Quality of Service im Internet

DISSERTATION
der Universität St. Gallen,
Hochschule für Wirtschafts-,
Rechts- und Sozialwissenschaften (HSG)
zur Erlangung der Würde eines
Doktors der Wirtschaftswissenschaften

vorgelegt von

Thorsten Hau

aus

Deutschland

Genehmigt auf Antrag der Herren

Prof. Dr. Walter Brenner

und

Prof. Dr. Rüdiger Zarnekow

Dissertation Nr. 3736

D-Druck-Spescha, St. Gallen, 2009

Die Universität St. Gallen, Hochschule für Wirtschafts-, Rechts- und Sozialwissenschaften (HSG), gestattet hiermit die Drucklegung der vorliegenden Dissertation, ohne damit zu den darin ausgesprochenen Anschauungen Stellung zu nehmen.

St. Gallen, den 16. November 2009

Der Rektor:

Prof. Dr. Ernst Mohr, PhD

Für Eva.

Vorwort

Diese Arbeit entstand zwischen 2007 und 2009 am Lehrstuhl von Prof. Dr. Walter Brenner am Institut für Wirtschaftsinformatik der Universität St. Gallen (IWI4). Während dieser Zeit habe ich in unterschiedlichsten Bereichen gearbeitet. Mein erstes Praxisprojekt am IWI4 hatte den Namen „Potenziale von SOA in der Chemischen Industrie“. Aus den Fallstudien, die ich für dieses Projekt erhob, entstand meine erste Veröffentlichung. Anschliessend arbeitete ich im Competence Center Industrialisierung des Informationsmanagements (CC IIM) an der Beschreibung von IT-Dienstleistungen. Meine dritte Station am IWI4 war das Projekt „Quality of Service im Internet“. Diesen neuen Forschungsbereich übernahm ich Anfang 2008 und arbeitete von dort an eng mit der TU-Berlin zusammen an mehreren Praxisprojekten und Veröffentlichungen.

Während meiner Promotion habe ich mit vielen Menschen zusammengearbeitet, wurde oftmals unterstützt und habe viel gelernt. Zuerst gebührt mein Dank meinem Referenten Prof. Dr. Walter Brenner für die Unterstützung bei meinen ersten akademischen Gehversuchen. Ohne seine Hilfe wäre die vorliegende Arbeit nicht möglich gewesen. Ebenso danke ich Prof. Dr. Rüdiger Zarnekow für die Übernahme des Korreferats und für die Zusammenarbeit und Unterstützung im Rahmen des Projekts „Quality of Service im Internet“.

Zu Dank verpflichtet bin ich auch Kollegen und Unternehmenspartnern, mit denen ich zusammen an Publikationen gearbeitet habe. Dr. Axel Hochstein und Dr. Matthias Büger danke ich für Ihr Engagement beim Thema „Outsourcing und Spieltheorie“. Jochen Wulf danke ich für viele Reviews, konstruktive Anmerkungen und die Koautorenschaft diverser Artikel, Bernhard Schindlholzer und Henrik Brocke für die Arbeit an unserer gemeinsamen Veröffentlichung.

Ein herzliches Dankeschön gebührt Barbara Rohner für die Unterstützung beim „Prof. Management“. Bei Carlos Bravo-Sanchez, Alexander Ritschel, Veit Schulz, Alexander Vogedes, Nico Ebert, Quyen Nguyen, Friedrich Köster, Sebastian Dudek und Rouven Kadura bedanke ich mich für die wissenschaftlichen Kaffepausen in freundschaftlicher Atmosphäre.

Meinen Eltern, meinen beiden Brüdern sowie meinen Schwiegereltern sage ich Danke für die Unterstützung und das Verständnis, die mir vor und während der Promotionszeit zuteil wurden.

Zuletzt möchte ich meiner Frau Eva danken. Sie begleitete und unterstützte mich in sämtlichen Phasen der Promotion. Ihr widme ich diese Arbeit.

St. Gallen, im November 2009 Thorsten Hau

Inhaltsverzeichnis

Zusammenfassung

Quality of Service (QoS) bedeutet, dass ein Kommunikationsnetzwerk unterschiedliche Qualitätsniveaus der Datenübertragung garantieren kann. Das Internet in seiner heutigen Form hat diese Fähigkeit nicht. Alle Daten, die über das Internet übertragen werden, werden von den Netzknoten gleich behandelt. Diese Situation stellt kein Problem dar, solange das Netzwerk nur schwach ausgelastet ist. Sobald aber mehr Datenverkehr anfällt, als das Netzwerk verarbeiten kann, müssen die Netzknoten entscheiden, welche Daten sie weiterleiten und welche Daten warten müssen. Das Problem, dass überlastete Netzknoten wichtige Daten nicht so schnell wie möglich weiterleiten, tritt vor allem an den Verbindungspunkten verschiedener Netzwerke (Peerings) auf. Kommerzielle Anbieter haben daher Content Delivery Networks (CDN) und Multi-Homing (MH) als Technologien zur Umgehung der Peerings etabliert.

In zwei Beiträgen untersucht diese Arbeit, wie sich CDN und MH auf das wirtschaftliche Kalkül der Internetanbieter auswirken. Es wird gezeigt, dass beide Technologien den Internetanbietern grössere Spielräume zur Preisgestaltung eröffnen, als dies normalerweise der Fall ist. Der zentrale Grund hierfür ist die Identifizierbarkeit der Quelle des Datenverkehrs und damit die Möglichkeit, eine Geschäftsbeziehung einzugehen. Ein interessantes Ergebnis ist, dass es durchaus möglich ist, dass der Preis für den Internetzugang für Endkunden auf null sinkt. Internetanbieter könnten sich nur über Einnahmen aus dem Geschäft mit Inhalteanbietern finanzieren und diese Einnahmen zur Subventionierung der Endkunden benutzen. Der Grund für solch ein Verhalten liegt im Wettbewerb um Eyeballs, also Endkunden, für deren Erreichbarkeit Werbetreibende eine hohe Zahlungsbereitschaft haben.

Drei Beiträge, die vor den gerade genannten Arbeiten über QoS entstanden sind, beschäftigen sich mit den IT-Trends Serviceorientierung, Industrialisierung und Outsourcing. Diese Arbeiten sind in erster Linie der Arbeit in den jeweiligen Projekten mit Praxispartnern geschuldet, lassen sich aber auch als Kontext zu den QoS-Überlegungen interpretieren. Eine Analyse dieser Entwicklungen zeigt nämlich, dass sie auf zuverlässige QoS angewiesen sind und somit Treiber für die Weiterentwicklung des Internet darstellen.

Diese Argumentation wird in einem Beitrag in Form einer Synthese der beiden Themengebiete QoS und Serviceorientierung vertieft. Der Artikel argumentiert, dass das Internet als Teil der Wertschöpfungskette einer Dienstleistung ein unkalkulierbares Risiko darstellt und daher als Handicap für die weitere Verbreitung geschäftskritischer Services gesehen werden muss. Somit ist es notwendig, bestimmte Qualitätsniveaus der Datenübertragung garantieren zu können, um die Vision des weltumspannenden Service-Netzwerks Realität werden zu lassen.

Abstract

Quality of Service (QoS) means that a communication network can guarantee reliable levels of data transmission quality. The Internet in its present form does not have this capability. All data that is sent over the Internet is being treated equally by all nodes of the network. This situation does not pose a problem as long as the network is only lightly loaded. As soon as there is more demand than there is capacity for data transmission, the nodes must prioritize certain data and delay other. The problem that overloaded network nodes do not forward important data as fast as possible primarily arises at the points of interconnection (peerings) between different networks. Commercial providers have therefore invented Content Delivery Networks (CDN) and Multi-Homing (MH). Both are technologies to circumvent peerings.

In two contributions, this work analyzes how CDN and MH influence the economic incentives of Internet service providers (ISPs). It is shown that both technologies open up greater possibilities for pricing than commonly available. The central reason for this is the possibility of a business relationship between the ISP and the originator of data traffic which becomes possible because the originator can be identified by the ISP. A further interesting result is the possibility of the end user price for Internet access falling to zero. ISPs could possibly be put in a situation in which it is feasible that all their income is generated on the content provider side and that revenues from this business are used to subsidize end user access. Such a situation is quite feasible when taking into account the competition for eyeballs, i.e. end users which advertisers want to reach.

Three contributions that were written prior to the just mentioned works on QoS treat three important IT-trends: Service orientation, industrialization and outsourcing. These articles should be seen primarily as results of the respective projects with industry partners. However, they can also be interpreted as contextual to the QoS works. An analysis of the three mentioned trends shows that they are all highly dependent on reliable QoS and thus can be seen as business drivers for the advancement of the Internet infrastructure towards QoS capabilities.

This argument is further developed in an article that attempts the synthesis of the two disjunct issues of ‘QoS’ and ‘service orientation’. The article argues that the Internet as a part of the service value chain presents an uncontrollable risk factor and is thus a handicap for the further distribution of business critical applications via the Internet. It is necessary to be able to guarantee certain quality levels of data transmission in order to reach the goal of a worldwide network of distributed services.

1 Einleitung

Das Internet hat in der kurzen Zeit seiner Existenz unsere Lebensgewohnheiten drastisch verändert. In den jüngeren Generationen ist es dabei, das Fernsehen abzulösen. Auch im geschäftlichen Bereich ist das Internet nicht mehr wegzudenken. Kaum ein Unternehmen würde heutzutage ohne E-Mail funktionieren.

Obwohl über das Internet heute jede erdenkliche Anwendung angeboten wird und oft auch gut funktioniert, ist das Internet eigentlich ein "Legacy System"; eine Informationstechnische Altlast, die früher oder später durch ein dem aktuellen Stand der Technologie entsprechendes System ersetzt werden muss. Vorschläge hierzu gibt es zuhauf, und "clean slate design", also die Erfindung eines neuen Internet als ob es keine historischen Einschränkungen gäbe, ist eine eigene Forschungsdisziplin [Feldmann 2007].

Eine wesentliche Schwäche des Internet in seiner heutigen Form ist die Unfähigkeit verschiedene Datenströme unterschiedlich zu behandeln. Obwohl dieses Problem technische weitgehend gelöst ist [Wang 2001], gibt es bis heute kein Beispiel für ein dem Internet ähnliches, offenes Netzwerk, in dem unterschiedliche Dienstklassen unterschiedlich behandelt werden. Hierunter leidet vor allem die Verbreitung von Geschäftsanwendungen, die ein konstant hohes Qualitätsniveau bieten müssen.

Next Generation Networks (NGN) sind ein Schlagwort, unter dem die Bestrebungen vieler Infrastrukturunternehmen zur Beseitigung der Schwachpunkte des Internet zusammengefasst werden [Elixmann et al. 2008b]. Leider ist der Begriff kaum definierbar, drei Kernpunkte lassen sich aber beschreiben:

- Die Versorgung der Haushalte mit breitbandigen Anschlüssen im Bereich 100MBit.
- Die Differenzierung von unterschiedlichen Dienstklassen und –güten.
- Die Abwicklung aller Dienste auf einer einheitlichen logischen Infrastruktur. Der Dienst soll nicht merken, ober er auf einem Arbeitsplatz-PC oder einem Mobiltelefon ausgeführt wird.

Kritisch anzumerken ist, dass sich diese Konzepte auf technische Aspekte beschränken. Ökonomische und Betriebswirtschaftliche Zusammenhänge werden kaum thematisiert, obwohl wirtschaftliche Gründe bisher eine Verbreitung von qualitätsgesichertem Datentransport im Internet verhindert haben. Vor allem wurde keine Lösung für bereits heute bestehende Anreizprobleme bei der Zusammenschaltung von Netzwerken gefunden [Amante et al. 2006,Marcus/Elixmann 2008].

1.1 Forschungsfrage

Diese Arbeit beschäftigt sich mit der Frage, warum es im Internet keine Quality of Service (QoS) – also die zuverlässige Garantie von Qualitätsparametern – gibt. Diese Fragestellung ist bedeutsam, weil die Nichtverfügbarkeit von garantierter Übertra-

gungsqualität Wohlfahrtsverluste bedeutet und innovative Dienste im Internet verhindert. Ein Beispiel hierfür ist Software as a Service (SaaS). Hierbei sollen Anwendungen von einem globalen Anbieter, der Software und Daten vorhält, betrieben werden und über das Internet zugänglich gemacht werden. Die so mögliche Spezialisierung und Arbeitsteilung verspricht enorme Produktivitätspotentiale zu erschliessen. Auf der heutigen Internet-Infrastruktur ist diese Vision aber nur schwer realisierbar. ERP-Anwendungen z.B. werden von fast allen Unternehmen zur Steuerung interner und externer Abläufe genutzt. Wird die Bedienung eines solchen ERP Systems aber zu langsam, weil die Datenübertragung im Netzwerk zu lange dauert, kann dies Produktivitätseinbussen bedeuten, da Mitarbeiter auf das System warten, statt Eingaben zu tätigen. Die schwankenden „Geschwindigkeit des Internet“ stellt also eine Hürde für die Akzeptanz solch innovativer Technologien dar.

Sowohl in der Praxis als auch in der Forschung wird der enge Zusammenhang von technischen und wirtschaftlichen Fragestellungen oft ignoriert und nur eine der beiden Seiten wird intensiv thematisiert. Darüber hinaus gibt es einen starken Überhang an technischen Lösungsvorschlägen. Während Quality of Service Mechanismen auf technischer Ebene gut verstanden sind, gibt es weit weniger Arbeiten, die sich mit ökonomischen und betriebswirtschaftlichen Fragestellungen auseinandersetzen. Das Thema dieser Arbeit ist daher die Untersuchung der betriebswirtschaftlichen und ökonomischen Zusammenhänge unter Berücksichtigung technischer Gegebenheiten.

Die Beiträge, die Teil dieser kumulativen Dissertation sind, gliedern sich in zwei Teile. Die früheren Artikel bilden den Kontext für das eigentliche Forschungsthema. Sie beschäftigen sich mit unterschiedlichen IT-Themen und haben nur peripheren Bezug zum Thema QoS. Diese Artikel sind einerseits wichtiger Hintergrund, andererseits können sie auch als Übungen verstanden werden, in denen sich der Autor der Publikationstätigkeit nähert. Die andere Hälfte der Artikel thematisiert direkt QoS-Themen, wobei sich der Artikel [Hau et al. 2009] um eine Synthese der Themen QoS und Serviceorientierung bemüht.

1.2 Forschungsprozess

Diese Arbeit entstand während der Mitarbeit des Autors in drei verschiedenen Kompetenzzentren. Kompetenzzentren am Institut für Wirtschaftsinformatik sind Organisationseinheiten eines Lehrstuhls, die sich in Zusammenarbeit mit Partnern aus der Privatwirtschaft über mehrere Jahre hinweg mit einem Forschungsthema auseinandersetzen. Durch diese organisatorische Aufstellung ist von vorneherein eine gewisse Spannung zwischen Wissenschaft und Praxis gegeben. Während die Partner aus den Unternehmen umsetzbare Vorschläge erwarten, stellt eine wissenschaftliche Publikation gänzlich andere Anforderungen an Allgemeinheit, Nachvollziehbarkeit und Abstützung auf existierenden Arbeiten. Der Autor sieht die Praxisprojekte im Rahmen der Kompetenzzentren daher als wertvolle Ideengeber für eine Forschungsagenda. Etwaige Publikationen müssen sich aber vom konkreten Praxiskontext lösen, da das simple

Niederschreiben von Projekterfahrungen nur ausnahmsweise erfolgversprechend ist. Vielmehr müssen Erfahrungen aus der Praxis abstrahiert und formalisiert werden um relevante Beiträge zum wissenschaftlichen Diskurs zu leisten.

Bedingt durch die wechselnden Projekterfahrungen in verschiedenen Kompetenzzentren, enthält diese Arbeit neben Aufsätzen, die sich mit der Fragestellung „warum gibt es kein QoS im Internet?“ beschäftigen, auch Artikel, die früheren Forschungsprojekten entstammen. Diese Arbeiten liefern einen Beitrag zum Verständnis der Bedürfnisse von Unternehmen. Sie sind nicht als direkter Beitrag zur Beantwortung der Forschungsfrage zu verstehen, sondern bilden das Hintergrundwissen über mögliche Anwendungsszenarien von QoS. Diese Artikel sind in Kapitel 3.2 „IT Trends“ zusammengefasst.

Die erste Veröffentlichung [Hau et al. 2008b] zum Thema Service-Orientierte-Architekturen (SOA) entstand aus dem Projekt „SOA in der Chemischen Industrie“ heraus. Im Rahmen dieses Projektes sollten in verschiedenen Unternehmen der deutschen Chemie-Industrie Fallstudien erhoben werden, um zu zeigen, welche Vorteile SOA [Erl 2005] für diese Unternehmen haben könnte. Während der Zusammenarbeit mit den Unternehmenspartnern entwickelt sich ein Verständnis für die praktischen Schwierigkeiten, eine SOA-Initiative zu starten. Die Veröffentlichung thematisiert dieses Problem und gibt Hinweise dazu, welche Projekte sich für eine Einführung von SOA eignen. Eine für die vorliegende Arbeit relevante Erkenntnis aus diesem Projekt ist, dass eine Service Infrastruktur, die sich nicht nur auf ein zentral gemanagtes Netzwerk beschränkt, sondern weltweit auf verteilte Ressourcen bzw. Service Anbieter zugreifen soll, nur mit einer verlässlichen und Service-Level-Agreement-fähigen Netzwerk-Infrastruktur realisierbar ist.

Nach Abschluss des SOA-Projekts arbeitete der Autor im Competence Center Industrialisiertes Informationsmanagement (CC-IIM) mit. Auch hier stand die Zusammenarbeit mit Partnern aus der Praxis im Mittelpunkt. Ein bedeutendes Ergebnis dieser Arbeit war die Erarbeitung von Beschreibungen für IT-Produkte. Anhand eines Konzepts und mehrerer Beispiele wurde gezeigt, dass sich komplexe IT-Produkte aus einzelnen einfachen Leitungen zusammensetzen lassen, und dass sich die Beschreibung der komplexen Produkte ebenfalls aus einfachen Bausteinen zusammensetzen lässt. Aus dieser Arbeit heraus entstand die Publikation [Brocke et al. 2009], die die Systematik der Beschreibung darlegt.

Die Arbeit über Outsourcing entstand am Rande des CC-IIM aus einem Studentenprojekt heraus. In Zusammenarbeit mit einem der Praxispartner entstand so ein Papier, das nur indirekt mit der Arbeit im CC-IIM zu tun hatte. Die Arbeit beleuchtet die Zusammenhänge zwischen Outsourcing und Standardisierung von IT.

Die dritte Station am IWI war das Projekt „Quality of Service im Internet“. Dieses Projekt wurde im Rahmen einer Kooperation zwischen der TU-Berlin, der Universität St. Gallen und der Deutschen Telekom durchgeführt. In mehreren Teilprojekten wur-

den verschiedene Fragestellungen zum Thema Quality of Service bearbeitet. Hierzu gehörten die Erarbeitung von QoS-Geschäftsmodellen und die Analyse der Synergiemöglichkeiten zwischen Content Delivery Networks (CDN) und Netzunternehmen. Im Rahmen dieser Zusammenarbeit entstanden die Hauptarbeiten dieser Dissertation, die sich mit QoS im Internet und speziell mit den Auswirkungen von CDN auf den Markt für Konnektivität zwischen Providern befassen.

1.3 Forschungsmethodik

Seit längerer Zeit schon tobt zwischen der deutschen Wirtschaftsinformatik (WI) Forschung und der angloamerikanischen Information Systems (IS)-Forschung ein Kampf um die „richtige“ Forschungsmethode. Die WI propagiert hierbei einen praxisorientierten Ansatz, der auf konsensorientiertem Wissensverständnis und Deduktion zur Erkenntnisgewinnung basiert. Gerne wird der Begriff „ingenieurmässig“ verwendet, mit dem gemeint ist, dass weniger der Weg zur Erkenntnis als vielmehr die Erkenntnis an sich in Form einer Lösung für ein gegebenes Problem relevant ist. Die Lösung wird als „richtig“ verstanden, wenn eine relevante Gruppe von Personen der Meinung ist, dass die Lösung sinnvoll ist. Im Gegensatz hierzu dominiert im IS ein behavioristisch-induktiver Forschungsansatz. Hierbei wird viel Wert auf methodisch breit abgestütztes, empirisches Vorgehen gelegt, wodurch die Wahrscheinlichkeit, dass eine Aussage mit der Realität korrespondiert, maximiert werden soll. Die Debatte über die Vor- und Nachteile beider Ansätze wird unter dem Schlagwort „Rigor versus Relevance“ zwischen den Disziplinen geführt. Hierbei wird die Diskussion vereinfachend auf die Fragestellung reduziert, ob Forschung einen Einfluss auf die Realität haben sollte, indem sie Ergebnisse produziert, die von der Praxis als relevant angesehen werden, oder ob sie durch den Einsatz empirischer Methoden Ergebnisse erzielen soll, die verlässliche Aussagen über die Realität machen. Die Auswirkungen der Fokussierung auf jeweils nur ein Forschungsparadigma auf den Zustand der beiden Disziplinen fasst [Frank 2006] wie folgt zusammen: „Information Systems: Global Success but State of Crisis“ und: „Wirtschaftsinformatik: State of Comfort, but Need for Change“. In diesen beiden Überschriften wird das Dilemma der Information Systems Research (ISR, subsumiert WI und IS) deutlich. Während IS das Prädikat der Wissenschaftlichkeit für sich reklamiert und der WI vorwirft, entsprechend „unwissenschaftlich“ zu sein, gelingt es der WI wesentlich besser private Mittel einzuwerben und ihre Studienabsolventen auf dem Arbeitsmarkt zu platzieren [Frank 2006].

