УДК 004.32

О. Я. Матов
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Оптимізація надання послуг обчислювальними ресурсами адаптивної хмарної інфраструктури

Розглянуто інфраструктуру хмарних обчислень (ХО) як об'єкт адаптації і процес адаптації хмарних обчислень як оптимізаційний. Викладено загальну постановку задачі адаптації дисципліни надання обчислювальних ресурсів користувачам ХО. Запропоновано технологію динамічної адаптивної змішаної дисципліни надання обчислювальних ресурсів користувачам ХО. Наведено напрямок вирішення задачі оптимізації динамічної адаптивної змішаної дисципліни. Запропоновано відомий функціонал оптимізації, який базується на припущенні, що результати використання обчислювальних ресурсів користувачем (вирішення задач користувача) знецінюються пропорційно часу їхнього знаходження в черзі на вирішення та самого вирішення в системі ХО. Можливі й інші функціонали з часовими обмеженнями. Це є актуальним для сучасних глобальних інформаційно-аналітичних систем реального часу із застосуванням технологій хмарних обчислень і може бути критичним при обмежених обчислювальних ресурсах ХО. Наведено, що задача оптимізації вирішується ітераційним методом з використанням відповідних аналітичних моделей функціонування ХО.

Ключові слова: *хмарні обчислення, дисципліна надання обчислювальних ресурсів, адаптація та оптимізація дисциплін обслуговування, ефективність адаптації, змішана дисципліна обслуговування, математична модель.*

Вступ

Створення адаптивних інфраструктур хмарних обчислень (ХО), які здатні динамічно адаптуватися до постійно діючих змін умов функціонування, та розробка відповідних методів організації обчислень є важливим напрямком розвитку сучасних глобальних інформаційно-аналітичних систем із застосуванням технологій хмарних обчислень.

Адаптація, як керування, є вторинною стосовно основного контуру керування. Якщо керування виконує основні цілі, реалізація яких забезпечує функціонуван-

ня об'єкта, то адаптація забезпечує якість цього функціонування. Тому, коли виникає потреба в підвищенні (чи підтримці на необхідному рівні) якості функціонування об'єкта, завжди виникає необхідність адаптації.

Особливість системи адаптації полягає в тому, що працювати їй приходиться в умовах значної невизначеності зовнішнього середовища та поводження об'єкта.

Невизначеність середовища і об'єкта є характерною рисою, що дозволяє розглядати адаптацію як специфічний вид керування. При цьому ступінь невизначеності визначає важливість рішення задачі адаптації: чим більше невизначеність, тим більше необхідність адаптації.

Хмарні обчислення як об'єкт адаптації

Хмарні обчислення є об'єктами з високим рівнем невизначеності процесу функціонування. Тут зовнішню невизначеність потоку запитів на обчислювальні ресурси (ОР) (середовища) доповнює внутрішня невизначеність ХО (об'єкта), що зв'язана з наявністю чи відсутністю необхідних ОР, випадкових несправностей системи ХО, а також необхідність забезпечення певних часових характеристик для багатьох клієнтів. Саме це визначає необхідність введення адаптації у процес функціонування ХО.

Крім того, введення адаптації у процес функціонування ХО зв'язано з необхідністю підтримки системи в оптимальному, а іноді і просто працездатному стані, незалежно від численних факторів зовнішнього та внутрішнього характеру, що виводять ХО з необхідного цільового стану.

Усе сказане однаковою мірою може бути віднесено і до обчислювального процесу як об'єкта адаптації, тому що він розвивається у ХО і є його невід'ємним атрибутом.

У поняття адаптації як активної дії (керування) звичайно вкладають два змісти: пристосування об'єкта до фіксованого середовища (пасивна адаптація) і пошук середовища, адекватного даному об'єкту (активна адаптація) [1]. У першому випадку об'єкт, що адаптується, функціонує так, щоб виконати свою мету в даному середовищі щонайкраще, тобто максимізує свою ефективність у даному середовищі. Активна адаптація, навпаки, має на увазі зміну середовища з метою максимізації ефективності функціонування об'єкта.

Стосовно до ХО, як системи масового обслуговування (СМО), активна адаптація може розглядатись як зміна інтенсивності чи кількості вхідних потоків заявок, а також законів розподілу процесів надходження заявок.

