Ertragsmanagementmodelle in serviceorientierten IT-Landschaften

Thomas Setzer, Martin Bichler

Lehrstuhl für Internetbasierte Geschäftssysteme (IBIS)
Fakultät für Informatik, TU München
Boltzmannstr. 3
85748 Garching bei München
thomas.setzer@in.tum.de
martin.bichler@in.tum.de

Abstract: Serviceorientierte Architekturen ermöglichen die dynamische Bereitstellung und Nutzung von IT-Dienstleistungen, was oftmals auch als On-Demand- oder Utility-Computing bezeichnet wird. Analysten sehen hierin den zukünftig am stärksten wachsenden Markt für IT-Dienstleister. Entsprechende IT-Infrastrukturen sollen hohe Skalierbarkeiten und hohe durchschnittliche Ressourcenauslastungen ermöglichen. Neuere Virtualisierungskonzepte, wie beispielsweise bei Grid-basierten Infrastrukturen eingesetzt, erlauben IT-Dienstleistern die effiziente Bereitstellung von Diensten, da diese auf gemeinsame Ressourcenpools zugreifen (engl. shared resources). Engpässe durch hohes Lastaufkommen verursachen jedoch auch hier die Abweisung von Dienstanfragen. Da Dienstannahmen mit temporären Ressourcenbelegungen einhergehen, sind diese somit neben Erträgen auch mit Opportunitätskosten verbunden. Diese Arbeit stellt Ertragsmanagementmodelle für IT-Dienstleister vor, die solche virtualisierten IT-Infrastrukturen verwenden. Die Modelle entscheiden anhand eines Vergleichs von Dienstpreisen und Opportunitätskosten über Akzeptanz, Pufferung oder Ablehnung von Anfragen, wobei die vorausschauende Ablehnung einer Reservierung von Ressourcen für erwartete, ertragreichere Anfragen entspricht. Ziel ist die Ertragsmaximierung abhängig von prognostizierter Nachfrage nach Diensten, deren Preisen, Service Level Agreements (SLAs), Verbrauch an Ressourcen sowie aktuellen Ressourcenauslastungen. Kombinatorische Effekte aufgrund der Belegung mehrere Ressourcen durch Dienste finden in den Modellformulierungen Berücksichtigung. Diese Modelle wurden mittels Monte-Carlo-Simulationen evaluiert und können in Entscheidungslogiken von Load-Balancing- oder Zugriffskontrollkomponenten integriert werden.

1 Einleitung

Serviceorientierte Architekturen ermöglichen flexible und dynamische Bereitstellung und Nutzung von IT-Dienstleistungen, was oftmals als On-Demand- oder Utility-Computing bezeichnet wird.

Analysten sehen hierin den zukünftig am stärksten wachsenden Markt für IT-

Dienstleister.[1] Effiziente Bereitstellung von Diensten setzt jedoch auch in flexiblen, serviceorientierten IT-Landschaften hohe Skalierbarkeiten und hohe durchschnittliche Ressourcenauslastungen voraus. IT-Dienstleister stellen heutzutage in ihren Rechenzentren meist dedizierte Hard- und Softwaresysteme sowie dedizierte Backupsysteme für bestimmte Kunden oder Dienste bereit. Um auch in Phasen hoher Last Dienste mit vereinbarten Dienstgüten erbringen zu können, werden Systeme nach Lastspitzen ausgerichtet, was zu geringen Durchschnittsauslastungen führen kann. Derzeit spricht man von einem durchschnittlichen Nutzungsgrad von IT-Ressourcen von 10-25%.[23]

Virtualisierungskonzepte ermöglichen die Dimensionierung von Ressourcen nach Durchschnittslasten, da zusätzlich benötigte Ressourcen Diensten dynamisch aus Ressourcenpools zugewiesen werden. Im Gegensatz zu klassischem Application-Service-Provisioning (ASP) erlauben Virtualisierungskonzepte die Erbringung derartiger Dienstleistungen auf Basis einer gemeinsamen IT-Infrastruktur, die von Kunden und Diensten gemeinschaftlich genutzt wird (engl. *shared services*).[4] Typische Beispiele, die heute schon als On-Demand-Service angeboten werden, sind CRM-Anwendungen (z.B. SalesForce.com), die für viele Kunden in verschiedenen Dienstklassen angeboten werden, bis hin zu ganzen ERP-Systemen (z.B. SAP On-Demand).