Einen Versuch, beide Ansätze zu integrieren und so einen Fortschritt der Forschungsmethoden der Information Systems Research zu bewirken stellt [Hevner et al. 2004] dar. Die Autoren propagieren einen Forschungsprozess, in dem sich Konstruktion und Validierung abwechseln und ergänzen. Hierbei dürfen verschiedene Methoden zu Validierung angewandt werden. [Hevner et al. 2004] erwähnen explizit auch analytisches, simulierendes oder argumentatives Vorgehen und beschränkt sich somit nicht allein auf empirische Methoden zur Validierung. Der geforderte Zyklus aus Konstruktion

und Validierung definiert allerdings sehr hohe Anforderungen an die Fähigkeiten eines einzelnen Forschenden. Spitzenleistungen in mehreren Disziplinen, wie der Konstruktion und er empirischen Validierung, sind wohl nur von sehr wenigen Personen zu erwarten.

Darüber hinaus zwingt der Trend zur reinen Journalpublikation zur Konzentration auf die Least Publishable Units [Broad 1981] als kleinste Einheit publizierbaren Materials. Die Veröffentlichung von Artikeln ist eine Notwendigkeit für jeden Forschenden, und die Publikation von mehr als dem Mindestmass an Ergebnissen in einem einzelnen Artikel ist eine Verschwendung hinsichtlich des möglichen Reputationsgewinns durch Publikationen. Hierdurch besteht ein starker Anreiz, Konstruktion und Validierung gerade nicht in der gleichen Publikation abzuhandeln, sondern auf zwei Publikationen zu verteilen. Aus diesem Anreiz zur Verteilung wiederum entsteht der Anreiz, sich auf ein einzelnes Gebiet zu spezialisieren, da so die Wahrscheinlichkeit steigt, seine Publikationen platzieren zu können. Es ist aus Sicht einzelner Autoren also höchst rational, das von [Hevner et al. 2004] vorgeschlagene Vorgehen gerade nicht zu wählen, da es erhebliche Karriererisiken birgt.

Für den Autor resultiert aus der Verfolgung der Debatte über die „richtige“ Forschungsmethodik die Erkenntnis, dass es sich hier in Teilen um eine dogmatische Diskussion handelt. Die Frage, welcher Ansatz der richtige ist, ist nicht generell entscheidbar, sondern sollte fallweise beantwortet werden. Verschiedene Ansätze erscheinen sinnvoll und notwendig, um neue Erkenntnisse zu generieren. Deshalb plädiert der Autor für eine Pluralität der Forschungsmethoden, wie zum Beispiel von Zelewski in [Winter et al. 2009] für die Wirtschaftsinformatik gefordert. Der Notwendigkeit zur Erzeugung von Publikationen kann sich allerdings kein Forschender entziehen. Die Antwort auf die Frage, wie sich der Widerspruch zwischen Anspruch an die Vielfalt der eigenen Forschungsarbeit und dem Zwang zur Spezialisierung auflösen lässt, muss der Autor (vorerst) schuldig bleiben.

Die Vielfalt denkbarer Forschungsmethoden spiegelt sich in den Arbeiten, die Teil dieser kumulativen Dissertation sind, wider. [Hau et al. 2008a] benutzt einen induktiven, eher dem IS gemässen Ansatz zur Generierung von Wissen, indem versucht wird aus mehreren Fallstudien eine Gesetzmässigkeit abzuleiten. [Brocke et al. 2009] basiert auf einem konstruktionsorientierten Vorgehen, wie es in der WI üblich ist, indem ein Artefakt vorgestellt wird, dessen Validität nur durch seine Anwendung in einem einzigen Projekt gezeigt wurde. [Hau et al. 2008c] andererseits formuliert ein mathematisches Modell, um beobachtete Phänomene zu beschreiben, zu verstehen und zu generalisieren. Dieses Vorgehen kann am ehesten als abduktiv [Reichertz 2003], also als kreativer Prozess, bezeichnet werden. Die einzelnen Arbeiten, die Teil dieser Dissertation sind, können also auch als Experimente verstanden werden, in denen der Autor verschiedene Forschungsmethoden erprobt. Dieses Vorgehen ist zwar höchst ineffizient, da jedesmal neues Wissen aufgebaut werden muss, dient aber gerade am An-

fang einer forscherischen Laufbahn der Orientierung und dient damit einem Zweck, der unmittelbaren Publikationszielen übergeordnet ist.

1.4 Aufbau der Arbeit

Zunächst werden im folgenden Kapitel die relevanten theoretischen Grundlagen geschlossen dargestellt. Danach wird jeder einzelne Artikel vorgestellt, und seine Bedeutung im Kontext der Forschungsfrage dieser Arbeit wird erläutert. Im Anhang sind diejenigen Arbeiten aufgeführt, die noch nicht veröffentlicht wurden.

2 Allgemeine Grundlagen

In diesem Teil werden die theoretischen Grundlagen für alle Artikel dargestellt. Durch diese geschlossene Darstellung wird einerseits die Grundlage für die folgenden Ausführungen und Artikel geschaffen, andererseits wird der innere Zusammenhang zwischen den einzelnen Artikeln besser verständlich. Ein ausführlicherer Teil befasst sich ausserdem mit einer Darstellung der technischen Zusammenhänge, da diese wichtig für das Verständnis sind, aber in keinem der Beiträge explizit thematisiert werden.

Dieser Teil der Arbeit gliedert sich in drei Abschnitte. Zunächst werden die wirtschaftswissenschaftlichen Grundlagen erarbeitet. Es folgen Grundlagen zu wirtschaftlichen Zusammenhängen im Internet. Zuletzt werden technische Aspekte des Internet erläutert.

2.1 Wirtschaftswissenschaftliche Grundlagen

Ein wesentlicher Teil dieser Arbeit befasst sich mit der Frage welche wirtschaftlichen Gründe es dafür gibt, dass Quality of Service im Internet noch nicht weit verbreitet ist. Die im Folgenden erläuterten Theorien bilden die Werkzeuge, mit denen versucht wurde, sich dieser Fragestellung zu nähern.

Vertikale Integration

Man spricht von vertikaler Bindung, wenn ein Unternehmen ein anderes, in der Wertschöpfungskette vor- oder nachgelagertes, Unternehmen kontrollieren kann [Tirole 1988]. Das einfachste Beispiel hierfür ist ein Unternehmen, das ein Monopol für ein Produkt besitzt. Handelt dieses Unternehmen nicht direkt mit den Endkunden, sondern ist ein Zwischenhändler eingeschaltet, so kann der Monopolist verschiedene Preissetzungsstrategien verfolgen. Das einfachste Modell ist die Preisbindung zweiter Hand. Hierbei wird angenommen, dass der Monopolist dem Zwischenhändler Absatzpreise vorschreiben kann. In diesem Szenario wird der Monopolist dem Zwischenhändler den Monopolpreis vorschreiben und durch überhöhte Zwischenhandelspreise den Gewinn des Zwischenhändlers absorbieren.

Ein anderer Fall ist gegeben, wenn sowohl der Hersteller als auch der Zwischenhändler Monopole auf ihren Märkten besitzen, aber keine Kontrolle aufeinander ausüben können. Im schlimmsten Fall treten nun doppelte Gewinnaufschläge [Tirole 1988] auf. Der Hersteller verlangt den überhöhten Monopolpreis von Zwischenhändler und dieser wiederum benutzt diesen überhöhten Preis zur Berechnung seiner Kosten. Mit diesen Kosten kalkuliert der Zwischenhändler dann ebenfalls den Monopolpreis gegenüber den Endkunden. Das Resultat einer solchen Kette von Monopolen ist ein Endkundenpreis, der noch höher liegt als der einfache Monopolpreis. Es entstehen also massive Wohlfahrtsverluste und der Gesamtgewinn der beiden Firmen zusammen, ist geringer, als er es wäre, wenn sie sich koordinieren würden. Hieraus entsteht ein Anreiz zur ver-

tikalen Integration beider Unternehmen, da so der Gesamtgewinn gesteigert werden kann.

Ein weiteres Szenario ist die Erhebung einer Franchise Gebühr [Tirole 1988]. Dieses Modell kann angewendet werden, wenn der monopolistische Hersteller zwar über vollständige Information verfügt, aber keine Kontrolle über seinen monopolistischen Zwischenhändler ausüben kann. In diesem Fall ist es für den Hersteller sinnvoll, kostenbasierte Preise vom Zwischenhändler zu verlangen. Dieser wird diesen Preis als Basis für die Berechnung des Monopolpreises gegenüber den Endkunden heranziehen und so den gewinnmaximierenden Preis verlangen. Um nun den Profit vom Händler abzuschöpfen, kann der Hersteller eine Franchise Gebühr – eine einmalige fixe Zahlung – einführen, die genau so hoch sein muss, wie der Gewinn des Händlers. So umgeht der Hersteller das Problem der doppelten Gewinnaufschläge und kann Monopolgewinne abschöpfen.

Die geschilderten Situationen sind einfach auf Marktsituationen im Internet anwendbar. Hier sind die Anbieter von DSL- oder Kabelanschlüssen für Endkunden die Monopolisten, da sie ein Monopol über Zugang zum Endkunden haben. Die Position des Händlers nehmen die Anbieter von Inhalten ein. Sie können ihr Angebot nur erbringen, wenn sie das Produkt des Monopolisten (Zugang zu den Kunden) einkaufen können. Zwar dreht sich hierbei die Flussrichtung in der Wertschöpfungskette im Vergleich zum Hersteller-Händler Beispiel um, doch sind die Auswirkungen die gleichen wie dort.

Preisdifferenzierung

Preisdifferenzierung bedeutet, dass zwei identische Einheiten eines Gutes an unterschiedliche Kunden zu unterschiedlichen Preisen verkauft werden. Für eine genauere Diskussion siehe [Tirole 1988, Kap. I.3]. Für Unternehmen ist Preisdifferenzierung erstrebenswert, da sie eine bessere Abschöpfung der Zahlungsbereitschaft bei den Kunden ermöglicht. Grundsätzlich muss das Unternehmen hierzu versuchen, seine Kunden in einzelne Teilmärkte zu segmentieren. Ist dies geschehen, kann das Unternehmen auf den getrennten Märkten den jeweils optimalen Preis verlangen. Im Falle eines Monopolisten wird so der Gewinn höher sein, als im Fall eines einheitlichen Preises.

Im Extremfall der perfekten Preisdifferenzierung ist es dem Monopolisten möglich, die gesamte Konsumentenrente (also Nutzen der Kunden minus Preis) an sich zu ziehen, da er von jedem einzelnen Kunden den maximalen Preis verlangen kann. Dieser Fall kann im Internet z.B. eintreten, wenn ein ISP (Internet Service Provider: Eine Firma, die Internetzugang für Endkunden anbietet.) mit einem eindeutig identifizierbaren Inhalte-Anbieter verhandelt. Der ISP besitzt das Zugangsmonopol zum Konsumenten und kann diese Monopolmacht voll ausnutzen, da er den Inhalte-Anbieter identifi-

zieren kann und, falls er die individuelle Zahlungsbereitschaft des Anbieters kennt, diese komplett abschöpfen kann.

Wenn der Hersteller weniger Möglichkeiten hat, seine Abnehmer zu identifizieren, kann es immerhin noch möglich sein, dass er durch das Verhalten der Abnehmer auf deren Zahlungsbereitschaft schliessen kann. Die hierdurch mögliche Preisdifferenzierung zweiten Grades erlaubt die Segmentierung der Nachfrage in *n* Teilmärkte, auf denen der jeweils gewinnmaximierende Preis erhoben werden kann. Ein Beispiel hierfür sind die unterschiedlichen Klassen in Bahn oder Flugzeug. Wenn man „Transport von A nach B“ als das Produkt ansieht, dann kann das Unternehmen die Kunden durch Anreize (breitere Sessel) dazu bringen, sich entsprechend ihrer Zahlungsbereitschaft selbst in die „richtige“ Kategorie einzusortieren. Im Internet entspräche diese Art der Preisdifferenzierung einer Einführung von Qualitätsklassen. Jeder Inhalte-Anbieter könnte dann selbst auswählen, welche Variante des Produkts Datentransport seinen Bedürfnissen entspricht.

Preisdiskriminierung dritten Grades findet statt, wenn die Trennung in verschiedene Teilmärkte nicht durch Selbstselektion stattfindet, sondern wenn das Unternehmen a priori verschiedene Märkte definiert. Ein Beispiel hierfür sind die niedrigeren Tarife für Studenten und Rentner im Theater. Diese Gruppen mit niedriger Zahlungsbereitschaft können von der verdienenden Bevölkerung getrennt werden, und so kann mit einem speziellen Preis ihre Zahlungsbereitschaft besser abgeschöpft werden. Im Internet entspräche dies unterschiedlichen Preisen für unterschiedliche Inhalte-Anbieter, basierend auf der Art der Inhalte, die vertrieben werden sollen.

Zweiseitige Märkte

Im Gegensatz zu den beiden zuvor diskutierten Themen sind „zweiseitige Märkte“ noch kein fest etablierter Teil der industrieökonomischen Lehre. Man spricht von einem zweiseitigen Markt, wenn ein Unternehmen ein Gut anbietet, dass sich zugleich an zwei getrennte Kundengruppen richtet und wenn nicht nur die Höhe der Preise sondern auch ihre Struktur, also die Verteilung auf die beiden Nachfragergruppen eine Rolle spielt [Rochet 2004]. Ein prominentes Beispiel für ein Unternehmen, das einen zweiseitigen Markt bedient, ist die Auktionsplattform eBay. Dieses Unternehmen bedient Käufer und Verkäufer und generiert Gewinne, indem es beide Kundengruppen zusammenbringt. Will man nun ermitteln, welche Preise eBay von seinen Kunden verlangen sollte, so ist das Ergebnis, dass der Gewinn nicht über jede Kundengruppe einzeln maximiert werden soll, sondern dass man den aggregierten Gewinn maximieren muss. Abhängig davon, wie eine Kundengruppe nun auf die Präsenz der anderen Kundengruppe reagiert, können die Ergebnisse dieser Optimierung nun beliebig weit vom Ergebnis der getrennten Optimierung abweichen. Wenn z.B. eine Kundengruppe sehr viel Wert auf die Präsenz der anderen Gruppe legt, so ist es denkbar, dass diese Gruppe sehr hohe Preise bezahlen muss, während die begehrte Gruppe sogar subventioniert wird. Dies ist z.B. in Diskotheken der Fall. Um mehr zahlende männliche Besucher

anzulocken, ist der Eintritt für weibliche Besucher oft sehr günstig. Diese Preisgestaltung resultiert aus der Tatsache, dass der Gesamtgewinn der Diskothek durch diese asymmetrische Preisgestaltung höher ist, als wenn von beiden Kundengruppen gleich hohe Preise genommen würden. In diesem Fall gäbe es nämlich weniger weibliche Besucher, wodurch der Besuch für Männer ebenfalls weniger attraktiv wird.

Ein speziell in der Literatur über zweiseitige Märkte behandeltes Thema sind „competitive bottlenecks", also Marktgegebenheiten, die dem Unternehmen, das die Engstelle („bottleneck") kontrolliert, eine gewisse Monopolmacht geben [Armstrong 2006]. Üblicherweise besteht diese Engstelle aus einem Zugangsmonopol zu Endkunden. Im Internet haben zum Beispiel DSL- und Kabelanbieter (ISP) ein Zugangsmonopol zur ihren Privatkunden. Die Privatkunden haben typischerweise nur einen Anbieter für ihren Internetanschluss, und sie können diesen auch nur mit hohen Hürden wechseln. Gegenüber Inhalte-Anbietern besitzt der ISP also ein Zugangsmonopol zu seinen Endkunden. Dieses Monopol ohne Rücksicht auf die Endkunden auszubeuten ist aber nicht die optimale Lösung für den ISP. Stattdessen muss er beachten, wie viel Wert seine Endkunden auf die Inhalte-Anbieter legen. Sind diese sehr attraktiv, so kann der ISP durch viele Inhalte viele Endkunden anziehen und so wiederum attraktiver für Inhalte-Anbieter werden.

Spieltheorie

Die Spieltheorie ist ein Werkzeug zur Beschreibung von Entscheidungssituationen, die auftreten, wenn mehrere Agenten in gegenseitiger Abhängigkeit stehen und aus mehreren Handlungsalternativen die für sie optimale auswählen müssen. Grundsätzlich lässt sich die Spieltheorie in zwei Hauptgebiete unterteilen: Die kooperativen und die nicht-kooperativen Spiele. In kooperativen Spielen dürfen Gruppen von Spielern Allianzen formen und haben die Möglichkeit, innerhalb der Allianz Ausgleichszahlungen vorzunehmen [Osborne/Rubinstein 1994]. In diesem Gebiet liegt der Fokus also darauf, zu ermitteln, welche Allianzen sich bilden und ob diese stabil sind. In der nicht-kooperativen Spieltheorie hingegen, agiert jeder Spieler individuell, und bindende Absprachen, wie sie für eine Allianz notwendig sind, finden nicht statt. Im Folgenden wird es um die nicht-kooperative Spieltheorie gehen [Fudenberg/Tirole 1991].

Das klassische Lehrbeispiel der Spieltheorie ist das „Gefangenendilemma". In diesem Spiel müssen zwei Straftäter vorm Verhör getrennt entscheiden, ob sie die Tat gestehen oder abstreiten sollen. Streiten beide ab, so erhalten beide eine Geldstrafe für kleinere Vergehen. Gestehen beide, gehen beide für ein Jahr ins Gefängnis. Gesteht aber nur einer, während der andere die Tat abstreitet, so ist der geständige Täter frei, wohingegen sein Komplize für 5 Jahre ins Gefängnis muss. Müssen sich nun beide Verbrecher ohne Kenntnis der Entscheidung des jeweils anderen für eine Strategie entscheiden, so wählen beide die Strategie „gestehen", da dies individuell rational ist. Könnten die beiden aber eine verbindliche Absprache treffen, so würden sie sich beide für „abstreiten" entscheiden, da dies jedem einzelnen ein besseres Ergebnis sichern

würde. Die notwendige Absprache ist aber schwer durchzusetzen, da jeder Spieler durch einseitiges Abweichen die Geldstrafe vermeiden kann. Dieses Gleichgewicht, das sich einstellt, wenn alle Spieler ihre Strategie wählen und jeder individuell rational handelt, wird Nash-Gleichgewicht genannt und ist einer der Grundpfeiler der Spieltheorie. Es ist wird allgemein definiert als eine Spielsituation, in der sich einseitiges Abweichen zu einer anderen Strategie für keinen Spieler lohnt. Im Gefangenendilemma lohnt sich das einseitige abweichen von der Strategie „gestehen“ nicht, da man sein Ergebnis verschlechterte, wenn der andere nicht auch irrationalerweise abwiche.

Die Situation ändert sich etwas, wenn die Spieler sequentiell handeln können, und sie Informationen über die vorgelagerten Handlungen Ihrer Mitspieler haben. In diesem Fall müssen die einzelnen Spieler in jeder Entscheidungssituation abwägen, welches der erreichbaren Spielergebnis im aktuellen Teilspiel (alle möglichen Strategiekombinationen, die mit der gegebenen Vorgeschichte an Zügen der einzelnen Spieler noch erreichbar sind) für sie optimal ist. Ausserdem müssen sie die optimalen Züge der Gegenspieler beachten. Mit diesen beiden Informationen können sie dann ihre optimale Strategie ermitteln. Der Lösungsalgorithmus, der dieses Vorgehen formalisiert, wird Rückwärtsinduktion genannt. Hierbei geht man von den Endzuständen des Spiels aus und überlegt wie der Spieler, der am Zug ist, entscheiden würde. Man sucht als zunächst diejenigen Ergebnisse, die für den Spieler, der den letzten Zug machen kann, optimal sind. Abhängig von dieser Teilmenge an Ergebnissen ermittelt man die optimale Wahl des als zweitletztes Ziehenden und so weiter, bis man beim ersten Spielzug angelangt ist.

