПІ»: Інформатика, управління та обчислювальна техніка. – К.: «ВЕК+», 2011. – № 54. – С. 164–174.
12. Ролик А.И. Модель управления перераспределением ресурсов информационно-телекоммуникационной системы при изменении значимости бизнес-процессов// Автоматика. Автоматизация. Электротехнические комплексы и системы. – 2007. – №2 (20).– С. 73–82.
13. Клейнрок Л. Теория массового обслуживания / Л. Клейнрок. – М.: Машиностроение, 1979. – 432 с.
14. Острейковский В.А. Теория надежности / В.А. Острейковский. – М.: Высшая школа. – 2003. – 463 с.
15. Месарович М. Теория иерархических многоуровневых систем: пер. с англ. / М. Месарович, Д. Мако, И. Такахара; пер. с англ. под ред. И. Ф. Шахнова. – М.: «Мир», 1973. – 344 с.