Der Einsatz entsprechender Virtualisierungstechniken erfordert allerdings neue Konzepte für die Ressourcenallokation bei IT-Dienstleistern. Bei hoher Arbeitslast können Knappheiten in Ressourcenpools auftreten, wodurch die Beantwortung sämtlicher Dienstanfragen nicht mehr möglich ist. Die Annahme von Dienstanfragen und damit einhergehende, temporäre Ressourcenbelegungen sind somit neben Erträgen auch mit Opportunitätskosten verbunden. Im Gegensatz zum Fall dedizierter Zuordnung von Ressourcen zu bestimmten Diensten konkurrieren heterogene Dienste mit unterschiedlichen Preisen, SLAs und Ressourcenverbrauch um knappe Ressourcen. Auch bei (noch) unvollständiger Auslastung sinkt mit jeder Bedienung einer Anfrage für die Zeitdauer der Diensterbringung die verfügbare Restkapazität gemeinsam genutzter Ressourcen.

In diesem Artikel stellen wir Ertragsmanagementmodelle für IT-Dienstleister mit virtualisierter Infrastruktur vor, die diese Problematik lösen sollen. Die Modellformulierungen entsprechen Entscheidungen über Akzeptanz oder Ablehnung (bzw. Pufferung) von Dienstanfragen anhand eines Vergleichs von Dienstpreisen und Opportunitätskosten, wobei die vorausschauenden Ablehnung von Anfragen der Reservierung von Ressourcen für erwartete, lukrativere Anfragen entspricht. Ziel ist die Ertragsmaximierung abhängig von der prognostizierten Nachfrage nach Diensten, deren Preisen, SLAs, Verbrauch an Ressourcen sowie aktuellen Rerssourcenauslastungen. Neben stochastischen Einflüssen werden ebenfalls die kombinatorischen Einflüsse durch die Belegung verschiedener Ressourcen durch einzelne Dienste berücksichtigt [Wi92]. Die Modelle können in Entscheidungslogiken von Load-Balancing- oder Zugriffskontrollkomponenten integriert werden. Kapitel 2 beschreibt die Problemformulierung und verschiedene Modellierungs-

---

[1] Gartner Report: Market Trends 2004
[2] http://www.jayeckles.com/research/grid.html
[3] http://www.speicherguide.de/magazin/virtualisierung.asp?theID=358
[4] Ovum Report: On-demand outsourcing services: the competitive landscape in EMEA, .December 2004.

ansätze. In Kapitel 3 wird die Effizienz der Methode mittels Monte-Carlo-Simulationen analysiert. Kapitel 4 fasst die Arbeit zusammen und gibt einen Ausblick auf weitere Forschungsfragen und Einsatzszenarien.

2 Problemformulierung

Nachfolgend beschreiben wir ein grundlegendes Modell zum Ertragsmanagement bei virtualisierten IT-Ressourcen. Ausgehend von diesem Basismodell haben wir einige Modellerweiterungen und –verfeinerungen entwickelt, die zahlreiche Realweltbedingungen mit abbilden.

Das Dienst-Portfolio eines IT-Dienstleister bestehe aus I Diensten i (i = 1, …, I), die zu diskreten Zeitpunkten t_r (r = 0, …, ∞) im Durchschnitt D_i mal nachgefragt werden. D_i sei konstant und unabhängig von vergangenen Dienstnachfragezeitpunkten und -mengen (Annahme einer unendlichen Population). Die Belegungsdauer der im Zuge der Diensterbringung verwendeten Ressourcen und die der Diensterbringung selbst seien von konstanter Länge Δt ($\Delta t = t_{r+1} - t_r$) und somit zum jeweils nächstmöglichen Nachfragezeitpunkt t_{r+1} abgeschlossen. Die Ressourcenverbrauchskoeffizienten a_{ei} geben die Verbrauchsmenge bzw. Verwendungsmenge an Einheiten der Ressource e (e = 1, …, E) für die Zeitdauer Δt an. Eine Ressource e hat eine begrenzte Kapazität C_e. Abbildung 1 veranschaulicht die genannten Zusammenhänge.