Eine Voraussetzung für die beiden vorgestellten Lösungskonzepte ist die „common knowledge“ Annahme. Dies bedeutet, dass man davon ausgeht, dass alle Mitspieler alle spielrelevanten Informationen besitzen. Speziell sind also die Auszahlungen jedes Spielers allen bekannt und alle wissen dies.

Preissetzung in Kommunikationsnetzen

Die Preissetzung in Kommunikationsnetzen ist seit langer Zeit Gegenstand der ökonomischen Analyse. [MacKie-Mason/Varian 1995] war eine der ersten Arbeiten, die sich konkret mit der Bepreisung von Datennetzwerken - und speziell dem Internet - befasste. Die Autoren erstellten in ihrem Aufsatz praktisch das Benchmark für alle zukünftigen Preismodelle, indem sie die theoretisch optimale Lösung ermittelten. Ihr Netzwerkmodell besteht wie das Internet aus Routern, die miteinander verbunden sind. An jedem Router muss ein Datenpaket nun gegen die anderen ankommenden Datenpaketen um die Weiterleitung durch den Router konkurrieren. Welches Paket weitergeleitet wird, hängt hierbei von der Zahlungsbereitschaft des Versenders und der Auslastung des Routers ab. Konkret funktioniert dieser Mechanismus indem der Versender jedes Paket mit einem Gebot versieht. Hat der Router Kapazität für n Datenpakete, so lässt er die n höchstbietenden Pakete passieren und sie müssen den $(n+1)$ höchsten Preis bezahlen. Dieser Mechanismus entspricht einer Erweiterung der Zweitpreisauk-

tion nach [Vickrey 1961]. Da es sich hierbei um eine Auktion zum Einheitspreis (uniform price auction) [Krishna 2002] handelt, ist dieser Mechanismus zwar nicht optimal im ökonomischen Sinn, doch einfach zu berechnen und „ziemlich effizient“, was seinen Einsatz zur Paketpriorisierung in Internet Routern plausibel macht. Obwohl theoretisch ansprechend, gibt es keine kommerziellen Produkte, die diese Idee umsetzen.

Eine andere Richtung verfolgen die Arbeiten von Laffont et al. [Laffont et al. 1998b,Laffont et al. 1998a]. Hier war die zentrale Frage nicht, wie viel die Übertragung eines einzelnen Datenpaktes kostet, sondern wie konkurrierende Firmen gegenseitig die Weiterleitung von Telefonaten verrechnen. Das Kernelement dieser Diskussion ist der Zugangspreis (two-way access charge), den ein Anbieter für die Weiterleitung von Gesprächen des anderen Anbieters verlangt. [Dewenter/Haucap Access Pricing: Theory and Practice 2006] und besonders [Buehler/Schmutzler 2006] bieten einen aktuellen und umfassenden Überblick über den Stand der Forschung in diesem Bereich. [Buehler/Schmutzler 2006] legen hier dar, dass die Ergebnisse von [Laffont et al. 1998b] wenig robust gegenüber einer Modifikation der Annahmen sind. Die Strategie die Kosten des Rivalen durch eine Anhebung des Zugangspreises anzuheben, ist demnach nur unter sehr restriktiven Bedingungen optimal.

Der Forschungsschwerpunkte der Arbeiten über one-way access befasst sich mit dem Zugang einer Firma zur Infrastruktur eines ehemaligen Monopolisten. In Anlehnung an die Literatur über vertikale Integration und „essential facilities“ wird diskutiert, wie die Zugangspreise, die ein in den Markt eintretendes Unternehmen an den Incumbent (historischer Monopolist) bezahlen muss, zu gestalten sind. Ein wesentliches Thema hierbei ist die Erhaltung der Investitionsanreize auf Seite des Incumbent [Gans 2006].

Diese beiden Forschungsgebiete sind noch stark von der Telefonie geprägt. Dies zeigt sich vor allem dadurch, dass es eine abrechenbare Einheit (z.B. die Gesprächsminute) gibt, auf deren Basis sich Preise angeben lassen. [Courcoubetis/Weber 2003] sind im Gegensatz hierzu stärker auf tatsächliche Datennetze fokussiert und beschreiben verschiedenste Preissetzungsmechanismen; unter anderem auch „congestion pricing“, wie von [MacKie-Mason/Varian 1995] vorgeschlagen.

2.2 Internet Ökonomie

In diesem Kapitel wird auf zwei Themen eingegangen, die für das Verständnis der Artikel in Abschnitt 3.1 von zentraler Bedeutung sind. Es geht hierbei erstens um die Geschäftsmodelle von Inhalte-Anbietern. Zweitens wird die Vertragsgestaltung zwischen Internet Service Providern bei Zusammenschaltungsverträgen diskutiert.

Geschäftsmodelle von Inhalte-Anbietern

Ein Inhalte-Anbieter oder englisch Content Provider (CP) ist eine Firma, die einen Dienst über das Internet anbietet. Dies kann z.B. eine Webseite, eine Software-as-a-Service Dienstleistungen, ein Spiele-Server oder eine IP-Telefonie-Lösung sein. Wenn

wir von der konkreten Problematik, dass viele Angebote im Internet noch rein Wagniskapital-finanziert sind und (noch) keine Gewinne erwirtschaften, absehen, so gibt es zwei grundsätzliche Geschäftsmodelle für CPs. Das weitaus wichtigste Geschäftsmodell ist das werbefinanzierte Angebot. Hierbei versucht ein Unternehmen durch das Angebot attraktiver Inhalte viele Endbenutzer anzuziehen. Diese Endbenutzer bezahlen allerdings wenig oder nichts für das Angebot. Die Finanzierung des Angebots wird durch Werbetreibende erreicht, für die die Aufmerksamkeit der Endbenutzer einen Nutzen stiftet, für den sie bereit sind zu bezahlen. Abgeschöpft wird diese Zahlungsbereitschaft, indem der CP Werbeplatz z.B. auf seiner Webseite verkauft. In diesem Geschäftsmodell, das auch demjenigen traditioneller Zeitschriften oder Fernsehsender entspricht, ist der CP ein Plattformanbieter, der einen zweiseitigen Markt (Kapitel 2.1) bedient [Rochet 2003].

Im Gegensatz zum Geschäft mit Privatkunden, das von werbefinanzierten Geschäftsmodellen dominiert wird, sind bei Geschäftskunden bezahlte Inhalte eher üblich. In so einem Geschäftsmodell bezahlt der Kunde ganz klassisch für die Dienstleistung, die der CP erbringt.

Der für diese Arbeit relevante Unterschied zwischen beiden Varianten ist das Optimierungskalkül des CP. Im werbefinanzierten Modell optimiert er über zwei Kundengruppen, wohingegen er im Bezahlmodell seine Preisentscheidung allein auf seine Endkunden beziehen kann.

Die Artikel in Kapitel 3.1 arbeiten heraus, welche Rolle Internetanbieter als weitere vermittelnde Instanz einnehmen und welche Rolle hierbei neue Verteilmechanismen wie CDN und Multi-Homing spielen.

Koordination zwischen Internet Service Providern

Internet Service Provider (ISP) oder Carrier sind Firmen, die die Infrastruktur des Internet besitzen [Noam 2001]. Diese Firmen besitzen die Kabel und Wegerechte, ohne die kein Cyberspace existieren würde. Das Internet besteht aus den Netzwerken vieler unabhängiger Unternehmen, die ihre Netzwerke zusammenschalten, also physische Verbindungen herstellen. In Kapitel 2.3 werden wir genauer auf die technologischen Aspekte eingehen. Dieser Abschnitt behandelt hingegen die ökonomischen Aspekte wie Anreize und Vertragsgestaltung.

Einzelne ISPs sind grundsätzlich eigenständige Unternehmen, die gewinnorientiert handeln. Um aber die universelle Konnektivität, die das Internet bietet, herzustellen, müssen die ISPs zusammenarbeiten. Dies passiert, indem eine Gruppe von etwa zehn sogenannten „Tier 1“ (Ebene 1) ISPs untereinander Peeringvereinbarungen abschliessen. Hierdurch stellen sie untereinander ein komplett vermaschtes Netz her. Jeder dieser T1 ISPs wiederum schliesst Transitverträge mit Tier 2 ISPs. Historisch waren T1-ISP mit Ausnahme von NTT ausschliesslich US Firmen. Andere Firmen sind aber da-

bei, sich diesem Status anzunähern, indem sie versuchen, mit immer mehr T1-Providern Peeringverträge abzuschliessen.

Während Peering Verträge ohne finanziellen Ausgleich funktionieren – zwei T1-Provider einigen sich darauf, Einrichtungen zum Datenaustausch mit einer bestimmten Bandbreite vorzuhalten – begründen Transitverträge ein Anbieter-Kundenverhältnis zwischen T1- und T2-ISP. Der T2-ISP kauft hierbei „universelle Konnektivität“ vom T1-Carrier.

ISPs, die Kunden in einem Transitvertrag sind, haben den natürlichen wirtschaftlichen Anreiz, ihre Transitkosten zu minimieren. Da sich Transitkosten üblicherweise an zur Verfügung gestellter Bandbreite orientieren, müssen die T2-ISPs eine Abwägung treffen zwischen Qualität der Internetverbindung (dies entspricht der Bandbreite pro Endnutzer) und den Kosten hierfür. Der T2 ISP hat also offensichtlich einen Anreiz, die Bandbreite der Transitverbindung nicht beliebig hoch zu wählen. Dies entspricht der Wahl eines Qualitätsniveaus, das nicht perfekt ist.

Bei Peering Vereinbarungen sind die Zusammenhänge anders, doch das Ergebnis das gleiche. Theoretisch kann ein T2 beliebig grosse Bandbreiten von seinem T1 kaufen und so beliebig gute Qualität erreichen. Bei Peerings ist dies nicht möglich, da per Definition keine Zahlungen zwischen den Parteien vorkommen. In diesen Vereinbarungen muss also der Nutzen auf beiden Seiten annähernd gleich sein, damit es zu einer Vereinbarung kommt. Dies führt dazu, dass immer nur der Bedarf an Bandbreite des T1 mit dem niedrigeren Bedarf gedeckt wird, da es für diesen keinen Anreiz gibt in ein Peering über seine optimale Kapazität hinaus zu investieren. Im Allgemeinen sind also Peerings „zu klein“ [Cremer 2000,Noam 2001], wobei dieser Effekt noch dadurch verstärkt wird, dass diverse weitere Erwägungen die Entscheidung über Peering-Bandbreiten beeinflussen. Peerings sind zum Beispiel Machtinstrumente. Ein ISP kann einem anderen ISP schaden, indem er ihn de-peert, also die Datenübertragung blockiert. Im schlimmsten Fall sind so von einem ISP aus nicht alle Internetadressen erreichbar.

2.3 Internet Technologien

Dieses Kapitel rekapituliert die wichtigsten Grundlagen zum Thema Internet Technologie. Zuerst werden Datennetzwerke und Schichtenmodelle behandelt. Die dort geschilderten Zusammenhänge bilden die Basis für die folgenden Themen. Der folgende Abschnitt diskutiert die Struktur des Internet. Es folgt eine Erläuterung der Messgrössen für die Qualität von Datenübertragung und Massnahmen zu deren Einhaltung. Zum Schluss werden mit Content Delivery Networks und Multi-Homing zwei Technologien diskutiert, die von wachsender Bedeutung sind und die Struktur des Internet nachhaltig verändern.

Grundlagen zu Datennetzwerken

Es existieren im Wesentlichen zwei Modelle zur Abstraktion von Datennetzwerken, die von Relevanz sind. Das OSI Schichtenmodell ist ein weitgehend theoretisches Konstrukt. Das TCP/IP Referenzmodell hingegen ist ein Schichtenmodell, das sich an der realen Struktur des Internet orientiert.

Tabelle 1: Gegenüberstellung von OSI- und TCP/IP Referenzmodellen.

<table>
<tr><th>OSI-Schicht</th><th>TCP/IP-Schicht</th><th>Hybridreferenzmodell nach Tanenbaum</th><th>Protokollbeispiel</th></tr>
<tr><td>Anwendung</td><td rowspan="3">Anwendung</td><td rowspan="3">Anwendung</td><td rowspan="3">http</td></tr>
<tr><td>Darstellung</td></tr>
<tr><td>Sitzung</td></tr>
<tr><td>Transport</td><td>Transport</td><td>Transport</td><td>TCP</td></tr>
<tr><td>Vermittlung</td><td>Internet</td><td>Vermittlung</td><td>IP</td></tr>
<tr><td>Sicherung</td><td rowspan="2">Netzzugang</td><td>Sicherung</td><td rowspan="2">Ethernet, DSL</td></tr>
<tr><td>Bitübertragung</td><td>Bitübertragung</td></tr>
</table>

Kapitel 1 aus [Tanenbaum 2000] stellt die beiden Referenzmodelle für Netzwerkarchitekturen einander gegenüber und analysiert ihre Stärken und Schwächen. Der Tenor der Analyse ist, dass das OSI Modell vor allem theoretisch ansprechend ist. Positiv hervorzuheben ist vor allem die klare Definition der Konzepte Dienst, Schnittstelle und Protokoll. Ein Dienst beschreibt, welche Funktionalität eine Schicht zur Verfügung stellt. Die Schnittstellendefinition beschreibt, wie auf diese Funktionalität (den Dienst) zugegriffen werden kann. Das Protokoll beschreibt die interne Funktion einer Schicht, also wie ein Dienst funktioniert. Kritisch in Bezug auf das OSI-Modell merkt Tanenbaum an, dass es überaus komplex und schwer implementierbar ist und darüber hinaus Designfehler enthält, wie zum Beispiel die redundante Definition von Funktionen auf mehreren Ebenen.

Wie schon aus Tabelle 1 ersichtlich ist das TCP/IP-Referenzmodell um einiges weniger komplex, als das OSI-Modell. Darüber hinaus hat es den Vorteil, dass es eine reale Implementierung gibt, nämlich das Internet. Im Gegensatz zum OSI-Modell, das zuerst als theoretisches Konstrukt entstand, wurde das TCP/IP-Modell im Nachhinein so definiert, dass es auf die vorhandenen Protokolle passte. Daher passt es auch perfekt zu den im Internet anzutreffenden Protokollen, eignet sich aber kaum zur Beschreibung anderer Netze.

Um in der abstrakten Beschreibung von Datennetzwerken die Stärken beider Schichtenmodelle zu kombinieren, schlägt Tanenbaum die Verwendung eines Hybridmodells

vor, das in Tabelle 1 dargestellt ist. Die folgende Beschreibung der Funktionen eines Netzwerk-Protokollstapels orientiert sich an diesem Modell.

Auf der Bitübertragungsschicht werden grundlegende physikalische Parameter der Datenübertragung festgelegt. Innerhalb dieser Schicht werden Protokolle definiert, die die Übertragung über Kupfer- oder Glasfasern, terrestrische Funksysteme oder per Satellit regeln. Standards auf dieser Ebene beziehen sich z.B. auf physikalische Parameter des Übertragungsmediums wie die Dämpfung des Leiters oder den Frequenzbereich, in dem Daten übertragen werden. Eine wichtige Funktion auf dieser Ebene ist auch die Aufteilung der verfügbaren Übertragungskapazität auf verschiedene Sender. Grundsätzlich geschieht dies entweder indem verschiedene Sender verschiedene Übertragungsfrequenzen benutzen, oder indem verschiedene Sender nur zu bestimmten Zeiten senden dürfen.

Im kabelgebundenen Bereich für Internetzugang (Access, letzte Meile) spielen heute vor allem zwei Technologien eine Rolle. Über Twisted Pair, also die Telefonleitung, werden Daten per DSL übertragen [ITU 1999]. Bei der Datenübertragung über das Koaxialkabel des Kabelfernsehens wird der Standard DOCSIS (Data Over Cable Service Interface Specification) [ITU 1998] verwendet. In Zukunft werden diese Technologien von Glasfaserinfrastrukturen mit massiv höheren Bandbreiten abgelöst werden. Im drahtlosen Bereich, der für den Internetzugang immer mehr an Bedeutung gewinnt, sind GSM [ETSI 2001], UMTS [ITU 2000], LTE [3GPP 2000], W-LAN [IEEE 2004a] und WiMax [IEEE 2004b] die Standards, die die Eigenschaften der Bitübertragungsschicht definieren. In allen anderen Netzwerkbereichen, also den Aggregations- und Backbonenetzen, kommt heutzutage praktisch ausschliesslich Glasfaser zum Einsatz, und Ethernet [IEEE 1973] setzt sich als Protokoll mehr und mehr durch.

Die Sicherungsschicht hat die Aufgabe, für eine fehlerfreie Übertragung einer Nachricht zu sorgen. Der Dienst, den diese Schicht mit Hilfe ihrer Protokolle zur Verfügung stellt, ist also die zuverlässige Übertragung der an sie übergebenen Daten. Hierzu gehört die Erkennung und Behandlung von Fehlern sowie die Vermeidung einer Überlastung des Empfängers durch einen zu schnellen Sender. Um diese Funktionen zu erfüllen, teilt die Sicherungsschicht den Datenstrom, den sie übermitteln soll, in sogenannte Rahmen - also Datenwörter fester Länge – auf und berechnet Prüfsummen für diese. Daten samt Prüfsumme werden dann zum Empfänger übertragen, der anhand der Prüfsumme Übertragungsfehler erkennen kann. Ein im Internet angewandtes Protokoll, das diese Dienste zur Verfügung stellt, ist das Point-to-Point-Protocol (PPP). Es zerlegt Daten in Rahmen, kümmert sich um die Fehlererkennung und baut Verbindungen zwischen zwei Endpunkten auf.

Eine wichtige Funktion, die bisher ignoriert wurde, ist die Organisation des Zugriffs auf ein geteiltes Medium. Das gerade beschriebene PPP organisiert den Datenaustausch zwischen zwei Computern. Falls diese beiden Hosts aber nicht über eine exklusive Verbindung verfügen, muss geregelt werden, wann das Medium für wessen Da-

tenübertragung zur Verfügung steht. Diese Funktion wird Medium Access Control (MAC) genannt. Ein gängiges Verfahren hierbei ist, dass der Sender ins geteilte Medium lauscht und nur dann sendet, wenn dies kein anderer tut. Diese Protokolle werden CSMA (Carrier Sense Multiple Access) genannt.

Die nächsthöhere Schicht im Hybridmodell nach Tanenbaum ist die Vermittlungsschicht. Diese Schicht ist dafür zuständig, dass Daten den richtigen Weg durch ein Netz finden. Im Internet bedeutet dies dass diese Schicht garantieren muss, dass ein Datenpaket vom Sender über mehrere Router bis zum Empfänger übertragen werden kann. Hierbei muss die Vermittlungsschicht für die Transportschicht eine Abstraktion der Netzwerktopologie und -technologie bereitstellen. Eine wesentliche Designentscheidung, die hierbei getroffen werden muss, ist, ob das Netz verbindungsorientiert (wie zum Beispiel das in Telefonnetzen verwendete Protokoll ATM) oder verbindungslos (wie das Internet Protokoll) arbeiten soll. Verbindungsorientierte Protokolle reservieren vor der eigentlichen Datenübertragung Ressourcen zwischen Sender und Empfänger und erstellen so einen virtuellen Pfad, der von allen anschliessend übertragenen Paketen gewählt wird. Bei verbindungslosen Protokollen hingegen enthält jedes Datenpaket nur die Zieladresse und jeder Router zwischen Quelle und Senke entscheidet lokal zu welchem nächsten Router er das Datenpaket weiterleiten muss, damit es seinem Ziel näher kommt. Im Kontext dieser Arbeit ist hier vor allem interessant, dass verbindungsorientierte Protokolle Quality of Service automatisch garantieren, sobald eine Verbindung zustande kommt. Im Gegensatz hierzu bieten verbindungslose Netze keinerlei Garantie für die Qualitätsparameter des Datenverkehrs.

Auf der Vermittlungsschicht werden auch wesentliche Funktionen zur Steuerung von Überlast wahrgenommen. Router besitzen zum Beispiel Pufferspeicher, um temporäre Lastspitzen ausgleichen zu können. Kommen mehr Datenpakete an, als der Router weiterleiten kann, so werden die ankommenden Pakete zwischengespeichert, bis deren Weiterleitung möglich ist. Eine Massnahme zur Vermeidung von Überlastungen sind ausgefeilte Routing Algorithmen, die ein Datenpaket über diejenige Verbindung weiterschicken, die die schnellstmögliche Weiterleitung verspricht.