У практиці функціонування ХО найбільше часто використовується пасивна адаптація, при якій адаптуючий вплив може мати різний характер. Воно може привести до зміни або параметрів об'єкта адаптації (параметрична адаптація), або його структури (структурна адаптація). Як керовані параметри ХО можуть розглядатись інтенсивність і закони розподілу процесу обслуговування заявок, обмеження на тривалість очікування (перебування) заявок у черзі (системі), порядок обслуговування заявок (дисципліна обслуговування) тощо. Прикладом структурної адаптації, при якій змінюється число обслуговуючих приладів і зв'язку між ними, є переконфігурація багатосерверного ХО. Структурна адаптація більш радикальна і звичайно супроводжується параметричною, тому що кожна структура має свої параметри.

Залежно від того, є модель об'єкта адаптації чи ні, розрізняють два дуже важливих види адаптації: адаптацію з моделлю і без моделі (пошукову адаптацію), які істотно відрізняються один від одного [1].

За наявності адекватної моделі об'єкта для синтезу впливу, що адаптує, досить виміряти стан середовища, і використовуючи модель, визначити вплив, що повинен перевести об'єкт у необхідний стан.

Однак дуже часто об'єкт адаптації настільки складний, що неможливо побудувати його модель, а адекватну модель — тим більше. При цьому, мабуть, не можна скористатися методом адаптації з моделлю, що змушує звертатися до пошукової адаптації. Цей вид адаптації відрізняється наявністю пошуку — спеціально організованого процесу, що дозволяє визначити необхідний адаптуючий вплив, не маючи моделі об'єкта. Для пошукової адаптації характерні експерименти з об'єктом, у процесі яких одержують інформацію про його властивості. На основі цієї інформації визначаються адаптуючий вплив, підвищувальна ефективність об'єкта.

Сам по собі процес пошукової адаптації має послідовний багатоетапний характер — на кожному етапі приймаються заходи для підвищення ефективності об'єкта (на відміну від адаптації з моделлю, за якої можлива адаптація за один етап).

Якщо при адаптації з моделлю стан об'єкта потрібно вимірювати тільки для підстроювання його моделі і не потрібно для самої адаптації, то при пошуковій адаптації стан об'єкта несе основну інформацію для формування адаптуючого впливу.

Труднощі адаптації з моделлю полягають у синтезі моделі об'єкта, а сама адаптація — у рішенні оптимізаційної задачі вибору такого формування адаптуючого впливу, що задовольнило би цілям адаптації. При пошуковій адаптації виникають труднощі іншого роду — потрібно одночасно і експериментувати з об'єктом, і адаптувати його.

В усіх випадках, коли вдається побудувати адекватну модель об'єкта, питання про вибір виду адаптації зважується однозначно на користь адаптації з моделлю, тому що тільки наявність моделі дозволяє швидко адаптувати об'єкт.

Адаптація хмарних обчислень як оптимізація

Рішення задачі адаптації зводиться до визначення такого керуючого (адаптуючого) впливу, при якому досягається максимальна ефективність роботи об'єкта у сформованій ситуації.

Сформована ситуація характеризується двома факторами: станом середовища, у якій знаходиться об'єкт, і станом самого об'єкта адаптації.

Для ХО, як системи масового обслуговування, під станом середовища можна розуміти, наприклад, інтенсивність вхідних потоків заявок, а під станом об'єкта (системи) — число чи час очікування (перебування) заявок у черзі (системі), чи справність-несправність обслуговуючого приладу, рівень завантаження системи і т.п.

Залежно від сформованої ситуації повинен бути сформований адаптуючий вплив, що мінімізує середнє число чи середній час очікування (перебування) заявок у черзі (системі), або час входження системи в стаціонарний режим, або сумарну вартість за роботу системи, або вірогідність утрати заявок і т.д. Метою адапта-

ції можуть бути максимізація доходу від обслуговування заявок, ліквідація перевантаження системи та підтримка її у стаціонарному режимі функціонування.

Таким чином, адаптацію ХО можна розглядати як процес оптимізації роботи у сформованій ситуації.

Загальна постановка задачі адаптації дисципліни обслуговування

Задача адаптації дисципліни обслуговування у ХО виникає в зв'язку з непередбаченими і неконтрольованими змінами в середовищі та системі, що неминуче змінюють оптимальне настроювання дисципліни обслуговування, якщо така була реалізована в системі. Тому систематичне підстроювання (адаптація) дисципліни обслуговування неминуче при бажанні підтримувати систему в оптимальному режимі незалежно від змін, що відбуваються в середовищі та системі.