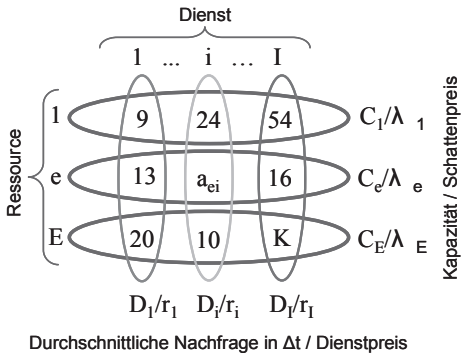

Abbildung 1 Ressourcenverbrauchsmatrix

Die zur Verfügung stehenden Ressourcen sollen nun in der profitabelsten Weise für die Beantwortung von Dienstanfragen eingesetzt werden. Unter den gegebenen Annahmen kann das Problem durch ein Ganzzahliges Lineares Programm (engl.: Integer Program (IP)) formuliert werden. Die ganzzahlige Variable x_i beschreibt die Anzahl der zu akzeptierenden Dienstanfragen für eine Zeitperiode Δt und r_i den Ertrag pro Dienst. Die stochastische Variable D_i wird in dieser Formulierung als deterministische Größe behandelt.

$$
\begin{aligned}
\max \quad & \sum_{i \le I} r_i \cdot x_i & & \\
s.t. \quad & \sum_{i \le I} a_{ei} x_i \le C_e & \forall e \le E & \quad \text{(IP)} \\
& x_i \le D_i & \forall i \le I & \\
& x_i \in Z_+ & \forall i \le I &
\end{aligned}
$$

Sind in einem Linearen Programm (LP) sämtliche Restriktionen der Entscheidungsvariablen sowie die rechte Seite des Gleichungssystems ganzzahlig, was hier zutrifft, sind dessen Lösungen stets ganzzahlig und das IP kann durch seine LP-Relaxation, einem Deterministischen Linearen Programm (DLP), gelöst werden [NM02]. Die dualen Variablen λ_e der Kapazitätsrestriktionen der LP-Relaxation können ökonomisch als Schattenpreise oder Opportunitätskosten der Verwendung einer Ressourceneinheit interpretiert werden. So genannte „Bid-Price-Controls“ stellen bei der Entscheidung über Annahme oder Ablehnung einer Dienstanfrage den Ertrag einer Dienstannahme den gesamten Opportunitätskosten (engl.: bid prices) gegenüber, die mit der Dienstannahme einhergehen. Opportunitätskosten eines Dienstes i können durch Addition der Produkte aus Ressourcenverbrauchskoeffizient und Opportunitätskosten pro Ressourceneinheit ($\sum_e a_{ei} \lambda_e$) berechnet werden [TR99]. Es werden ausschließlich Dienstanfragen akzeptiert, bei denen der Ertrag die Opportunitätskosten übersteigt.

Auf Grundlage dieses Basismodells wurden eine Reihe von Erweiterungen entwickelt, die zahlreiche Realweltbedingungen mit abbilden, jedoch aus Platzgründen hier nicht vollständig diskutiert werden können. Das Basismodell DLP unterstellt Dienstanfragen zu diskreten Zeitpunkten, deren Bearbeitungen zum nächsten diskreten Nachfragezeitpunkt abgeschlossen sind. In der Modellformulierung DLPc wird von der Annahme diskreter Nachfragezeitpunkte und konstanten Ressourcenbelegungszeitdauern durch Dienste abgewichen; Anfragen erfolgen kontinuierlich. Auch wurden etwaige Vertragsstrafen berücksichtig, die durch Verletzung von Service Level Agreements bei Anfragen-Ablehnungen resultieren. Schließlich wurden stochastische Formulierungen analysiert, mit der Modellierung der Nachfrage als stochastische Größe [Bo99].