Die Transportschicht stellt den Kern des vorgestellten Protokollstapels dar. Sie stellt für die Anwendungsschicht einen einfach benutzbaren Transportdienst bereit. Wie bei der Vermittlungsschicht unterscheidet man auch bei der Transportschicht zwischen verbindungsorientierten und verbindungslosen Diensten, die denjenigen der tieferen Schicht ähnlich sind. Der bedeutende Unterschied zwischen Vermittlungs- und Transportschicht ist, dass die Vermittlung „im Netz“, also von den Routern durchgeführt wird, während die Umsetzung der Transportschicht auf den Hosts, also den ans Netz angeschlossenen Endpunkten der Kommunikation ausgeführt wird. Hiermit ist auch der wichtige Aspekt der Kontrolle verbunden. Während Netzbetreiber die Vermittlungsschicht kontrollieren, haben die Nutzer die Kontrolle über die Protokolle der Transportschicht.

Die Transportschicht stellt (typischerweise als Teil des Betriebssystems) sogenannte Dienstprimitive für die Datenübertragung bereit. Hierdurch können Applikationen der Anwendungsschicht durch einfache Aufrufe von Betriebssystem Funktionen Verbindungen zu entfernten Rechnern aufbauen, mit diesen kommunizieren und die Verbindung wieder abbauen. Sämtliche Komplexität des Managements einer Verbindung zwischen zwei Rechnern, also speziell das nicht triviale Problem des Verbindungsauf und –abbaus, wird komplett gekapselt und vor der Anwendungsschicht versteckt.

Im Internet gibt es hauptsächlich zwei Transportprotokolle. Das User Datagramm Protokoll (UDP) ist ein verbindungsloses Transportprotokoll, das das Internet Protokoll der verbindungslosen Vermittlungsschicht nur minimal erweitert. UDP ist vor allem geeignet, wenn die Kontrolle über den Datenfluss grösstenteils in der Anwendungsschicht passieren soll. Angewendet wird dieses Protokoll zum Beispiel beim Streaming von Multimediainhalten, da es hier kaum ins Gewicht fällt, wenn ein Paket verloren geht. Viel problematischer wäre es, wenn eine Verbindung starke Verzögerung erfährt, weil ein bestimmtes Paket nicht übertragen werden kann.

Das Transmission Control Protocol (Übertragungskontrollprotokoll, TCP) stellt für die Anwendungsschicht einen verbindungsorientierten Dienst zur Verfügung. Da die Vermittlungsschicht im Internet nicht verbindungsorientiert ist, muss TCP aus einem verbindungslosen, unzuverlässigen Dienst einen verbindungsorientierten und zuverlässigen Dienst machen. Hierzu stellt TCP sicher, dass alle Pakete ankommen und dass sie in der richtigen Reihenfolge ankommen. Dies wird realisiert, indem der Empfänger von Datenpaketen zu jedem empfangenen Paket eine Bestätigung schickt. Geht ein Datenpaket verloren, so erkennt der Sender den Verlust am fehlenden Bestätigungspaket und kann nach einer gewissen Zeit das Datenpaket erneut senden. Da der Sender immer erst weiterarbeitet, wenn er eine Bestätigung für die letzte Übertragung erhalten hat, ist die Reihenfolge der Pakete ebenfalls sichergestellt.

TCP spielt auch eine wichtige Rolle bei der Behandlung von Überlastsituationen. Überlastung bei der Datenübertragung kann entweder dadurch entstehen, dass der Empfänger zu langsam ist, oder dass das Übertragungsnetz zu geringe Kapazität hat. Der erste Fall ist einfach lösbar, indem der Empfänger mitteilt, welche Datenrate er maximal verarbeiten kann. Schwieriger ist die Frage der Netzkapazität, da diese unbekannt ist. Hierzu macht der Sender bestimmte Annahmen über die Kapazität der Verbindung und beginnt mit einem „Slow Start“, indem er mit einer geringen Übertragungsrate anfängt und diese zunächst exponentiell, dann linear steigert, bis ein Datenpaket nicht innerhalb der gesetzten Frist bestätigt wird. Sobald ein Datenpaket nicht bestätigt wurde, reduziert der Sender die Datenrate auf den Anfangswert und steigert sie dann wieder, diesmal aber weniger schnell. Eine versteckte Annahme hinter diesem Vorgehen ist, dass ein verlorenes Datenpaket eine Überlastsituation signalisiert. Dies ist sinnvoll, da moderne Glasfaserleiter sehr geringe Verlustraten haben und Paketverlust also praktisch nur auftritt, wenn ein Router überfordert ist und ankommende Pake-

te verwirft. Diese Annahme ist aber nur in leitungsgebundenen Netzen sinnvoll. In Funknetzen gehen Pakete mit hoher Wahrscheinlichkeit verloren, was die Reduktion der Datenrate als Reaktion auf ein verlorenes Paket zu einer schlechten Strategie macht. In diesem Fall sollte der Sender mindestens mit der gleichen Geschwindigkeit weitersenden. Wird die vollständige Transparenz des Protokollstapels gewahrt, weiss die Transportschicht also nicht, um welche Art von Übertragungsmedium es sich handelt, kann sie keine optimale Strategie zur Behandlung von Paketverlusten wählen.

Die Anwendungsschicht zeichnet sich im Internet durch eine enorme Heterogenität aus. Hier finden sich Anwendungen wie das World Wide Web, Email, IP-Telefonie oder Multimedia Streaming. Ein wichtiges Protokoll der Anwendungsschicht, das besondere Erwähnung verdient, ist das Domain Name System (DNS). Dieses besteht aus Servern, die Adresslisten verwalten und so dafür sorgen, dass die für Menschen gut lesbare Adresse www.iwi.unisg.ch in die vom Internetprotokoll benötigte IP-Adresse 130.82.101.29 umgewandelt wird.

Struktur des Internet

Das Internet besteht aus vielen Teilnetzen, die zusammengeschaltet das Internet ausmachen. Im Internet-Jargon wird ein Teilnetz als Autonomes System (AS) bezeichnet, das durch eine eindeutige AS-Nummer identifiziert wird. Ein AS entspricht üblicherweise einem kommerziellen Netzbetreiber (ISP, Carrier, Provider) wie der Swisscom oder akademischen und staatlichen Institutionen (z.B. das Switch Netzwerk in der Schweiz). Die kommerziellen AS stellen hierbei die überwiegende Mehrheit. Die AS bilden eine Klassengesellschaft. Sogenannte Tier 1 (T1)-Carrier bilden eine Clique der etwa 10 grössten Netzbetreiber mit gegenseitigen Peering Vereinbarungen. Die T1-Carrier können so ohne dafür zu bezahlen jeden Internetteilnehmer der Welt erreichen. Unterhalb der T1-Carrier gibt es eine grosse Anzahl T2-Carrier, die von regionaler Bedeutung sind. Diese T2-Carrier haben Transit Vereinbarungen mit den T1-Carriern. Der T1 übernimmt den Verkehr des T2 und sorgt für die globale Konnektivität des T2, lässt sich dies aber bezahlen. Oft kommt es auch vor, dass T2-Carrier untereinander Peeringvereinbarungen schliessen, um so Transitkosten zu sparen. Sogenannte Internet Exchange Points (IX) stellen die physische Infrastruktur für die Verbindung der AS-Netze bereit. Dies sind Rechenzentren, in denen die AS-Netze Router betreiben. Sollen zwei AS-Netze verbunden werden, wird eine Verbindung zwischen zwei Routern hergestellt.

Neben dieser Hierarchie der Teilnetze gibt es auch eine Hierarchie der Netzwerktechnologien. Grob gesagt kann man hier das Kernnetz und das Zugangsnetz unterscheiden. Das Kernnetz besteht aus Glasfaserkabeln mit sehr hoher Kapazität. Diese Glasfaserkabel verbinden Städte, Länder und Kontinente miteinander. Durch den Bauboom der Dotcom Blase ist dieser Bereich heute gekennzeichnet durch massive Überkapazitäten. Diese Glasfaserkabel werden in den Städten in Verteilzentralen mit dem Zugangsnetz verbunden. Obwohl hier in Zukunft auch Glasfaser das dominierende Medi-

um sein wird, werden heutzutage hauptsächlich Twisted-Pair-Telefonkabel und Koaxialkabel eingesetzt.

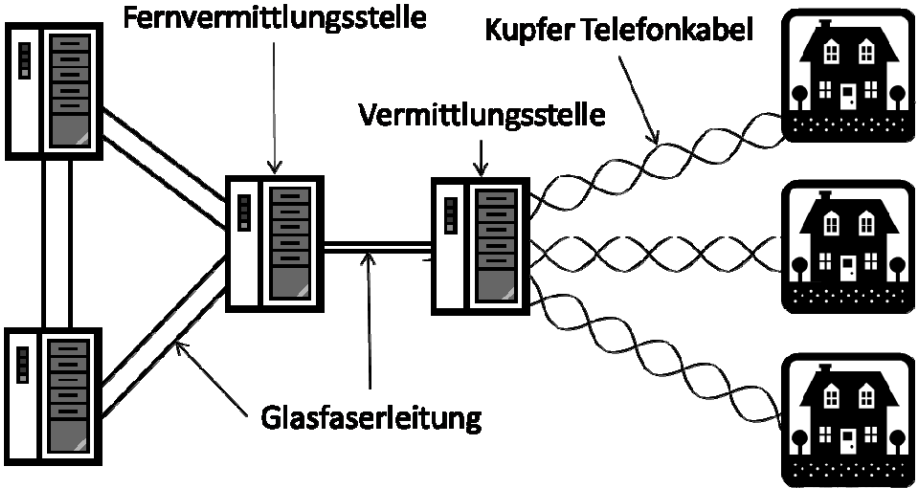


Abbildung 1: Struktur eines DSL-Zugangsnetzes (in Anlehnung an [Tanenbaum 2000])

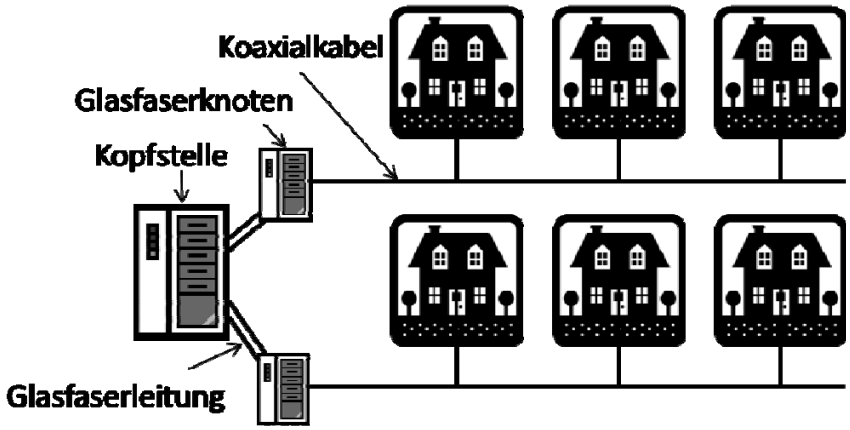


Abbildung 2: Struktur eines Kabel-Zugangsnetzes (in Anlehnung an [Tanenbaum 2000])

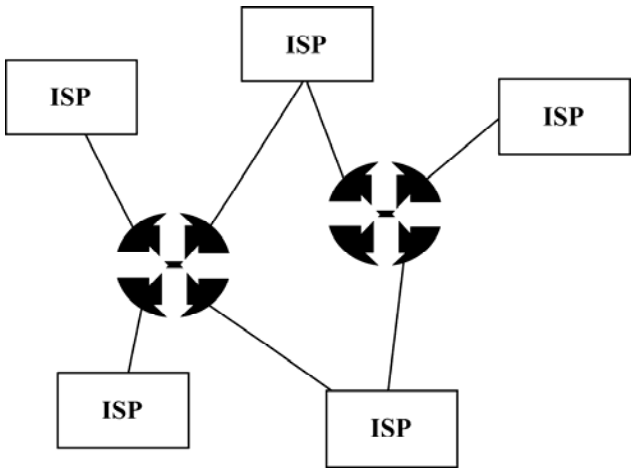


Abbildung 3: Schematische Darstellung des Internet. ISPs schalten ihre Backbones an Peering Punkten zusammen.

Qualitätskenngrössen

Bei der Diskussion über die Qualität von Datenübertragung ist es sinnvoll die zwei Aspekte Quality of Service (QoS) und Quality of Experience (QoE) zu trennen. QoS meint die technischen Parameter der Datenübertragung. Diese sind objektiv messbar und vertraglich regelbar. Im Gegensatz hierzu hängt die QoE, also die „gefühlte“ Qualität des Benutzers nicht nur von der QoS ab, sondern auch von der Art der Anwendung, die benutzt wird, und von den Ansprüchen des Benutzers. Um eine angemessene QoE für seine Kunden sicherzustellen, muss der ISP also pro Anwendungsklasse Qualitätsparameter definieren und kann diese eventuell noch nach Nutzergruppe differenzieren.

Qualitätskriterien im Sinne von QoS sind die Bandbreite, die Latenz, der Jitter und die Verlustrate. Die Bandbreite misst den theoretisch möglichen Datendurchsatz pro Zeiteinheit typischerweise in MBit/s. Die Latenz beschreibt die Verzögerung bei der Übertragung. Latenz entsteht durch die endliche Geschwindigkeit elektromagnetischer Wellen, die Verarbeitungsgeschwindigkeit von Netzwerkkomponenten und die Pufferspeicher von Internet-Routern. Sie wird in Millisekunden (ms) gemessen. Die Kenngrösse Jitter beschreibt, wie stark die Latenz schwankt und wird ebenfalls in ms gemessen. Die Verlustrate steht für den Anteil an Datenpaketen, der bei der Übertragung verloren geht. Diese Verluste sind heute zum einen durch das Übertragungsmedium verursacht, zum anderen durch Überlastsituationen, in denen Router neu ankommende Pakete verwerfen.

Die Anwendungen, die Nutzer über das Internet verwenden, lassen sich nach ihren Anforderungen an die Qualitätsparameter in verschiedene Gruppen einteilen. Zunächst sind elastische von unelastischen Diensten zu unterscheiden. Bei elastischen Diensten spielt die Latenz nur eine geringe Rolle. Bei Diensten wie Email oder ftp spielt es keine Rolle, ob ein Datenpaket 20 oder 200 ms bis zum Empfänger unterwegs ist. Ganz anders hingegen beim Betrachten von Webseiten oder bei interaktiven Diensten wie SaaS und VoIP. Der Nutzer merkt deutlich, wenn eine Website nur mit Verzögerung lädt. Man kann dies beobachten, wenn man die Ladezeiten von Webseiten in Europa und der Westküste der USA vergleicht. Noch extremer ist dieser Effekt bei interaktiven Diensten. Bei SaaS wird die Interaktion mit dem Dienst zäh, da das Programm nicht angemessen schnell auf Nutzereingaben reagiert, bei VoIP müssen die Gesprächspartner warten, bis sie sprechen dürfen. Ab einer bestimmten Latenz werden die interaktiven Dienste weitgehend unbrauchbar.

Ein weiteres Unterscheidungsmerkmal ist die Anforderung an die Bandbreite. Videodienste haben sehr hohe Anforderungen, während Email nur sehr geringe Datenvolumen überträgt.

Der Jitter ist bei Anwendungen wichtig, die Daten als Stream bereitstellen, also alle Audio- und Videoanwendungen. Bei hohem Jitter schwankt die Verzögerung der Datenübertragung und die Reihenfolge der ankommenden Datenpakete entspricht nicht

der Sendereihenfolge, oder es entstehen immer wieder längere Lücken im ankommenden Datenstrom. Bei einem Datei Download ist dies wenig relevant, da die Datei ohnehin erst am Ende des Downloadvorgangs angesehen werden kann. Beim Streaming werden allerdings mit sehr kurzem Zeitverzug die aktuell vorhandenen Daten angezeigt. Kommt ein Datenpaket mit Sprachinformationen also zu spät beim Empfänger an, so bemerkt dieser entweder eine Verzögerung oder, wenn einzelne Pakete fehlen, ein Nachlassen der Verbindungsqualität (z.B. „Knacken in der Leitung“).

Um eine allgemeine Klassifikation der Anforderungen von Anwendungen an die Übertragungsqualität vorzunehmen kann man Anwendungen in folgendes Schema einordnen.

Tabelle 2: Klassifizierung von Anwendungen nach Qualitätsanforderungen.

Klasse	**Beschreibung**	**Beispiel**
Interaktiv	Niedrige Latenz ermöglicht flüssige Interaktion mit anderen Personen oder Programmen über das Netz.	Videotelefonie, SaaS, Spiele
Teilelastisch	Daten werden von einem Server zu einem Kunden gesandt und mit kurzer Verzögerung widergegeben. Pufferungsstrategien beim Kunden gleichen Qualitätsschwankungen der Übertragung aus. Geringe Verzögerungen sind ohne weiteres tolerierbar.	Video on Demand, WWW
Elastisch	Völlig elastischer Dienst, bei dem nur die durchschnittliche effektive Bandbreite zählt.	Download einer Datei, Email

An diesem Schema wird schnell klar, welche technischen Möglichkeiten das Internet heute bietet. Elastische und teilelastische Dienste sind weitgehend etabliert und erfreuen sich grösster Beliebtheit. Begrenzend wirken sich hier allenfalls die verfügbaren Bandbreiten in den Zugangsnetzen der ISPs aus. Dies spiegelt sich in der Debatte über die Finanzierung von „Fibre to the Home“ wider [Elixmann et al. 2008a]. Die interaktiven Anwendungen im Gegensatz fristen ein Nischendasein. Internet-Telefonie ist zwar in Form der Anwendung Skype weit verbreitet, aber nicht in geschäftstauglicher, gleichbleibender Qualität. Betrachtet man die geringe Nutzung interaktiver Anwendungen im Internet, so wird klar, dass hier Handlungsbedarf besteht. Abbildung 4 zeigt wie stark sich die übertragenen Datenvolumina in den unterschiedlichen Segmenten unterscheiden.

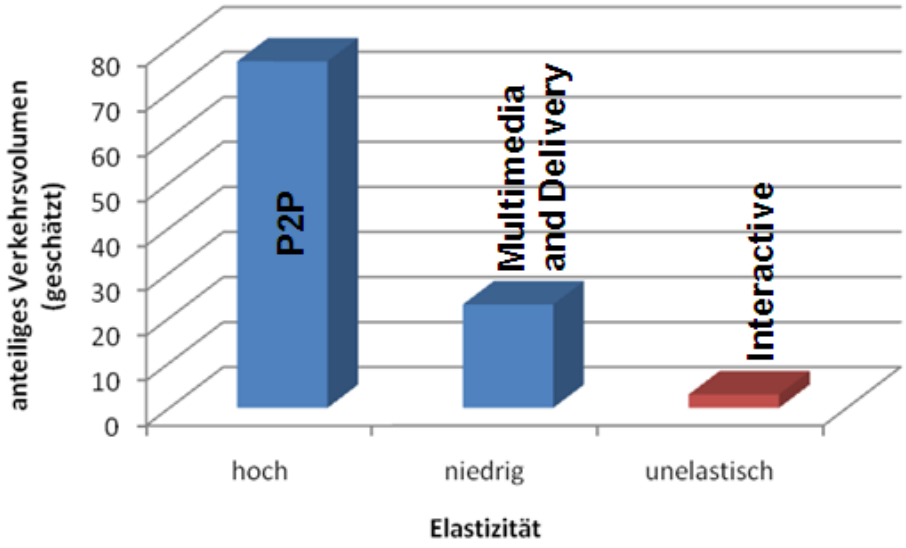


Abbildung 4: Verkehrsvolumen im Internet als Funktion der Elastizität des Dienstes. Quelle: [Ipoque 2007]

Hohe Übertragungslatenzen stellen ein Hindernis für die stärkere Verbreitung von interaktiven Diensten dar. In Zukunft wird die Latenz aber auch für die Qualität der teilelastischen Dienste von immer grösserer Bedeutung sein. Geht man davon aus, dass die Übertragungskapazitäten im Internet weiterhin sehr viel stärker wachsen werden, als die Verarbeitungsgeschwindigkeit von Computern [Tanenbaum 2000], so wird Bandbreite mittelfristig kein begrenzender Faktor für die Geschwindigkeit der Datenübertragung sein. Bei der Übertragung einer – relativ gesehen – kleinen Datenmenge über einen Kanal mit sehr hoher Kapazität, rückt die Latenz als der durchsatzbegrenzende Parameter in den Mittelpunkt. [Tanenbaum 2000, S. 621] gibt hierfür ein interessantes Beispiel: Wollen wir eine 1-MBit Datei durch eine 4000km lange Leitung mit 1 kBit/s Übertragungsgeschwindigkeit und einer Verzögerung von 40ms übertragen, so dauert dies im Idealfall 1000 Sekunden. Im Vergleich hierzu ist die Latenz sehr unbedeutend. Bei einem modernen Glasfaserleiter mit einer Kapazität von 1 GBit/s aber dauert die reine Übertragung von 1 MBit nur noch 1 ms. Nun dominieren auf einmal die 40 ms Latenz die Gesamtzeit der Übertragung und damit die effektive Bandbreite. Es ist somit klar, dass in Zukunft Latenzen zu einem relevanten Problem für die Datenübertragung werden. Sie müssen daher auf das physikalische Minimum reduziert werden.