Сформулюємо в загальному вигляді задачу адаптації дисципліни обслуговування [1].

Нехай X та E — контрольований і неконтрольований стани середовища. Пара $\{X, E\}$ однозначно описує середовище, в якій знаходиться ХО. Наприклад, X — паспортні дані заявок з обслуговування, а E — інтенсивність їхнього надходження.

Аналогічно пара $\{Y, H\}$ описує стан системи. Тут Y та H — відповідно контрольовані та неконтрольовані фактори. Наприклад, Y — довжина черг заявок, а H — інтенсивність їхнього обслуговування чи інтенсивність відмовлень системи.

Показник ефективності системи носить екстремальний характер. Він визначений на контрольованих станах середовища та системи:

$$Э = Э(X, Y). \tag{1}$$

Як показники ефективності системи можуть виступати середній час перебування (очікування) заявок у системі (черги), середня довжина черги заявок, середня сумарна вартість очікування (перебування) заявок у черзі (системі) і т.п.

Стан системи Y залежить від X, E та H, а також від дисципліни обслуговування S:

$$Y = F(X, E, H, S), \tag{2}$$

де F — оператор системи.

Під дисципліною обслуговування розуміють правило вибору заявок на обслуговування залежно від станів середовища та системи:

$$S = S(X, Y). \tag{3}$$

Оптимальність дисципліни S у більшості випадків зв'язана з екстремізацією показника ефективності функціонування системи (1). Це означає, що для синтезу оптимальної дисципліни S^0 необхідно вирішити наступну оптимізаційну задачу:

$$Э[X, F(X, E, H, S)] \to \underset{S \in S^*}{extr} \Rightarrow S^0, \tag{4}$$

де S^* — обмеження, що накладаються на вибір дисципліни обслуговування S.

Ці обмеження можуть бути зв'язані, наприклад, з визначеним набором заздалегідь заданих дисциплін обслуговування і т.п.

Очевидно, що вирішити задачу (4) на стадії проектування обчислювальної системи (ОС) неможливо, тому що апріорі невідомі фактори E та H. Усереднювання за цими факторами вводити не можна, тому що вони можуть мати нестаціонарний характер.

Тому задачу синтезу оптимальної дисципліни S^0 варто вирішувати шляхом адаптації ХО, тобто в режимі їхньої експлуатації. Тоді адаптація зводиться до рішення задачі

$$Э(S) \to \underset{S \in S^*}{extr} \Rightarrow S^0$$

за локальними спостереженнями оцінок значення показника ефективності при різних дисциплінах:

$$\hat{Э}_1 = \hat{Э}(S_1), ..., \hat{Э}_\xi = \hat{Э}(S_\xi).$$

Алгоритм адаптації повинний указати послідовність переходу від однієї дисципліни до іншої: $S_1 \to S_2 \to \cdots \to S_\xi \to \cdots$, яка приводить до рішення S^0, що є оптимальним у сформованій ситуації.

Використання для адаптації технології динамічної адаптивної змішаної дисципліни надання обчислювальних ресурсів користувачам хмарних обчислень

На даний час відома велика кількість різних дисциплін обслуговування. З них у ХО широко застосовуються дисципліни обслуговування з відносним і абсолютним пріоритетами. Однак ці дисципліни є статичними і внаслідок цього мають ряд істотних недоліків, що знижують ефективність обчислювальних систем (процесів) в умовах невизначеності зовнішнього середовища та поводження самих систем.

При використанні дисципліни з відносним пріоритетом вибір чергової заявки для обслуговування може бути здійснений тільки після завершення поточного обслуговування, навіть якщо заявка, що обслуговується, має нижчий пріоритет. Унаслідок цього тривалість перебування у ХО деяких найбільш важливих заявок може виявитися неприпустимо великою. Зменшення затримки в обслуговуванні важливих заявок досягається за рахунок переривання, тобто введення для цих заявок абсолютного пріоритету. Однак при цьому зростає тривалість перебування у ХО заявок низьких пріоритетів і, в ряді випадків, при інтенсивному надходженні важливих заявок може відбутися блокування процесу обслуговування заявок низьких пріоритетів, що також знижує ефективність ХО в цілому.