3 Evaluierung

Die Modelle des letzten Abschnitts berechnen Opportunitätskosten für Dienste und können im Rahmen von Zugriffskontroll- und Load-Balancing-Verfahren zum Einsatz kommen. Um Aussagen über die Effizienz der Ertragsmanagementmodelle zu treffen, wurde die DLP- sowie die DLPc-Formulierung in Monte-Carlo-Simulationen evaluiert. Effizienz-Kriterium war die Höhe des bei verschiedenen Kapazitäten und volatiler Nachfrage in einem Zeitraum erzielten Gesamtertrags. Hierfür wurden, entsprechend stochastischen Nachfrageverteilungen, für bestimmte Zeitintervalle Dienstanfragen auf verschiedene, hinsichtlich Preisen, SLAs und Ressourcenbelegungsmengen heterogen zusammengesetzte Dienstleistungsportfolios generiert.

Startend mit für den Gesamtbedarf ausreichender Kapazität wurde sukzessiv die Kapazität einer Ressource pro Zeitintervall reduziert. Es wurde jeweils der Ertrag unter Zugriffskontrolle mit demjenigen verglichen, der sich ohne Kontrolle ergeben hätte (Dienste werden nur abgelehnt, wenn zum Anfragezeitpunkt zu wenig Kapazität frei ist). Die Ergebnisse zeigen, dass mit zunehmender Last die ökonomischen Vorteile durch den Einsatz geeigneter Zugriffskontrollverfahren steigen. So konnten bei realitätsnahen Szenarien Ertragssteigerungen von 60% und mehr erzielt werden. Durchgängig zeigten sich positive Korrelationen zwischen Ertragssteigerung und Dienstpreisdifferenzen, Nachfragevolatilität und der Größe des Dienstleistungsportfolios. Die Überlegenheit des DLPc über DLP stieg auf Grund seiner Adaptivität an kurzfristige Schwankungen mit zunehmender Nachfragevolatilität.

4 Zusammenfassung und Ausblick

In dieser Arbeit wurden Ertragsmanagement-Methoden bei IT-Dienstleistern mit virtualisierten IT-Ressourcen, bzw. einem Ressourcen-Grid, vorgestellt. Die Ergebnisse der Simulation zeigen, dass mit zunehmender Last auf gemeinsam genutzte Ressourcen der Ertrag durch den Einsatz der vorgestellten Optimierungsmodelle steigt.

Wir konzentrieren uns in unserer derzeitigen Forschung auf mehrschichtige Server-Cluster, die dynamische Ressourcenzuteilungen zu Dienstanfragen durchführen. Diese kommen typischerweise bei Webapplikationen zum Einsatz. Die Belegungsdauer und –menge an Ressourcen mehrschichtiger Dienste wird gemessen, um daraus Ressourcenverbrauchskoeffizienten für die beschriebenen Optimierungsmodelle abzuleiten. Die vorgestellten Modelle werden derzeit prototypisch als Teil von Zugangskontrollkomponenten für On-Demand-Services implementiert.

Die Fortführung des Projektes erfolgt mit freundlicher Unterstützung durch Siemens Business Services und in Kooperation mit der BMW Group

Literaturverzeichnis

[Bo99] S. de Boer, R. Freling, and N. Piersma, "Stochastic Programming for Multiple-Leg Network Revenue Management," Erasmus University Rotterdam, Rotterdam Technical Report EI-9935/A, 1999.

[NM02] Neumann, K.; Morlock, M.: Operations Research. Hanser Verlag, München, 2002.

[TR99] Talluri, K. T.; van Ryzin, G. J.: An Analysis of Bid-Price Controls for Network Revenue Management. In: Management Science 44, 1999.

[Wi92] E. L. Williamson, "Airline Network Seat Control." MIT Cambridge, MA, USA, 1992.