Qualitätsmechanismen

Dieses Kapitel erörtert Methoden zur Sicherstellung einer bestimmten Übertragungsqualität im Internet. Zuerst werden hierzu generische, allgemein gültige Strategien erläutert ([Tanenbaum 2000] S.439ff). Anschliessend werden zwei Protokolle diskutiert, die im Internet Qualität ermöglichen könnten. Zuletzt werden mit CDN und Multi-Homing zwei tatsächlich verwendete Qualitätsmechanismen vorgestellt.

Die einfachste Methode um die Qualität der Datenübertragung sicherzustellen, ist die Bereitstellung einer sehr grossen Menge an Ressourcen. Diese Strategie wurde z.B. im Telefonnetz benutzt. Hier kommt es kaum vor, dass man den Besetztton hört, weil das Netz überlastet ist. Egal, wie nah das Telefonnetz an seiner Kapazitätsgrenze operiert, die Gesprächsqualität ist immer gleich. Dies wird durch Zugangskontrolle realisiert. Sind alle Kapazitäten ausgeschöpft, so wird kein neuer Verkehr angenommen, sondern es wird das Besetztsignal gesendet.

Speziell für Streaming Dienste interessant sind Pufferungsstrategien zur Qualitätsverbesserung. Hierbei spielt der Empfänger den empfangenen Datenstrom mit einigen Sekunden Verzögerung ab, um so Schwankungen der Übertragungsqualität – speziell Jitter – auszugleichen. Ist die Verzögerung der Datenübertragung grösser als die Länge der gepufferten Daten, kommt es aber auch hier zum Aussetzen der Widergabe.

Traffic Shaping glättet den Verkehr, den eine Quelle erzeugt und kann zum Einsatz kommen, wenn z.B. ein Sender mit einer maximalen Datenrate sendet, die über der maximalen Verarbeitungsgeschwindigkeit eines Routers liegt. Sendet der Kunde kurzzeitig zu viele Daten, so werden diese vom Router gepuffert, und nur mit der niedrigeren Datenrate ins Netz weitergeleitet. Die gepufferten Pakete erfahren natürlich eine Verzögerung, was die Qualität für diesen einen Sender schlechter macht, aber die verfügbare Bandbreite für andere Kunden erhöht, was wiederum einen positiven Einfluss auf deren Qualität hat.

Durch die Reservierung von Ressourcen in einem Netzwerk kann die Übertragungsqualität eines Datenstroms gesteuert werden, indem ein Sender im Voraus berechnet, welche Ressourcen für die Übertragung notwendig sind, und diese reserviert. In diesem Fall kann der Sender beeinflussen, wie schnell seine Daten beim Empfänger ankommen und muss nicht mit Beeinträchtigungen durch zufällig ankommende fremde Datenpakete rechnen.

Zwei für das Internet entwickelte Protokolle zur Sicherstellung von Qualität sind Integrated Services (IntServ) und Differentiated Services (DiffServ) [Wang 2001]. Bei IntServ wird vor dem Versand von Daten eine Route durchs Netz reserviert und alle Pakete werden entlang dieser Route versandt. Dieses Vorgehen entspricht dem Schalten einer virtuellen Leitung durch das Internet und kann benutzt werden, wenn absolute Qualitätszusagen notwendig sind. So kann ein Sender z.B. einen 1 MBit/s Kanal zum Empfänger reservieren und über diese Kapazität verfügen solange sie für ihn reserviert ist. Im Unterschied hierzu reserviert DiffServ keine Kapazitäten, sondern markiert die versandten Pakete. Hierzu gibt es Felder im IP-Protokoll in denen die Qualitätsklasse des Datenpaketes angegeben werden kann. Die Router im Internet leiten dann im Fall einer Überlast diejenigen Pakete mit einer höheren Priorität schneller weiter, als die Pakete mit niederer Priorität. Mit DiffServ sind keine absoluten Qualitätszusagen möglich, da die tatsächliche Qualität nicht nur von der Paketmarkierung abhängt, sondern auch von der Auslastung des Netzwerks. Übersteigt die anfallende

Datenmenge an Paketen der höchsten Qualitätsklasse die Kapazität eines Routers, so erfahren diese Pakete ebenfalls eine Verzögerung. Während also bei IntServ durch die Vorabreservation der Ressourcen eine Zugangskontrolle sicherstellt, dass nur Kapazität zugesagt wird, die auch vorhanden ist, kann DiffServ keine solch feste Zusage machen. Allerdings kann man bei DiffServ aufgrund der Qualitätsklasse, der Konfiguration der Router und des erwarteten Verkehrsaufkommens statistische Aussagen über die zu erwartende Qualität machen. Dies ist für die allermeisten Anwendungen ausreichend, da es der ISP durch die Auslegung der Netzwerkkapazität in der Hand hat, wie oft die Qualitätszusagen nicht eingehalten werden.

Obwohl beide Ansätze vielversprechend sind und speziell fürs Internet entwickelt wurden, ist es nie zu einer flächendeckenden Verbreitung gekommen. Neben Protokollproblemen wie der mangelnden Skalierbarkeit von IntServ (jeder Router muss sich alle Reservationen merken) und der nicht bedachten Zugangskontrolle (wer darf wann welche Qualitätsklasse benutzen und wie wird abgerechnet?) spielten hier auch politische Probleme eine Rolle. Um eine internetweite Weiterleitung von DiffServ Paketen zu ermöglichen, hätten sich alle ISPs der Welt auf einheitliche Qualitätsklassen und zugehörige Qualitätsparameter einigen müssen, was nicht einfach gewesen wäre und auch nie ernsthaft versucht wurde.

Content Delivery Networks (CDN) und Multi-Homing (MH) sind zwei real existierende Technologien, die im Internet benutzt werden, um die Übertragungsqualität zu steigern. CDNs betreiben sogenannte Edge Server. Dies sind Server, die über das gesamte Internet verteilt direkt bei lokalen ISPs in deren Rechenzentren untergebracht sind. Von einem Edge Server aus müssen also nur kurze Wege durch die Infrastruktur eines einzelnen ISP zurückgelegt werden, um Daten an einen Endkunden auszuliefern. Die Firma, die das CDN betreibt, verkauft dann Kapazität auf diesen Servern an Inhalteanbieter, die so ihre Kunden schneller bedienen können. Ein typisches Beispiel ist z.B. die Verbreitung von Multimedia-Inhalten über ein CDN statt nur über einen einzelnen Server. CDNs erzielen also einen Qualitätsgewinn durch die Vermeidung von langen wegen durch das Internet.

Unter MH versteht man die Strategie eines Inhalteanbieters, nicht nur über einen ISP Zugang zum Internet zu erhalten, sondern gleich mehrere ISPs zu haben. Dies erhöht zum einen die Zuverlässigkeit, mit der der Inhalteanbieter Zugang zum Internet hat, andererseits kann diese Strategie auch verwendet werden, um, genau wie beim CDN, direkteren Zugang zu den Endkunden des ISP zu erhalten. Der Inhalteanbieter baut sich praktisch sein eigenes CDN auf.

Während sowohl CDN als auch MH in der Praxis verifizierte Methoden zur Verbesserung der Übertragungsqualität sind, haben sie natürlich bedeutende Nachteile. Vor allem ist zu erwähnen, dass beide Technologien allein darauf beruhen, dass Nadelöhre im Internet wie z.B. die Peerings zwischen ISPs umgangen werden. Garantierte Quali-

tät wie bei IntServ oder wenigstens statistische Zusagen wie bei DiffServ sind nicht möglich.

3 Übersicht der einzelnen Arbeiten

Das vorangegangene Kapitel hat dem Leser eine nichttechnische Einführung in die Konzepte gegeben, die dieser Arbeit zugrundeliegen. Da sich die Arbeit an der Schnittstelle zwischen Technologie und Ökonomie bewegt, wurden diese beiden Themenblöcke behandelt. Der wesentliche Beitrag der obigen Ausführung ist die Bildung eines Rahmens um die folgenden Arbeiten.

In diesem Abschnitt werden nun die einzelnen Artikel vorgestellt. Sie werden kurz zusammengefasst, und ihre Bedeutung im Gesamtkontext der Dissertation wird herausgestellt. In Tabelle 3 folgt zunächst eine chronologische Übersicht über alle Arbeiten und deren Veröffentlichungsstatus. In den folgenden Unterkapiteln werden die einzelnen Artikel thematisch gruppiert besprochen.

Tabelle 3: Arbeiten, die als Teil der kumulativen Dissertation gelten.

Lfd. Nr.	Titel	Autoren	Outlet	Status
1	Where to Start With SOA? Criteria for Selecting SOA Projects	Hau, Ebert, Hochstein, Brenner	HICSS 2008	Published Best Paper Nominee
2	Design Rules for User Oriented IT Service Descriptions	Brocke, Hau, Vogedes, Schindlholzer, Uebernickel, Brenner	HICSS 2009	Published
3	Economic Effects of Multi-Homing and Content Delivery Networks	Hau, Wulf, Zarnekow, Brenner	ITS Regional Conference 2008, Rom	Published
4	Optimizing Investment in Standardization in the Context of Outsourcing – A Game-Theoretical Approach	Hau, Büger, Hochstein, Brenner	Working Paper	Unpublished
5	Quality of Service Delivery: Economics of Multi-Homing and Content Delivery Networks	Hau, Wulf, Zarnekow, Brenner	WI 2009, Wien	Published Best Paper Nominee

Lfd. Nr.	Titel	Autoren	Outlet	Status
6	Price Setting in Two-sided markets for Internet Connectivity	Hau, Brenner	ICQT 2009, Aachen	Accepted Best Paper Award

Die thematische Klammer um sämtliche Beiträge ist die Frage nach der Notwendigkeit und Umsetzbarketi von QoS im Internet. Die Artikel über CDN und Multi-Homing leisten einen Beitrag zum Verständnis real eingesetzter QoS Strategien und deren Auswirkungen auf das Internet. Die Artikel zu Trends in der IT können auf verschiedene Weisen interpretiert werden. Einerseits leisten sie einen Beitrag zum Verständnis wesentlicher Entwicklungen in der Verwendung des Internet und damit einer eventuellen QoS-Fähigkeit. Andererseits können sie auch als Schreibübungen und Orientierungsversuche gesehen werden. Der Artikel in Abschnitt 3.3 stellt den Zusammenhang zwischen den Themenkomplexen her, indem explizit auf die Interdependenz von Serviceorientierung und Qualität der Datenverbindung eingegangen wird.

3.1 Qualität im Internet

Im Englischen wird der Begriff „Peering“ entweder allgemein für „Zusammenschaltungsvereinbarung“ (Peering und Transit) oder für „Netzzusammenschaltung ohne Abrechnung“ (nicht Transit) verwendet. Hier werden beide Bedeutungen verwendet, wobei meist aus dem Kontext hervorgeht, welche Bedeutung die richtige ist. Ist dies nicht der Fall, so wird die Bedeutung präzisiert. Dieser Abschnitt behandelt generell die wirtschaftlichen Aspekte von Zusammenschaltungsvereinbarungen.

Tabelle 4: Arbeiten zu wirtschaftlichen Aspekten von Peerings

Lfd. Nr.	Titel	Autoren	Outlet	Status
3	Economic Effects of Multi-Homing and Content Delivery Networks	Hau, Wulf, Zarnekow, Brenner	ITS Regional Conference 2008, Rom	Published
6	Price Setting in Two-sided markets for Internet Connectivity	Hau, Brenner	ICQT 2009, Aachen	Accepted Best Paper Award

Der Artikel „Economic Effects of Multi-Homing and Content Delivery Networks“ betrachtet Peerings als wesentliche Nadelöhre im Internet und analysiert die Auswirkungen von Umgehungsstrategien. Hierzu wird die Wertschöpfungskette aus Inhalteanbieter, ISP des Inhalteanbieters und ISP des Endkunden analysiert. Speziell wird

untersucht, wie sich die Handlungsmöglichkeiten der einzelnen Teilnehmer verändern, wenn, neben den normalen Peering- und Transitvereinbarungen, auch CDN-Angebote verfügbar sind oder Multi-Homing möglich ist. Diese Situationen werden in einem vereinfachenden Modell diskutiert, das zwei wesentliche Einschränkungen hat. Erstens können Endkunden ihren ISP nicht wechseln, wodurch dieser ein Zugangsmonopol zu „seinen“ Endkunden erhält. Zweitens gilt die Analyse nur für den Markt für bezahlte Inhalte, weil Zahlungen zwischen Inhalteanbieter und Endkunden notwendig sind.

Der Artikel „Price Setting in Two-sided markets for Internet Connectivity“ behebt die letztere der gerade genannten Einschränkungen. In diesem Artikel wird analysiert, wie sich der ISP verhält, wenn er als Plattformanbieter agiert, der versucht, Inhalteanbieter dazu zu bewegen, über CDN oder MH direkteren Zugang zu seinen Endkunden zu erhalten. Hier werden jegliche Zahlungen zwischen Endkunden und Inhalteanbieter explizit ausgeschlossen. Das Ergebnis dieser Analyse ist, dass Endkunden durchaus von solch einem Szenario profitieren könnten, da der ISP seine Gewinne bei den Inhalteanbietern dazu benutzen könnte, den Endkundenpreis für den Internetzugang zu senken.

3.2 IT Trends

Neben der Fragestellung, wie QoS zu realisieren ist und wo die Hindernisse liegen, ist es auch interessant zu fragen, warum QoS überhaupt wichtig ist. Aus der Auseinandersetzung mit drei unterschiedlichen Themengebieten sind drei Artikel entstanden, die spezielle Fragestellungen in unterschiedlichen Forschungsbereichen der IT analysieren. Diese Artikel entstanden vor den Artikeln über QoS. Sie sind daher auch als erste Gehversuche im „Publikationsgeschäft“ zu verstehen und wurden nicht im Hinblick auf QoS verfasst. Trotzdem leisten sie als Analyse möglicher QoS Anwendungsszenarien einen Beitrag zum Umfassenden Verständnis von QoS.

Die Themen Serviceorientierung, Industrialisierung und Outsourcing weisen alle eine gewisse Überschneidung auf. Forschung im Bereich Serviceorientierung beschäftigt sich mit dem Engineering effizienter IT Prozesse. Die Vision besteht darin, einen beliebigen Prozess durch das einfach Zusammenfügen von verteilt abrufbaren Servicebausteinen zusammenstellen zu können. Die Forschung im Bereich Industrialisierung der IT hat zum Ziel, das Management von IT effizient zu gestalten. Statt zu fragen „wie kann man Service vernetzen?“ wird hier gefragt „wie kann man Dienstleistungen effizient erbringen?“. Die Forschung über Outsourcing setzt wieder eine Ebene höher bei der Strategie an und fragt „welche Leistungen werden selbst erstellt und welche werden eingekauft?“. Die drei im Folgenden vorgestellten Artikel geben also einen Querschnitt über Forschungsfragen auf sämtlichen Ebenen der IT. Keiner der Artikel beschäftigt sich explizit mit der Frage nach QoS. Vielmehr sind sie Zeugen einer intensiven Beschäftigung mit den einzelnen Themenkomplexen.

Der Artikel „Where to Start With SOA? Criteria for Selecting SOA Projects“ entstand aus einem Industrieprojekt, das sich mit der Frage auseinandersetzte, wie Serviceorientierung für Unternehmen von Nutzen sein kann. Während dieser Artikel sich nicht mit der Zuverlässigkeit von Services beschäftigt, war dies ein Punkt, der im Projektverlauf aufgeworfen wurde und einen Nachhall in der unten beschriebenen Publikation „Quality of Service Delivery: Economics of Multi-Homing and Content Delivery Networks“ fand.

Der Artikel „Design Rules for User Oriented IT Service Descriptions“ beschäftigt sich mit der Frage, wie Services so beschrieben werden, dass ein Kunde die für ihn relevanten Informationen erhält. Diese Arbeit entstand im Rahmen eines Projektes zur Industrialisierung des Informationsmanagements. Eine wesentliche Erkenntnis hieraus war, dass es für den Erbringer einer Dienstleistung von fundamentaler Bedeutung ist, alle seine Input Faktoren in Ihrer Qualität kontrollieren zu können.

Die Arbeit „Optimizing Investment in Standardization in the Context of Outsourcing – A Game-Theoretical Approach“ benutzt spieltheoretische Methoden um zu verstehen, wie Outsourcingentscheidungen und Standardisierungsentscheidungen zusammenhängen. Die Erkenntnis für diese Arbeit besteht vor allem darin, dass Wertschöpfungsketten räumlich immer weiter auseinanderfallen und daher durch Kommunikationstechnik zusammengehalten werden müssen.

Tabelle 5: Arbeiten, die sich mit Trends der IT befassen.

Lfd. Nr.	Titel	Autoren	Outlet	Status
1	Where to Start With SOA? Criteria for Selecting SOA Projects	Hau, Ebert, Hochstein, Brenner	HICSS 2008	Published Best Paper Nominee
2	Design Rules for User Oriented IT Service Descriptions	Brocke, Hau, Vogedes, Schindlholzer, Uebernickel, Brenner	HICSS 2009	Published
4	Optimizing Investment in Standardization in the Context of Outsourcing – A Game-Theoretical Approach	Hau, Büger, Hochstein, Brenner	Working Paper	Unpublished

3.3 Zusammenführung der Themenkomplexe

Während sich die vorangegangenen Artikel entweder mit der Qualität der Datenübertragung im Internet oder IT-Trends, die die Verwendung des Internet beeinflussen, beschäftigen, ist es das Ziel des Artikels „Quality of Service Delivery: Economics of Multi-Homing and Content Delivery Networks“, diese beiden Themenkomplexe zu-

sammenzuführen. Hierzu wird argumentiert, dass die QoS-Fähigkeit des Internet eine notwendige Voraussetzung für viele absehbare Entwicklungen der IT ist. Weder Serviceorientierung noch das effiziente, industrialisierte Angebot von Software-as-a-Service-Dienstleistungen wird ohne steuerbare Qualitätsniveaus auskommen.

Tabelle 6: Zusammenführung beider Themenkomplexe.

Lfd. Nr.	Titel	Autoren	Outlet	Status
5	Quality of Service Delivery: Economics of Multi-Homing and Content Delivery Networks	Hau, Wulf, Zarnekow, Brenner	WI 2009, Wien	Published Best Paper Nominee

4 Zusammenfassung und Ausblick

Die Beiträge zu dieser Arbeit entsprechen den Praxisprojekten, an denen der Autor beteiligt war. Ein Artikel ist aus dem Projekt „Nutzenpotentiale von SOA in der Chemischen Industrie“ entstanden. Zwei Arbeiten entstanden aus der Arbeit im Rahmen des Kompetenzzentrums „Industrialisiertes Informationsmanagement“. Diese drei Aufsätze bilden den Kontext für die Untersuchung von Quality of Service im Internet. Sie leisten jeweils einen eigenständigen Beitrag zum thematisierten Forschungsbereich ohne direkt Bezug auf QoS zu nehmen.

Die beiden Artikel über CDN und MH thematisieren weitverbreitete Methoden zur Verbesserung der QoS im Internet. Sie zeigen auf, wie sich ISPs verhalten können, um auf Basis der Nachfrage nach Übertragungsqualität Gewinne zu erwirtschaften. Sie leisten damit einen Beitrag zum Verständnis der Faktoren, die die Verbreitung von QoS Technologien beeinflussen.

Der Artikel „Quality of Service Delivery: Economics of Multi-Homing and Content Delivery Networks“ bildet den ersten Schritt zur Zusammenführung der beiden genannten disjunkten Forschungsschwerpunkte. Hier wird argumentiert, dass viele der aktuellen IT-Trends auf QoS angewiesen sind.