Для компенсації недоліків, які властиві дисциплінам обслуговування з відносним і абсолютним пріоритетами й обліком їхніх переваг, доцільно реалізувати у ХО змішані дисципліни обслуговування, що використовують як відносний, так і абсолютний пріоритети.

Розглянемо одну зі змішаних дисциплін обслуговування. Нехай N вхідних потоків заявок відповідно до їхньої важливості та терміновості в обслуговуванні розбито на M груп, між якими діє абсолютний пріоритет, а всередині — відносний. Це означає, що заявки будь-якого потоку з групи $m(m=\overline{1,M})$ переривають обслуговування заявок, які належать потокам із груп з номерами $\overline{m+1,M}$. У кожній групі знаходиться N_m потоків, заявки яких не переривають один одного. Очевидно, що $\sum_{m=1}^{M} N_m = N$. Пріоритет будь-якої заявки в системі з такою дисципліною обслуговування може бути описаний парою чисел m та n, де $n=\overline{1,N_m}$ і визначає номер потоку заявок у групі з номером m.

Описана змішана дисципліна обслуговування дозволяє за рахунок її адаптації більш гнучко реагувати на різні ситуації, що виникають у процесі функціонування ОС. При цьому адаптація дисципліни складається зі зміни кількості та положення границь, що розділяють потоки заявок на групи абсолютного пріоритету, тобто зі зміни кількості груп і кількості потоків у групах. Варіанти групування будемо називати розбивками. Загальна кількість розбивок Φ визначається числом потоків заявок $N: \Phi = 2^{N-1}$. Кожна розбивка $\varphi(\varphi \in \Phi)$ задається сукупністю чисел $\{N_1, N_2, \ldots, N_M\}$.

Така дисципліна правомірно названа динамічною адаптивною змішаною дисципліною надання обчислювальних ресурсів користувачам ХО.

Уведена в розгляд змішана дисципліна обслуговування представляє відомий практичний інтерес, оскільки вона при оптимальному виборі розбивки потоків по групах у принципі забезпечує не гірше обслуговування порівняно з «*чистими*» дисциплінами (з відносним і абсолютним пріоритетами). Так, при $M=N$, $N_m=1$ для всіх $m=\overline{1,M}$ виходить дисципліна обслуговування з абсолютним пріоритетом, а при $M=1$, $N_1=N$ — з відносним.

Задачі динамічної адаптивної змішаної дисципліни надання обчислювальних ресурсів з моделлю

Розглянемо дві практичні задачі динамічної адаптивної змішаної дисципліни надання обчислювальних ресурсів (змішаної дисципліни обслуговування) з моделлю.

Одними із основних показників ефективності ХО є показники, що базуються на оцінці часових характеристик цих систем. Такі показники можуть задаватися договором між постачальником і користувачем ОР ХО і здобувають особливе значення для систем, що функціонують у реальному масштабі часу.

Унаслідок випадкового характеру обчислювального процесу виникають додаткові затримки в обробці інформації, порушуються припустимі обмеження на час її перебування в ХО, що негативно позначається на ефективності рішення цільових задач користувачів.

Для забезпечення необхідної ефективності ХО в таких ситуаціях необхідно підтримувати часові характеристики системи на заданому рівні. В умовах дефіци-

ту обчислювальних ресурсів це можливо тільки за рахунок підвищення ефективності обчислювального процесу, зокрема, за рахунок адаптації дисципліни обслуговування. Поряд з цим виникає задача найбільш ефективного використання наявних обчислювальних ресурсів у кожен момент часу функціонування керуючої ХО. Цю задачу також можна вирішити шляхом адаптації дисципліни обслуговування.

У зв'язку з викладеним як показник ефективності ХО виберемо середню сумарну вартість часу надання (очікування в чергах і часу використання, тобто перебування в ХО як у СМО) ОР за заявками (вимогами) користувачів. Для цього використаємо відомий функціонал [1] — середню сумарну вартість часу надання ОР:

$$C^{(S)} = \sum_{i=1}^{n} \alpha_i \lambda_i v_i^{(s)},$$

з чого маємо

$$C^{(\varphi)} = \sum_{m=1}^{M} \sum_{n=1}^{N} \alpha(m,n)\lambda(m,n)v^{(\varphi)}(m,n), \qquad (5)$$

де α_i — вартість за одиницю часу ОР для *i*-го типу заявок користувачів;

λ_i — інтенсивність *i*-го потоку заявок;

$v_i^{(S)}$ — середній час надання ОР заявок *i*-го потоку;

n — кількість типів заявок;

s — параметр, що характеризує спосіб організації обчислювального процесу;

$v^{(\varphi)}(m,n)(m=\overline{1,M}, n=\overline{1,N_m})$ — середній час надання ОР в ХО заявкам (*m*, *n*)-го потоку;

$\alpha(m, n)$ — вартість одиниці часу надання ОР в ХО заявкам (*m*, *n*)-го потоку;

$\lambda(m, n)$ — інтенсивність (*m*, *n*)-потоку надання ОР в ХО.