Die vorliegende Arbeit bietet eine integrierte Sicht auf betriebswirtschaftliche, ökonomische und technische Zusammenhänge aus der Sicht von Herstellern und Nutzern von Datentransportleistungen. Dem Autor ist keine Arbeit bekannt, die wirtschaftliche und technische Aspekte des Internet integriert und zugleich wissenschaftlich fundiert betrachtet. Die verfügbare Literatur abstrahiert immer von jeweils einem der beiden Aspekte, um den jeweils anderen zu analysieren. Des Weiteren wird sowohl die Perspektive der Anwender (in den Artikeln über IT-Trends), als auch die Sicht der ISPs als Produzenten von QoS (in den Artikeln über MH und CDN) beachtet. Damit hebt sich die vorliegende Dissertation in ihrer Gesamtheit von den bisher verfügbaren Arbeiten über QoS ab und liefert einen wesentlichen Beitrag zur Debatte.

Eine einfache Antwort auf die Frage, wann und wie das Internet QoS-fähig werden wird, ist leider nicht möglich. Auf Betriebswirtschaftlicher Ebene gibt es ein Henne-Ei Problem: QoS wird nicht nachgefragt und damit existieren keine Anreize für ISPs in ihre Netzwerke zu investieren. Wie bereits erörtert bringt NGN hier nicht die endgültige Lösung. Bei QoS-Zusammenschaltung existiert ebenfalls ein Anreizproblem, da sich die Investition, die ein ISP tätigen müsste, immer auch für den Wettbewerber lohnen würde. Auf technologischer Ebene sind vor allem Fragen der Zugangskontrolle und Abrechnung noch ungeklärt. Es existiert keine betriebswirtschaftlich plausible Lösung, die dem Endnutzer eine QoS-Kontrolle ermöglichen würde.

Im Gegensatz zu diesen Hindernissen gibt es im Markt verschiedene Entwicklungen zu beobachten, die ganz klar darauf hindeuten, dass das Internet schon heute und noch

mehr in naher Zukunft verschiedene Qualitätsniveaus bietet und bieten wird. Diese unterschiedlichen Qualitätsniveaus werden aber nicht, wie dies idealtypisch erwartet würde, über Netzwerkprotokolle realisiert, sondern durch CDN und MH, also Technologien, die die neuralgischen Punkte des Internet bewusst vermeiden. Mit diesen Technologien sind zwar keine Qualitätszusagen auf dem Niveau von typischen SLAs möglich, aber sehr viele Anwendungen, vor allem Video Streaming und normale Webseiten, lassen sich mit massiv verbesserter Quality of Experience anbieten. Hiermit stehen professionellen Inhalteanbietern also zwei Möglichkeiten zur Verfügung, die einen sehr grossen Teil des Marktes für verbesserte Übertragungsqualität abdecken. Eventuell wird echtes QoS so zu einem Nischenmarkt.

Aus ökonomischer Sicht hat QoS ein weiteres Problem. Die ISPs bieten momentan sehr margenträchtige Produkte in potentiellen QoS-Anwendungsbereichen an. Dies sind vor allem Sprachtelefonie und Mietleitungen für Unternehmen. Jeder ISP wird sich daher genau überlegen, ob es sinnvoll ist, QoS anzubieten, wodurch es jedem beliebigen Anbieter möglich würde, in diese Märkte einzudringen. CDN und MH im Gegensatz bedienen Märkte, bei denen die ISPs keine Kanibalisierung befürchten müssen. ISPs können sogar durch Joint Ventures mit CDN-Anbietern höhere Gewinne erwitschaften, als dies mit der Netzleistung alleine möglich ist.

Die Aussichten für eine globale QoS Fähigkeit des Internet sind also durchaus kritisch zu bewerten. Im Rahmen von NGN werden die Provider ihre Fähigkeiten diesbezüglich zwar verbessern und zumindest netzintern QoS anbieten können, so dies denn in ihrem Interesse ist, doch das Problem der QoS-Zusammenschaltung wird auch in den NGN-Konzepten nicht gelöst. QoS zwischen Providern wird also kaum auf breiter Front mit Produkten für Endkunden eingeführt werden. In den nächsten Jahren werden vielmehr lokale Lösungen und CDN-Anbieter die QoS-Debatte dominieren.

Die vorgestellten Arbeiten bilden nur den Startpunkt einer Forschungsagenda, die sich um ein besseres Verständnis der wirtschaftlichen Zusammenhänge rund um QoS bemüht. Eine Zusammenführung und Erweiterung der ökonomischen Artikel zu einer integrierten Darstellung der Anreize bei CDN und MH ist notwendig und die Argumentation, dass Services nicht ohne QoS auskommen, muss verfeinert werden. Weitere Schritte umfassen die Konstruktion eines funktionierenden Marktmechanismus für den Austausch von QoS-Verkehr zwischen Providern, sowie die Analyse der technischen Mängel heutiger QoS-Vorschläge. Speziell im Bereich NGN wurden, wie früher schon, ökonomische Fragestellungen und die Interaktion von Anreizen und technischen Möglichkeiten wenig berücksichtigt. Ein eher theoretisches Forschungsfeld tut sich in der weiteren ökonomischen Betrachtung von Peering-Vereinbarungen auf. Der State of the Art wird hier stark von einem Denken in den Dimensionen der Telefonie (speziell die Abrechnung nach Zeit oder Anzahl der Telefonate) geprägt. Ob diese Literatur im Kontext des Internet noch zeitgemäss ist, sollte hinterfragt werden.

Anhang A. Komplette Publikationsliste des Autors

Lfd. Nr.	Titel	Autoren	Outlet	Status
1	Where to Start With SOA? Criteria for Selecting SOA Projects	Hau, Ebert, Hochstein, Brenner	HICSS 2008	Published Best Paper Nominee
2	Design Rules for User Oriented IT Service Descriptions	Brocke, Hau, Vogedes, Schindlholzer, Uebernickel, Brenner	HICSS 2009	Published
3	Economic Effects of Multi-Homing and Content Delivery Networks	Hau, Wulf, Zarnekow, Brenner	ITS Regional Conference 2008, Rom	Published
4	Optimizing Investment in Standardization in the Context of Outsourcing – A Game-Theoretical Approach	Hau, Büger, Hochstein, Brenner	Working Paper	Unpublished
5	Quality of Service Delivery: Economics of Multi-Homing and Content Delivery Networks	Hau, Wulf, Zarnekow, Brenner	WI 2009, Wien	Published Best Paper Nominee
6	Price Setting in Two-sided markets for Internet Connectivity	Hau, Brenner	ICQT 2009, Aachen	Accepted Best Paper Award
7	Specifying Enabling Services in Telecommunications Services Systems	Wulf, Hau, Zarnekow, Brenner	AMCIS 2009	Accepted
8	Business Models for the IP-Based Distribution of Services	Wulf, Hau, Zarnekow, Brenner	Working Paper	Unpublished

9	Potenziale der Produktionsplanung und –steuerung bei IT-Dienstleistern	Ebert, Vogedes, Hau, Uebernickel, Brenner	10. Paderborner Frühjahrstagung	Published
10	Economic Effects of Multi-Homing and Content Delivery Networks	Hau, Wulf, Zarnekow, Brenner	Telecommunications Policy	Revise and Resubmit

Anhang B. Unveröffentlichte Arbeit

Lfd. Nr.	Titel	Autoren	Outlet	Status
4	Optimizing Investment in Standardization in the Context of Outsourcing – A Game-Theoretical Approach	Hau, Büger, Hochstein, Brenner	Working Paper	Unpublished

OPTIMIZING INVESTMENTS IN STANDARDIZATION IN THE CONTEXT OF OUTSOURCING - A GAME-THEORETICAL APPROACH

Abstract

Due to scale effects, outsourcing a standardized business process costs less than outsourcing a specific one. For companies running individual processes supported by legacy software, the question is whether an investment in standardizing that process prior to outsourcing is efficient. We model this problem as a multi stage game and derive the following results: Firstly, it is ex ante rational to invest in standardization to reduce the lock in with the internal provider. Secondly, the prices of the service providers do not matter for the ex ante investment decision. Thirdly, the price gap between internal and external provider will reflect exactly the lock-in in the form of switching costs.

We use an extensive game to model the multi period decision problem and then combine backward induction and optimization to find the optimal investment decision.

The principal implication of our research is the guidance we provide on which information is necessary for the standardization and outsourcing decisions at each stage of the game.

We are not aware of any prior work analyzing the interdependence between standardization and outsourcing that gives clear guidance on which information is necessary for an optimal decision.

Keywords: C72 - Noncooperative Games, M15 - IT Management, L86 - Information and Internet Services; Computer Software.

Introduction

As a benchmark for service provisioning, internal IT departments have to face the price of external, market driven firms. However, the internal supplier has an edge over the external provider by being able to support the customer's individual processes. If the customer wants to externally source any of the IT-services supporting their individual processes it needs to invest in standardizing its processes, and thus making them compatible with solutions readily available on the market. We assert that standardization lowers the cost of switching to an external provider since non-standardized processes require strong customization resulting in high costs.

This work supposes that the outsourcing process has three stages. First, the customer decides about whether to standardize a certain process or not. Then the in-house provider that has been providing the service so far and a prospective external provider get to make bids for the price for which they will provide that service. Lastly the customer decides whether or not to outsource. Put in a nutshell, the research question this paper

answers is:”How strongly should processes be standardized when outsourcing of the IT-services is an option?”

In order to answer this question, this paper employs methods borrowed from game theory to analyze the decision that needs to be made by the outsourcing firm. With this tool we are able to generate results that have relevant implications for today’s business practices.

Our first finding is that the optimal amount of money is invested in standardization if the sum of standardization cost and switching cost is minimal. In particular, this means that the decision is independent of the prices the providers offer. Secondly, no matter how the outsourcing organization behaves, it will always pay a markup on the market price – either through elevated internal prices or through costs for switching to the other provider. Lastly, it is rational for the potential outsourcer to simulate low switching costs and high investments in standardization in order to increase perceived pressure on the internal IT providers.

The paper is structured as follows: First we review relevant literature on game theory, outsourcing and standardization. Then we present a formal model which we will use to analyze the decision problem. In the subsequent section we discuss our model and demonstrate several important implications as well as discussing the limitations of the model. We also provide two extensions that increase the practical relevance of our model. We show how the base model can be extended to a fully integrated internal IT provider and external IT service markets without perfect competition. We then conclude and give some hints to possible extension of this research.

Theoretical Background

Game Theory

We only make slight reference to game theory and use it mainly for the description of our problem. The most important building block of this work that is borrowed from game theory is the concept of extensive form games and backward induction as their main solution method. For all of the following material refer to [Fudenberg/Tirole 1991] for further details.

Loosely speaking, a game is a situation in which a set of involved parties (called players) make decisions (called strategies) and finally receive a payoff, which depends on the decisions of all players. It is assumed that each player behaves rational is the sense that he aims to maximize his payoff. The set of strategies of each player might be discrete (e.g. saying yes or no) or continuous (e.g. choosing a price). Players might know about the decisions other players have already made or not (this can be used to model a situation in which players make their decision at the same time). Our example will be of the type where players make their decisions in several consecutive steps and we assume that each player has full information about the decisions taken so far. Hence,

the game can be illustrated in a decision tree (which is the so called extensive form of the game).

Analyzing the game, we want to identify stable situations which are likely to be the outcome when all parties behave rational. Here, we use the concept of Nash equilibria. Such an equilibrium is a situation, where no party is able to increase its payoff by changing its own strategy unilaterally.

In order to give an example, suppose a company A holds a monopoly in a certain market resulting in a profit of 10 units. A new player B contemplates entering the market which would cost a investment of 2 units. In case B enters, A can either cooperate with the new player and share the profit or start a fight for market shares. The fight would reduce the total profit, and we assume that the profit would collapse to 2 units, one for each.

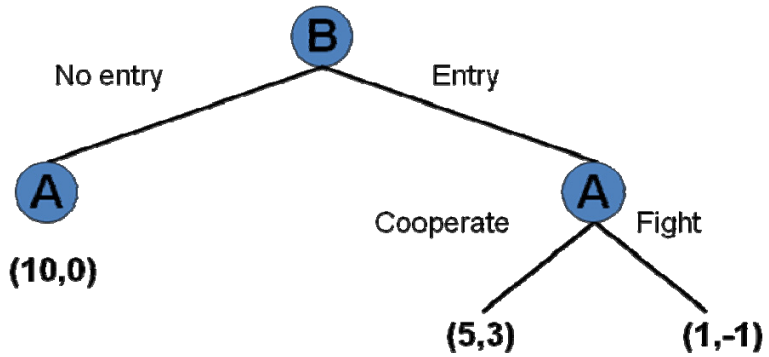


Figure 1

In case B decides to enter the market, A can fight or cooperate. A fight would drive B in a situation (payoff -1) worse than if she had stayed out (payoff 0), but it would also hurt A. Since we assumed that A is rational and wants to maximize his outcome, A will certainly prefer cooperation (with a payoff of 5) rather than a fight (with a payoff of 1). Now, we can analyze with a backtracking algorithm: B knows that A is a rational player and will, therefore, prefer cooperation. Hence, B will obtain a payoff of 3 in case she enters the market, which is more than the payoff 0 in case she stays out. Hence, B will enter the market and A will cooperate.

This game has the Nash equilibrium in the strategy ‘B enters the market and A cooperates’. If either A or B chose a different strategy, each will reduce the payoff (A would obtain 1 instead of 5 and B 0 instead of 3). It turns out that to fight against a possible market entry is an empty threat.

Outsourcing, Standardization and Switching Costs

[Dibbern et al. 2004] offer a comprehensive overview of the burgeoning literature on outsourcing. The authors present a comprehensive overview of the research on out-

sourcing of information systems. Research questions and corresponding works with relevance for this paper are:

- Why to outsource ([Cheon et al. 1995], [Earl 1996], [McLellan et al. 1995])
- Which kind of outsourcing to choose ([Chalos/Sung 1998], [Lacity 1995])
- How to outsource ([Klotz/Chatterjee 1995], [Quinn 2000], [Whang 1992])
- What are the effects of outsourcing ([Aubert et al. 1996], [Lacity et al. 1996], [Hirschheim/Lacity 1998])
- Which kind of IS functions to outsource ([Fitzgerald/Willcocks 1994], [Grover et al. 1994], [King/Malhotra 2000])

Dibbern et al. (2004) also analyze the toolset used to analyze these questions. The most frequently used tool is empirical analysis of outsourcing deals. All other tools such as agency theory or game theory and in general formal modelling are more rarely used to understand outsourcing.

In context of this research, especially the link between outsourcing success and degree of standardization is relevant. Literature suggests that there is a positive impact of process standardization on the success of outsourcing [Wüllenweber et al. 2008]. This implies that in case of non-standardized processes, there is an incentive for the outsourcing company to first invest in process standardization and then outsource the corresponding IS functions. However, the literature doesn't give any advice on investment decisions regarding process standardization prior to outsourcing deals.

The work on "a transaction cost model of IT outsourcing" by [Aubert et al. 1996] provides the foundation for our modeling of the transaction costs of outsourcing (the variables w(s) and s below). Their work establishes that investments in outsourcing are largely sunk and the amount is negatively correlated with the intensity of outsourcing.

Lock in and switching costs are the main keywords when searching for the effects of standardization. Good sources for this body of literature are [Klemperer 1995]. However these works are largely concerned with consumer markets in which two providers are faced with questions of compatibility of their products. Phenomena such as the standard wars for video recording can be explained within these frameworks.

The relevance of that work for ours is established by [Whitten/Wakefield 2006] who attempt to measure "switching costs in IT outsourcing". They disaggregate switching costs into costs for search and evaluation of a new solution, uncertainty costs and so on. They thus provide a more in depth analysis of the cost w(s) introduced in our model below. Their concept of switching costs corresponds to the aggregate $w(s)+s$ explained below.

A Formal Model to Capture the Standardization Decision

This section introduces an extensive game to model the interaction between the company that is considering outsourcing of an IT service (the customer C), the internal IT

provider (called P_o) and an external IT provider (called P_1). The service to be outsourced may be any standardizable business process that is supported by IT. For the purpose of this model we assume the internal IT department to be a separate entity that optimizes its revenue function regardless of the overall objective function of the company it is part of. This is plausible in so far as many IT service organizations are run as independent parts of the organization (possibly a shared service center). This independent unit then optimizes its targets which are set by a central entity.

While it is possibly difficult to set the right targets for these internal IT providers, they have two advantages over any external provider of IT services: Firstly they have exclusive knowledge about their customer that an external company cannot come by. Secondly, the internal provider can be considered an incumbent; it has been there before any external competitor has had the chance to compete with it. The external provider can be any firm able to provide the IT service in appropriate quality. In a first analysis we will consider this external firm to be active in a perfectly competitive market. Thus the price p it offers can be considered the lowest possible (price = marginal costs) and one decision dimension is removed from the game. The external provider is said to have no knowledge about the internal structure and will, thus, offer a market price p independent on the decisions of the other parties. It can be assumed that p is below the price which C had to pay to his internal provider in the past – simply because p will be at the marginal cost level of the external provider which is reasonable for a service provider competing with other external providers for the customer's business.

Hence, we only have two players in the game, the company C acting as customer and the internal provider P_o choosing its offered price.

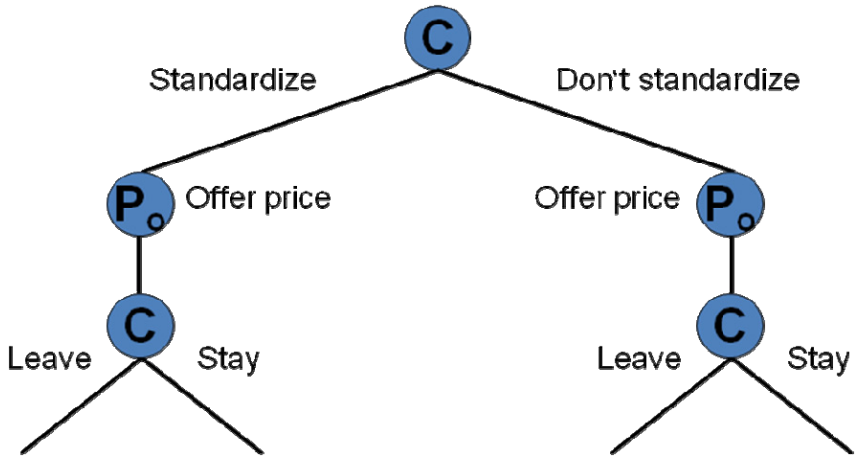


Figure 2

In the topmost vertex the customer C gets to decide whether to standardize and, in case C standardizes, how much money s ($s \geq 0$) is invested. For example, the process, which should be standardized, could be a payment process which could be modified so that it can be supported by available ERP solutions without major modifications to the ERP suite. Of course it would take time and money to change a legacy process supported by

a custom system to conform to some industry standard codified by the process variants supported by the ERP. Clearly, investment in standardization lowers the switching effort w(s) which C has to pay if she decides to accept the offer of the external provider.

In the next step, the providers have to offer a price for providing the IT-service to the customer C. As said above, the external provider will choose the market price p. Knowing the standardization effort C has made, the potential switching cost w(s), his internal cost base c_0 and the market price p, the internal provider chooses its offer p_0. Once this offer is made, the customer C gets to decide whether it stays with its internal provider or switches to the external one. This is depicted by the last pair of vertices.

This setup defines the sequence of actions the players in this game may take. The next step is to define the possible outcomes of this game. This is depicted in figure 3:

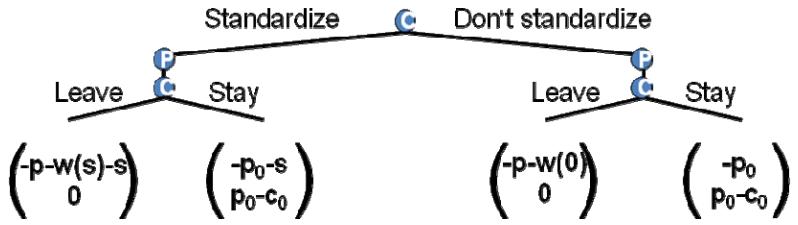


Figure 3

The upper line within each pair of braces shows the payoff to the customer C, the lower line that to the internal provider P_0. If the customer C decides to stay, P_0 wins the offer and has a benefit of $p_0 - c_0$, which is the difference between the money she gets and his cost base. In this case, C has to pay the offered price p_0 plus the prior standardization effort s. If C decides to leave, P_0 is out of business – which is shown by the 0 payoff. In this case, C pays the market price *p*, the switching cost *w* and the potential standardization effort *s*. Note that this is true for both sub-games ("standardize" as well as "don't standardize"). They payoffs on both sides are identical except for the fact that on the right side $s = 0$.