Показник ефективності базується на припущенні, що результати використання ОР користувачем знецінюються пропорційно часу їхнього перебування в системі ХО. Тоді цілями адаптації змішаної дисципліни обслуговування будуть або задоволення вимог вчасного перебування (*m*, *n*) заявок у системі, що задаються припустимими значеннями цього часу $v_{Д}(m,n)$, або мінімізація функціоналу (5). Ця мета досягається шляхом відшукання відповідних оптимальних розбивок φ^0, тобто задачі адаптації змішаної дисципліни обслуговування з відносно-абсолютним пріоритетом являють собою оптимізаційні задачі, загальна постановка яких розглянута вище.

Оскільки сформульовані вище цілі адаптації змішаної дисципліни обслуговування можуть бути досягнуті при декількох різних розбивках потоків заявок на групи абсолютного пріоритету, то виникає необхідність уведення додаткового обмеження на вибір розбивки φ.

Наявність у ХО абсолютного пріоритету потребує деяких технологічних утрат ОР, які пропорційні числу груп (рівнів) абсолютного пріоритету. У зв'язку з

цим оптимальною необхідно вважати таку розбивку, яка забезпечує досягнення цілей адаптації при мінімальній кількості груп абсолютного пріоритету M.

Тоді розглянуті задачі адаптації змішаної дисципліни обслуговування можуть бути формально поставлені в такий спосіб:

$$\begin{gathered} v^{(\varphi)}(m,n) \le v_{Д}(m,n) \Rightarrow \varphi^{0}, \\ \varphi \in \Phi, \\ M = \min. \end{gathered} \tag{6}$$

$$\begin{gathered} C^{(\varphi)} \to \min \Rightarrow \varphi^{0}, \\ \varphi \in \Phi, \\ M = \min. \end{gathered} \tag{7}$$

Рішення задач відшукування оптимальної розбивки (6) і (7) за допомогою відомих аналітичних методів оптимізації не представляється можливим. Єдиний шлях рішення цих задач — евристичний підхід, що не має формального обґрунтування, а спирається лише на специфіку задач (математичних моделей) і зв'язані з ними розуміння.

З виразів (5)–(7) випливає, що досягнення цілей адаптації змішаної дисципліни обслуговування сполучено з потребою оцінки значення середнього часу перебування в системі заявок (m, n)-типу — $v(m, n)$. Тому виникає необхідність синтезу математичної моделі ХО зі змішаною дисципліною обслуговування.

Висновки

Відшукування оптимальної дисципліни обслуговування не завжди зв'язано з екстремізацією показника ефективності функціонування ХО. Метою адаптації також може бути задоволення обмежень на показник ефективності, що задані у вигляді рівностей чи нерівностей. У будь-якому випадку така постановка задачі адаптації припускає необхідність реалізації у ХО декількох чи однієї змішаної дисципліни обслуговування.

1. Матов А.Я., Шпилев В.Н., Комов А.Д. и др. Организация вычислительных процессов в АСУ/под ред. А.Я.Матова. Киев, 1989. 200 с.

2. Матов О.Я., Храмова І.О. Проблеми користування і математичне моделювання хмарних обчислень для інтегрованої інформаційно-аналітичної системи державного управління. *Реєстрація, зберігання і оброб. даних*. 2010. Т. 12. № 2. С. 113–127.

3. Матов О.Я., Храмова І.О. Сучасні технології інтеграції інформаційних ресурсів. *Реєстрація, зберігання і оброб. даних*. 2009. Т. 11. № 1. С. 33–42.

4. Matov A.Y. Optimum operational training in systems containing components subject to failure. *Engineering Cybernetics*. 1975. **13**(5). С. 87–90.

Надійшла до редакції 06.09.2018