Note also that the choice of p_0 might depend on the investment in standardization, which means that $p_0 + s$ is not necessarily above p_0. This is because p_0 is a function of $w(s)$. The more the customer has invested to lower her switching costs, the lower the price the internal provider can charge.

Given this payoff structure we now apply backward induction. Starting with the final payoffs we determine the rational course of action of each player contingent on what she knows about the other player's actions. Let's first have a look at the left half of the game tree. Since the customer has the last decision we first look at her payoffs. Her decision criterion is

$$\min\{p + w(s) + s, p_o + s\}. \tag{1}$$

Therefore she will choose to leave if the internal price p_o is greater than $p + w(s)$. With this information the two providers set their offer prices. Assuming that the external

price p is a fixed number given as the outcome of competition, the internal provider will set a price $p_o = p + w(s)$ if this is feasible (price is higher than its cost: $p_o > c_o$).

Now look at the right half of the game tree. Again we first look at the payoffs to the customer. Her target function now is:

$$\min\{p + w(0), p_o\} \tag{2}$$

Since she has not invested in standardization ($s = 0$), w attains its maximum value at this point. The decision of the customer depends again on the very same reasoning as above. She will stay with the internal provider only, if $p_o \leq p + w(0)$. In the following step the internal provider will set its price according to $p_o = p + w(0)$.

Now with both sub-games (half-trees) solved we can put the result together and determine the criterion for the standardization decision of the customer. In both sub-games her payoff is $-(p + w(s) + s)$ (where $s = 0$ on the right side).

The first conclusion we can draw from this result is, that this standardization decision does not at all depend on the projected behavior of the internal or external IT-providers. In both games the payoff for the customer only depends on the market price for the IT-service and the standardization effort measured by the investment s. Therefore, in order to make an optimal standardization decision the customer only needs to solve the problem

$$\min_s w(s) + s \tag{3}$$

since C cannot influence the market price.

Interpretation

Equation (3) offers an important insight since it shows that standardization of an IT-service should not be based on the price of the providers. Neither the cost for the internal provider's service nor the possible market price is relevant to an optimal decision. The sole decision criterion is that shown in equation (3). The target function can be interpreted as the sum of the lock-in effect (i.e. height of the switching costs) $w(s)$ and the cost for reducing that lock-in s. Thus the customer minimizes the "pain" of buying from its internal IT-department but will in effect always stay with the internal supplier as long as that supplier is able to offer a price that fulfills the condition $c_0 < p_0 \leq p + w(s)$ and is thus able to stay below the external price without losing money. If this is not given, the internal provider can be considered to be "much more" inefficient than the external market and the customer should switch. Viewed from another perspective this implies that a customer switches from the internal to an external provider if and only if the internal provider is truly inefficient.

The second interesting insight from the model is that there is a "natural" markup by a rational internal IT-provider on the external price. This is not necessarily caused by inefficiency but is a purely rational result of the action space we defined for the game.

This markup is measurable in real life as the difference between the internal and external price of an IT-service. Since we assumed that the internal provider knows about the customer's cost of switching to an external provider, the comparison of prices yields a very easy estimation of the size of the lock in, because $p_0 = p + w(s) \Leftrightarrow p_0 - p = w(s)$. Thus, the difference $p_0 - p$ can be seen as an indicator for the switching cost $\mathrm{w(s)}$. We note, however, that C should be able to derive $\mathrm{w(s)}$ from its own IT knowledge. The reason is that the equation $p_0 = p + w(s)$ holds only if the internal provider P_0 assumes that C knows $\mathrm{w(s)}$. In case C does not know about its real switching cost, P_0 has a chance to cheat, making C believe that the switching cost is higher than it really is, and offering $p_0 = p + w(s) + \Delta$ with some positive Δ.

However, the customer might also try to cheat in letting the provider believe that she invested a higher amount $\mathrm{s'}$ than she actually did. Since $w(s)$ decreases as a function of s, we obtain $w(s') < w(s)$. If P_0 believes in s', she will offer a price $p_0 = p + w(s')$ and has a disadvantage of $\Delta' = w(s) - w(s') > 0$.

By analyzing the model presented we have established that the customer always pays $w(s) + s$ above the market price p. Therefore the question "should the customer standardize its process?" must be rephrased into "how much should the customer invest in standardization". The answer to this question is that the customer should invest that amount s in standardization which minimizes the sum of the switching costs $w(s)$ and the investment s. The target function is shown in equation (3).

Extension of the Model

One Balance Sheet

We have just established that one should always invest such an amount s that equation (3) is minimized. However, this result is only true with an internal provider whose revenue is not part of the customer's balance sheet. With the internal providers profit being part of the customer's revenue, we have to reinterpret the payoffs in figure 2. The customer C does not only consider her payoffs but the sum of the firm's payoff and that of the provider as the decision relevant figure. Numbering the outcomes from left to right with one to four, we have the following results as also shown in figure 4:

- $-p - w(s) - s$
- $-p_0 - s + p_0 - c_0 = -s - c_0$
- $-p - w(0)$
- $-p_0 + p_0 - c_0 = -c_0$.

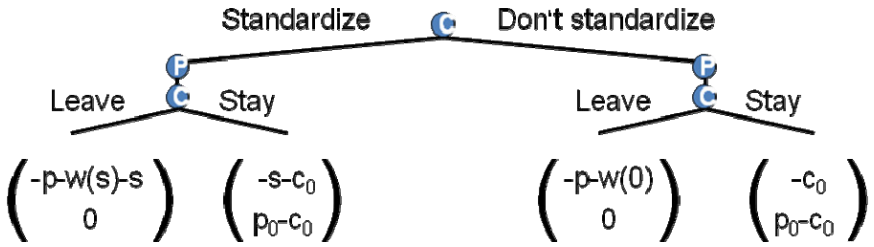


Figure 4: Aggregate payoffs at the firm level and perceived provider profit.

If C stays with her internal provider, she pays $c_0 + s$. Since c_0 is a given constant, C can simply optimize by not investing in standardization at all, i.e. $s=0$. If C decides to switch to an external provider, the total cost is again $p + w(s) + s$ and s should be chosen in a way that $w(s) + s$ is minimal. Assume that s^* minimizes $w(s) + s$. Hence, C pays either c_0 (in case she stays with the internal provider) or $p + w(s^*) + s^*$ (in case she switches). Thus, C has to compare these two values and stay with the internal provider without doing any standardization if

$$c_0 \leq p + w(s^*) + s^* \tag{4}$$

and invest s^* and switch to an external provider if $c_0 > p + w(s^*) + s^*$. It follows that in this situation, the customer has to evaluate first which is the optimal level of standardization but then also compare the market price to the cost of internal production and make sure that the gap is greater than the switching cost. Clearly, the internal price p_0 is not relevant for the consolidated balance sheet of C and is, thus, not relevant for the decision. However, in a real firm, p_0 influences the level of IT usage since it is the price perceived by the users of the IT services.

Small Markets

In a small market there is no market price that can be assumed to be equal to marginal costs and given. However, this does not change the model very much. Consider a second price procurement auction where the lowest bidding firm wins the contract. It has been shown by [Vickrey 1961] that in a second price auction it is a dominant strategy for all bidders to bid their true value, that is they bid the amount at which their profit is zero. If they win the auction they get the amount bid by the second lowest bid. In the scenario with a perfect outside market and at least two external competitors (we do not allow side agreements or any other complications), they both bid their marginal costs and one of them gets the contract by chance getting marginal cost and thus making zero profits.

On the other hand, if there is only one external provider of the service, the situation is more difficult to analyze by identical in outcome to the base case analyzed above. Consider first that the customer announces that it will grant the contract to the lowest bid at the second lowest (the other) price (There are plenty of collusive outcomes that

we will ignore). Now the internal provider might feel tempted to lower its price down to the market price and thus the customer would get away by paying the internal provider the market price without the markup. However, this is not a credible course of action by the customer. It has a commitment problem. If the internal price is less or equal to the external price plus the cost of outsourcing, it is not rational to leave. Therefore this case can be ruled out. Then the only credible strategy for the customer is to conduct the auction based on total cost, i.e. on price plus outsourcing costs. Now the situation is as in the base case. The external provider must bid its marginal cost (see [Vickrey 1961] for why this is the only rational course of action) and the internal provider does not even need to be strategic since it can just bid any value below external price plus switching costs and always win the contract for $p + w(s)$. Thus we have shown that the assumption of a given external price is no restriction of the model. It holds for any number of external providers between one and infinity.

Conclusions and Outlook

This paper provides clear guidance on the amount of up front investments in standardization of processes that might be outsourced in a later period. The base case analyzes this problem under the assumption that the internal provider is rather independent of its internal customer without a central authority that dictates prices. In such a scenario, the information requirements necessary for an optimal investment strategy are very low. Optimizing the target function $\min_s w(s) + s$ can be done without any knowledge of market prices for services. Secondly, it is clear that a customer with an internal provider will always pay the switching cost w(s) independent of her decision to outsource / switch. The consequence of this is that a markup the internal provider charges is not necessarily a sign of inefficiency but a rational choice. However, this markup might lead to inefficient IT usage within the firm since the price is higher than the efficient price which sends the wrong signal to the users of the IT services. Thus, the markup can cause a competitive disadvantage from inefficient resource usage.

Lastly, leaving the model world and interpreting the results further, it is rational for a real world customer to make the providers believe that w(s) is very low, i.e. that she has invested a lot to lower switching costs, while internal providers gain an advantage in overestimating the switching costs.

The first extension shows, that information requirements are tougher when a central entity optimizes across the internal provider's profit as well as the price paid internally. In addition to optimizing its investment level as above, the customer now needs to know internal and external prices and compare these savings w(s). Only if equation (4) holds, outsourcing and the upfront investment s are actually beneficial.

The small markets extension illustrates, that the results hold for external market prices as well as for only one or a few external providers. This section draws from auction theory to proof that there are no rational outcomes but those presented. This is an im-

portant extension, since frequently there is no real market for outsourcing but only a few vendors compete for a contract.

Limitations of this work are inherent in the modeling approach. A mathematical model can never capture the richness of reality. That a clear cut division between two decision periods can never be made in practice is self understood. Furthermore we assumed that the prices offered and the costs of outsourcing are common knowledge. While this might be an approximation of reality, these pieces of information will be among the best guarded ones in any real world scenario.

These limitations lead to possible extensions of the presented work. Two major extensions appear important. Firstly an iterated game would more realistically model the nature of outsourcing decisions. In such a scenario the customer continuously evaluates her options and decides on how to proceed in the next period. Such an extension would reduce the abstraction artifacts due to the simple two period model. The second major extension would involve a modification of the information structure of the game. Allowing less knowledge for the participants would probably alter their action space and give raise to many new strategic considerations. However, the clear and simple results obtained under complete information would probably not be possible.

In contrast to the available literature which is mostly descriptive or explanatory in nature, this work is a step towards giving decision makers clear guidance on their investment strategy for outsourceable business processes. To the authors knowledge there is no prior work that achieves a similar outcome.

Literaturverzeichnis

[3GPP 2000]

3GPP, TS 36.201, http://www.3gpp.org/ftp/Specs/archive/36series/36.201/, 27.03.2009

[Amante et al. 2006]

Amante, S., Bitar, N., Bjorkman, N., Callon, R., Chan, K., Charny, A., Davie, B., McDysan, D., Fang, L., Inter-provider Quality of Service, Quality of Service Working Group, MIT Communications Futures Program (CFP), 2006

[Armstrong 2006]

Armstrong, M., Competition in two-sided markets., in: RAND Journal of Economics, 37, 2006, Nr. S. 668-691

[Aubert et al. 1996]

Aubert, B. A., Rivard, S., Patry, M., A transaction cost approach to outsourcing behavior: Some empirical evidence, in: Information & Management, 30, 1996, Nr. 2, S. 51-64

[Broad 1981]

Broad, W. J., The publishing game: getting more for less, in: Science, 13, 1981, Nr. S. 1137-1139

[Brocke et al. 2009]

Brocke, H., Hau, T., Vogedes, A., Schindlholzer, B., Uebernickel, F., Brenner, W., Design Rules for User-Oriented IT Service Descriptions, Proceedings of the 42nd Annual Hawaii International Conference on System Sciences (CDROM), Hawaii, Computer Society Press, 2009.

[Buehler/Schmutzler 2006]

Buehler, S., Schmutzler, A., On the Role of Access Charges Under Network Competition, in: Dewenter, R., Haucap, J. (Hrsg.), Access Pricing: Theory and Practice, Emerald Group Publishing, 2006, S. 121-148

[Chalos/Sung 1998]

Chalos, P., Sung, J., Outsourcing Decisions and Managerial Incentives, in: Decision Sciences, 29, 1998, Nr. 4, S. 901-919

[Cheon et al. 1995]

Cheon, M. J., Grover, V., Teng, J. T. C., Theoretical perspectives on the outsourcing of information systems, in: JIT. Journal of information technology(Print), 10, 1995, Nr. 4, S. 209-219

[Courcoubetis/Weber 2003]

Courcoubetis, C., Weber, R., Pricing communication networks, Wiley Hoboken, NJ, 2003

[Cremer 2000]

Cremer, J. R., Patrick & Tirole, Jean, Connectivity in the Commercial Internet, in: The Journal of Industrial Economics, 48, 2000, Nr. 4, S. 433-472

[Dewenter/Haucap Access Pricing: Theory and Practice 2006]

Dewenter, R., Haucap, J., Access Pricing: Theory and Practice, Emerald Group Publishing, 2006

[Dibbern et al. 2004]

Dibbern, J., Goles, T., Hirschheim, R., Jayatilaka, B., Information systems outsourcing: a survey and analysis of the literature, in: ACM SIGMIS Database, 35, 2004, Nr. 4, S. 6-102

[Earl 1996]

Earl, M., The Risks of Outsourcing IT, in: Sloan Management Review, 37, 1996, Nr. 3, S. 26-32

[Elixmann et al. 2008a]

Elixmann, D., Ilic, D., Neumann, D. K.-H., Plückenbaum, D. T., The Economics of Next Generation Access - Final Report, 2008a

[Elixmann et al. 2008b]

Elixmann, D., Kühling, J., Marcus, S., Neumann, K., Plückebaum, T., Vogelsang, I., Anforderungen der Next Generation Networks an Politik und Regulierung, 2008b

[Erl 2005]

Erl, T., Service-oriented architecture: concepts, technology, and design, Prentice Hall PTR Upper Saddle River, NJ, USA, 2005

[ETSI 2001]

ETSI, 3GPP TS 45.001: GSM, http://www.etsi.org/WebSite/Technologies/gsm.aspx, 27.03.2009

[Feldmann 2007]

Feldmann, A., Internet clean-slate design: what and why?, in: ACM SIGCOMM Computer Communication Review, 37, 2007, Nr. 3, S. 59-64

[Fitzgerald/Willcocks 1994]

Fitzgerald, G., Willcocks, L., Contracts and Partnerships in the Outsourcing of IT, SOCIETY FOR INFORMATION MANAGEMENT, 1994.

[Frank 2006]

Frank, U., Towards a Pluralistic Conception of Research Methods in Information Systems Research, ICB, 2006

[Fudenberg/Tirole 1991]

Fudenberg, D., Tirole, J., Game Theory, MIT Press, 1991

[Gans 2006]

Gans, J. S., Access Pricing and Infrastructure Investment, in: Dewenter, R., Haucap, J. (Hrsg.), Access Pricing: Theory and Practice, Emerald Group Publishing, 2006, S.

[Grover et al. 1994]

Grover, V., Cheon, M., Teng, J., A descriptive study on the outsourcing of information systems functions, in: Information and Management, 27, 1994, Nr. 1, S. 33-44

[Hau et al. 2008a]

Hau, T., Ebert, N., Hochstein, A., Brenner, W., Where to Start With SOA? Criteria for Selecting SOA Projects, Proceedings of the 41st Annual Hawaii International Conference on System Sciences (CDROM), Hawaii, Computer Society Press, 2008a.

[Hau et al. 2008b]

Hau, T., Hochstein, A., Brenner, W., Nutzenpotentiale von SOA in der Chemischen Industrie, Vertrauliches Projektergebnis,

[Hau et al. 2008c]

Hau, T., Wulf, J., Zarnekow, R., Brenner, W., Economic Effects of Mult Homing and Content Delivery Networks on the Internet, in: Proceedings of the 19th ITS European Regional Conference in Rome, 2008c, Nr. S.

[Hau et al. 2009]

Hau, T., Wulf, J., Zarnekow, R., Brenner, W., Quality of Service Delivery: Economics of Multi-Homing and Content Delivery Networks, Wien, 2009.

[Hevner et al. 2004]

Hevner, A., March, S., Park, J., Ram, S., Design science in information systems research, in: Management Information Systems Quarterly, 28, 2004, Nr. 1, S. 75-106

[Hirschheim/Lacity 1998]

Hirschheim, R., Lacity, M., Reducing information systems costs through insourcing: experiencesfrom the field, 6, 1998.

[IEEE 1973]

IEEE, 802.3 Ethernet, http://www.ieee802.org/3/, 27.03.2009

[IEEE 2004a]

IEEE, IEEE 802.11

LAN/MAN Wireless LANS, http://standards.ieee.org/getieee802/802.11.html, 27.03.2009

[IEEE 2004b]

IEEE, IEEE 802.16: Broadband Wireless Metropolitan Area Network, http://standards.ieee.org/getieee802/802.16.html, 27.03.2009

[Ipoque 2007]

Ipoque, Internet Traffic Study, http://www.ipoque.com/userfiles/file/p2p_study_2007_abstract_de.pdf, 27.03.2009

[ITU 1998]

ITU, Recommendation J.112: DOCSIS, http://www.itu.int/rec/T-REC-J.112/en, 27.03.2009

[ITU 1999]

ITU, Recommendation G.992.1: ADSL, http://www.itu.int/rec/T-REC-G.992.1/en, 27.03.2009

[ITU 2000]

ITU, Recommendation M.1457: 3G, http://www.itu.int/rec/R-REC-M.1457/e, 27.03.2009

[King/Malhotra 2000]

King, W., Malhotra, Y., Developing a framework for analyzing IS sourcing, in: Information & Management, 37, 2000, Nr. 6, S. 323-334

[Klemperer 1995]

Klemperer, P., Competition when Consumers have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade, in: Review of Economic Studies, 62, 1995, Nr. S. 515-539

[Klotz/Chatterjee 1995]

Klotz, D. E., Chatterjee, K., Dual sourcing in repeated procurement competitions, in: Management science, 41, 1995, Nr. 8, S. 1317-1327

[Krishna 2002]

Krishna, V., Auction theory, Academic press, 2002

[Lacity et al. 1996]

Lacity, M., Willcocks, L., Feeny, D., The Value of Selective IT Sourcing, in: Sloan Management Review, 37, 1996, Nr. S. 13-25

[Lacity 1995]

Lacity, M. C. W., Leslie P., Interpreting information technology sourcing decisions from a transaction cost perspective: Findings and critique, in: Accounting, Management and Information Technologies, 5, 1995, Nr. S. 203-244

[Laffont et al. 1998a]

Laffont, J.-J., Rey, P., Tirole, J., Network Competition: I. Overview and Nondiscriminatory Pricing, in: The RAND Journal of Economics, 29, 1998a, Nr. S. 1-37

[Laffont et al. 1998b]

Laffont, J.-J., Rey, P., Tirole, J., Network Competition: II. Price Discrimination, in: The RAND Journal of Economics, 29, 1998b, Nr. S. 38-56

[MacKie-Mason/Varian 1995]

MacKie-Mason, J. K., Varian, H. R., Pricing the Internet, in: Kahin, B., Keller, J. (Hrsg.), Public Access to the Internet, London, UK: Prentice Hall, 1995, S. 269-314

[Marcus/Elixmann 2008]

Marcus, S., Elixmann, D., The Future of IP Interconnection, 2008

[McLellan et al. 1995]

McLellan, K., Marcolin, B. L., Beamish, P. W., Financial and strategic motivations behind IS outsourcing, in: JIT. Journal of information technology(Print), 10, 1995, Nr. 4, S. 299-321

[Noam 2001]

Noam, E., Interconnecting the Network of Networks, MIT Press, 2001

[Osborne/Rubinstein 1994]

Osborne, M., Rubinstein, A., A Course in Game Theory, MIT Press, 1994

[Quinn 2000]

Quinn, J., Outsourcing Innovation: The New Engine of Growth, in: Sloan Management Review, 41, 2000, Nr. 4, S. 13-28

[Reichertz 2003]

Reichertz, J., Die Abduktion in der qualitativen Sozialforschung, Leske+ Budrich, 2003

[Rochet 2003]

Rochet, J. C. T., J., Platform Competition in Two-Sided Markets, in: Journal of the European Economic Association, 1, 2003, Nr. S. 990-1029

[Rochet 2004]

Rochet, J. C. T., J., Defining Two-Sided Markets, Toulouse, France: IDEI, mimeo, January 2004

[Tanenbaum 2000]

Tanenbaum, A. S., Computernetzwerke, Pearson Studium, 2000

[Tirole 1988]

Tirole, J., The Theory of Industrial Organization, MIT Press, 1988

[Vickrey 1961]

Vickrey, W., Counterspeculation, auctions, and competitive sealed tenders, in: Journal of Finance, 16, 1961, Nr. 1, S. 8-37

[Wang 2001]

Wang, Z., Internet QoS: Architectures and Mechanisms for Quality of Service, Morgan Kaufmann, 2001

[Whang 1992]

Whang, S., Contracting for software development, in: Management Science, 38, 1992, Nr. 3, S. 307-324

[Whitten/Wakefield 2006]

Whitten, D., Wakefield, R. L., Measuring switching costs in IT outsourcing services, in: The Journal of Strategic Information Systems, 15, 2006, Nr. 3, S. 219-248

[Winter et al. 2009]

Winter, R., Krcmar, H., Sinz, E., Zelewski, S., Hevner, A., Was ist eigentlich Grundlagenforschung in der Wirtschaftsinformatik?, in: Wirtschaftsinfomatik, 2009, Nr. 2, S. 1-9

[Wüllenweber et al. 2008]

Wüllenweber, K., Beimborn, D., Weitzel, T., König, W., The impact of process standardization on business process outsourcing success, in: Information Systems Frontiers, 10, 2008, Nr. 2, S. 211-224

Anhang C. Veröffentlichte Arbeiten

Lfd. Nr.	Titel	Autoren	Outlet	Status
1	Where to Start With SOA? Criteria for Selecting SOA Projects	Hau, Ebert, Hochstein, Brenner	HICSS 2008	Published Best Paper Nominee
2	Design Rules for User Oriented IT Service Descriptions	Brocke, Hau, Vogedes, Schindlholzer, Uebernickel, Brenner	HICSS 2009	Published
3	Economic Effects of Multi-Homing and Content Delivery Networks	Hau, Wulf, Zarnekow, Brenner	ITS Regional Conference 2008, Rom	Published
5	Quality of Service Delivery: Economics of Multi-Homing and Content Delivery Networks	Hau, Wulf, Zarnekow, Brenner	WI 2009, Wien	Published Best Paper Nominee
6	Price Setting in Two-sided markets for Internet Connectivity	Hau, Brenner	ICQT 2009, Aachen	Accepted Best Paper Award

Im Folgenden sind die bereits veröffentlichten Arbeiten im Format des Zielmediums beigefügt.

Where to Start with SOA

Criteria for Selecting SOA Projects

Thorsten Hau, Nico Ebert, Axel Hochstein and Walter Brenner
Institute of Information Management
Iniversity of St. Gallen, Switzerland
Email: thorsten.hau@unisg.ch

***Abstract*—The concept of service oriented architectures has received considerable attention lately. Despite the hype our experience from five major German chemicals companies shows that firms do not blindly adopt the concept but want to see its value for their organizations. This paper identifies criteria for projects that should serve especially well as first proof of concept SOA implementations. Therefore the paper compares project goals to SOA benefits and requirements and deduces decision supporting criteria. Complex projects with many involved stakeholders and high risk of changing requirements, to name a few of the criteria, are more likely to profit from SOA than small simple projects with few involved people. The criteria are then applied in five cases to evaluate projects the authors analyzed. The application shows that the criteria give a good indication for or against SOA adoption in a project.**

I. Introduction

Service Oriented Architecture (SOA) is more than a buzzword used by marketing departments to sell IT products. Further down in chapter II-B we will show that the relevant aspects of SOA are those that enable a firm to modularize its IT application landscape. While this modularization makes managing an IT landscape less hazardous, the process of disentangling today's tightly coupled applications is a tedious one.

Assuming that SOA will enter the average organization through small and rather isolated projects, there is an obvious conflict of interest between project management that needs to achieve short term goals and the extra investment needed to adhere to SOA design principles. In this work we will derive a set of criteria that helps to find a starting point for the adoption of SOA ideas within a company. In order to achieve this we look for projects within a company for which SOA adoption yields immediate benefits. Our goal in this is to ease the tension between long term global benefits of SOA and short term local needs of project management.

A basic assumption made in this work is that there are two possible scenarios for the adoption of SOA ideas. In a top down scenario, top management would realize the potential of SOA and conduct a high level project to establish SOA paradigms within the organization. This scenario is not unrealistic and there are several examples like Deutsche Post (the German Post, a company that boasts its SOA competence) where SOA was introduced this way [1]. However, the bottom up approach appears to be the more relevant scenario because often SOA might not be considered to be of strategic importance. The examples of top level SOA projects mostly represent instances where the inability to modify the IT landscape had become a pressing problem which in turn made SOA the remedy to an acute problem top management was facing. Supposing that in most firms there are no such acute pain points and SOA in turn is not a top priority, one must consider a gradual scenario in which SOA enters an organization through many small projects with little resources to spare. In these projects SOA would typically be a second order goal. For example introducing a new logistics tool could have one requirement that reads: make functionality available via a specified interface. After several such projects one would have two parallel worlds: an old one with close coupling, dependencies and uncontrollable complexity and a new (SOA-) world that can easily be controlled and adapted. The new world would gradually expand while the old part would become less important. As promoted by [2] we consider this soft introduction of SOA the most feasible approach and this work develops a set of criteria which can be used to evaluate whether a project is suitable for a proof of concept SOA implementation.

II. Theoretical Background

A. Project Management

The "iron" goals of project management are time, scope and money. These criteria form the magic triangle within which the project management has to find an optimal compromise [3], [4]. These criteria have been modified, extended and rearranged. For example, [5] emphasized that the customers point of view must be integrated into a project success metric. They also argue, that success can not be easily defined and that it depends on the perspective of the judging person, whether a project can be considered successful or not. Interestingly the authors also name "preparing for the future" as a relevant aspect. However, despite all these efforts to provide different perspectives on project management, some sort of general agreement among practitioners seems to exist that the three goals named above are still the most important ones. One variation which this article will use is the split-up of scope into quantity (which we will call scope) and quality. With this differentiation we get the following four dimensions to judge a project's success:

Time: Can the project be finished on time?

1530-1605/08 $25.00 © 2008 IEEE

Scope: Can the project achieve the goals that were set? This is a quantitative measure that doesn't take into account the quality of a solution.
Money: Does the project stay within budget?
Quality: Is the delivered output of acceptable or the defined quality?

Project management will commonly be evaluated along these four criteria. Since the evaluation of a project's success usually happens upon its conclusion, we will assume the average project manager to optimize his short term success. It follows that everything that makes a project more expensive or longer will be judged as having a negative impact on his chances for success and thus be avoided. Tools, on the other hand, that reduce the spendings for a certain task or improve the quality that can be achieved, will be used if their potential can be made plausible.

B. SOA

Service oriented architecture (SOA) is a very ill defined buzzword in the IT community. This is largely due to the fact that every player in the market for IT uses its own definition of SOA. SAP for example sells its ERP as SOA, IBM sells databases and middleware as SOA and BEA sells a source code repository as SOA. To clarify our understanding of SOA we first explain its two building blocks:
Architecture is a term that can not be properly defined without putting a lot of contextual meaning into the definition. In the realm of large distributed software systems like the application landscapes of large companies the term architecture can have two meanings. The first means the model according to which different elements relate to each other [6]. The second meaning is the concrete instantiation of such a model and thus is synonymous to 'all applications, hardware and their relations to each other'. In the following we will use both meanings since the context will clarify which one is intended.
Services are the envisioned building blocks of an SOA (model and instantiation). A service encapsulates a certain functionality like providing customer data and makes it available through a well defined interface. This point of view switches the focus away from applications towards functions that need to be performed. Consequently a business process as a series of tasks to be performed can be composed of different services performing these tasks [7].

On the level of principles, SOA can be considered as the transfer of object oriented programming (OO) paradigms, i.e. objects hide information and encapsulate functions, to the world of enterprise systems. Translating the principles of OO-programming to SOA speech one gets a definition of SOA like the one provided by [2]: Service Oriented Architecture (SOA) combines component orientation with loose coupling and external process control into a method to design system landscapes. The objects in OO programming are the components of a system landscape, loose coupling is achieved through encapsulation, information hiding and external process control. Since external process control is rather a means to achieve loose coupling than and end in itself, we can modify the above definition and say that the basic principles underlying SOA are:

- Component Orientation and
- Loose Coupling.

These two paradigms are supposed to lead the way to more manageable system landscapes. However, in order to put these two abstract requirements into practice one must go one step further and define a set of concrete measures to achieve these two goals. Focusing on large software landscapes one can identify the following rules necessary to construct SOA compatible architectures [1]:
Interface Orientation corresponds to the OO paradigm of encapsulation. How a functionality is executed doesn't matter to the user of the functionality. In order to put this principle into practice it is necessary to have well defined service specifications that describe the service's interfaces, capabilities and availability.
Standardization between the components of an SOA is necessary to be able to integrate different elements into a new functionality. Standardized ways of interaction make sure that one software component can exchange data with another over the same interface. In analogy to electricity this means that all applications use the same plug and socket. Today it is common that one firm uses US-style plugs and European sockets within one domain.
Autonomy and Modularity are necessary to achieve component orientation. Autonomy in this context means that a Service is an entity that can execute a certain task without relying on other services. There should be no interdependencies between services and furthermore the communication between the services should be loosely coupled (i.e. message based and stateless) [7].
Business Orientation is the translation of object orientation to the world of enterprise systems. Object orientation tried to draw the link between the real world and the programming abstraction by using objects as representations of real things. In this line of thinking, a service should represent a meaningful abstraction of a business related real object. A service providing customer data might be such a meaningful entity.

Having discussed the different aspects of SOA we will now turn to the question why all this effort should be made. Therefore we follow [8] and name the four main benefits of adopting SOA:
More agility will make a firm able to adapt faster to new requirements. If a process changes, for example because of new regulatory requirements, one can simply use the available services in a different order without causing major changes in the software. Agility also finds its manifestation in smaller projects.
Less complexity makes IT infrastructure easier to handle because each component can be considered individually. Changes in the implementation of a component don't affect the users of that component since they only know the interface. Furthermore changes should not have impact anywhere else in the software landscape since there should be no uncontrollable dependencies.

Increased reusability is possible when software components can be used in several contexts and identical tasks don't need to be implemented several times. The service get_customer_data for example can be used by several consumers.
Better interoperability is a likely benefit of SOA because components of a software landscape can more easily interact and exchange information since only interfaces and middleware need to be considered instead of whole implementations.

One aspect all these benefits have in common is that they are likely to only arise in the long term and can not be realized instantaneously because the architecture of an enterprise system landscape needs to be service oriented in order to be less complex and more agile than they are today. Becoming service oriented, however, is a process that takes time because all systems must adhere to the above mentioned SOA rules.

III. Deriving Criteria for Good SOA Projects

A. The Adoption Scenario

As said already the authors believe that the top down SOA introduction approach where top management commits to SOA and a major effort is made, is far from reality in most companies. The realistic scenario in our opinion is an iterative development in small steps. This means that isolated projects are conducted which follow a subset of the SOA principles. After a few projects have implemented services it might be recognized that SOA makes project management more efficient and that one should consider coordinated action to broadly enforce the adoption of SOA methods. In such a setting project management is likely to evaluate the concepts available and use those that provide a superior cost benefit ratio. If SOA is to be chosen it must be better than other approaches or yield benefits that no other approach can provide.

SOA as frequently communicated can hardly be called a light weight approach since for example governance structure and IT infrastructure are needed to exploit all potential benefits of the concept. In the following paragraph we will analyze costs and benefits of SOA, compare them to project goals and constraints and identify those aspects of SOA that yield immediate benefits. The intention behind this approach is to establish SOA as a tool useful for daily business. Once this is achieved further steps towards the SOA vision might become feasible and an investment in infrastructure becomes an option.

B. SOA Benefits vs. Project Goals

As explained In part II-B the benefits of adopting SOA are more agility, complexity reduction, increased reusability and better interoperability. These benefits of adopting SOA practices are furthermore of mostly strategic and long term nature. In contrast, projects commonly have a duration of months to years and need to achieve very strictly defined goals within a set time frame. With this work we try to find a compromise to ease the tension between theses conflicting targets and benefits.

It might not be immediately clear why the SOA benefits are in opposition to project goals. Wouldn't agility be beneficial to achieve a timely project conclusion? The problem is that SOA benefits can only be achieved through adhering to certain design principles which possibly make a project longer and more expensive without yielding immediate benefits. In other words we have an incentive problem. Project management today must do certain things that project managers in the future will benefit from. In the following we will derive a set of criteria that help to identify projects and SOA practices that, in combination, promise superior performance. This paper is supposed to provide a set of guidelines to identify promising first step SOA projects and the SOA principles that should be applied. These projects provide immediate tangible benefits and partly implement SOA ideas.

In order to resolve the conflict of interest between project management's goals and long term SOA benefits one should first analyze which SOA benefits can be realized at least partly in the short run. Furthermore one should analyze which cost factors of SOA can be avoided in a first attempt.
Interface orientation, which is absolutely necessary for an SOA, is a consequence of implementation guidelines and will only cause minimal increases of programming effort. On the other hand even minor changes might punish too tightly interwoven parts of functionality. Therefore the benefits of interface orientation cost little, reduce the risk from changing requirements and simplify project management since different teams can work on the basis of defined interfaces which greatly reduces coordination effort. It follows that this design paradigm does not increase the cost or the duration of a project but has the potential to greatly reduce project duration and project risk in case of unexpected changes. The reduction of coordination effort further reduces project duration.
Standardization can probably be expected to be 'built in' in a project that implements a functionality from scratch. If, however, the extension of an existing function is the goal, project management should carefully analyze cost and benefits of a standardized way of communication between different modules. On the one hand standardization simplifies the task because only one standard needs to be understood but on the other hand it is possible that a lot of functions need to be wrapped in order to appear to be standardized elements. Besides wrapping elements that don't follow the set standard one could also let some sort of middleware perform the task of standard translation so that all modules can interact without having to implement different message types. Since both options cause costs because either wrappers need to be programmed or licenses need to be bought, the question of standardization has to be decided in the specific context. This is in opposition to the global perspective since standardized interfaces will certainly reduce integration effort in the future.
Autonomy and Modularity are necessary for an SOA because services as fundamental building blocks are supposed to provide certain functions to any requester. However, the reason for todays application landscapes' interdependencies is that often a quick and dirty close coupling solution is much faster

and cheaper than a properly executed separation of functions. Whether or not these paradigms provide benefits to a project therefore depends on the size and complexity of the project. Small projects are likely to be quicker with close coupling but large projects might profit from modularizing the tasks and thus making management easier.

Businesss Orientation: Correct granularity is an aspect which is of importance for the SOA idea because it is an important aspect of "business orientation". A service should be so large that it provides a meaningful business functionality. This requirement provides benefits on the small project scale because one service that provides a whole chunk of functionality is easier to handle than a multitude of small functions that need to be known on both sides of an interface. However, similar to the above said, bigger projects are likely to profit from this aspect while it adds overhead to small ones.

From what was said above one can identify projects for which SOA principles yield bigger benefits than for others. The bigger and more complex the project, the bigger the benefits from cutting it up into small chunks of functionality. The more time pressure on the project, the more important it is to start at different points and meet in the middle. It follows that the main drivers for potential SOA benefits in projects are complexity, size and time pressure because here SOA can help to cope with them. Comparing these to the project goals (time, scope, money) it appears to be feasible to see SOA as a means of achieving the project goals because it helps to control some of the main factors responsible for projects not meeting their targets.

In addition to the above mentioned inherent characteristics of promising SOA projects which we deduced from SOA paradigms, there are several characteristics of the projects that would profit from SOA ideas, which can be deduced from possible SOA benefits.

SOA **reduces complexity** by cutting things up into separate domains. Within each domain Services might not be extremely advantageous but they probably are for inter domain communication. It follows that it might be beneficial to consider projects or problems at the boundary of a domain. For example, trying to use SOA or simply services within SAP R/3 is a rather challenging task. Even though this might possibly be beneficial it is more feasible to start with a project that for example implements a functionality that gets data from R/3 to another application. At the domain boundary where two systems need to be connected it is much easier to find SOA benefits. Complexity reduction is especially important in very complex environments. Therefore it is important to find projects that work on a sufficiently complicated problem. Writing a plain html website is not very complex. Creating a website that provides information from different sources and presents it all on one screen is considerably more interesting and it might be considerably easier to provide the content with the use of services.

The most mentioned SOA benefit is **agility**. By using a service oriented approach to design ones IT landscape, a firm is supposed to be able to adapt to changes faster. Since this only plays a role in environments where changes actually do happen, one should pick an environment that is not static. Human resource operations for example are, at least at the core, very static. Managing personnel data has not changed fundamentally during the last years. Logistics operation on the other hand are constantly changing. As new possibilities for communication are developed the need for new channels of interaction arises. Furthermore, logistics have seen major changes in data availability which created the need for integrating more data sources into a more tightly woven net of data sinks. Along with the availability of data came the need to use it and implement more sophisticated instruments to control the flow of goods. These changes are still going on as innovations like RFID chips become more widely used. To sum up, it is most promising to look for SOA benefits in changing environments.

Interoperability is also mentioned in the context of SOA. Consequently one should look for situation where connectivity between different systems is a topic. For example, where two SAP systems need to be connected, a short script in proprietary format will almost always do the job. This way of interconnecting components will most definitely become expensive during the next release change, but since project managers are, as stated above, rather short sighted, this argument won't help. However, if two systems from different vendors need to be connected to each other chances are that middleware is required. Here it is feasible to argue for adherence to some standards from the SOA portfolio to reduce programming effort.The most important argument in favor of SOA is that one only needs people who understand the respective system. One doesn't need somebody who understands both sides. Furthermore the two sides don't need a lot of interaction because once the interface is agreed upon, they can do their job without caring for what the other side is doing. This is an important point because coordination of efforts costs a lot of time and specialists who know two systems are more expensive than those who only know one.

Reusability is the last SOA benefit we want to discuss here. SOA is supposed to provide services that can be used in more than one context. For services like "get_customer_master_data" this is immediately clear, but reuse is a feature that will mostly be considered a long term benefit. In the short run, however, reuse can also be interesting. For a project that has to implement different versions of similar tasks service orientation might provide substantial benefits. Consider for example a multi channel scenario. The same data has to be presented to users on different devices, or data has to be sent from different devices to on sink. Instead of programming all the different pipes to and from each device, the application could provide a single service that bundles the necessary functionality. The different devices can then access this service. This solution should be much more efficient than the "do it n times" approach.

Tables I and II summarize the above criteria for "good" SOA projects. Table I focuses on the SOA principles and lists

TABLE I
APPLYING SOA PRINCIPLES IS ESPECIALLY BENEFICIAL IF THE FOLLOWING CRITERIA ARE MET.

SOA Concept	Helps With
Interface Orientation	Size Complexity Risk of Changing Requirements Multiple Teams Time Constraints
Standardization	Building from Scratch
Autonomy/Modularity	Size Complexity Risk of Changing Requirements Multiple Teams Time Constraints
Business Orientation	Size

TABLE II
POTENTIAL SOA BENEFITS ARE MOST LIKELY TO BE REALIZED IF THESE CRITERIA ARE MET.

SOA Benefit	Helps With
Complexity Reduction	Complexity Risk of Changing Requirements Multiple Teams
Agility	Changing Environment
Interoperability	Domain Boundary Different Systems involved Scarce Expert Knowledge
Reusability	Shared Services Multi Channel Scenarios

project characteristics for which they are beneficial. Table II on the other hand focuses on the benefits of SOA and lists criteria where these benefits are most likely to have a profound impact. The redundancy of the criteria is removed in the tables in the following chapter that show the application of our ideas.

C. Explanation and Discussion of the Criteria

The criteria in the tables as such are not very telling. Therefore we now give a more detailed explanation of their intended meaning:

Size relates to the size of a project. Typical metrics would be number of people involved, duration or budget. A big project requires more coordination which makes well defined interfaces as single point of contact between different implementations beneficial.

Complexity relates to the degree of difficulty of a project. Metrics might be the number of function points, a subjective estimation of complexity on an arbitrary scale or the newness of the problem to be solved.

Risk of Changing Requirements: This is a constant threat to project management and needs to be subjectively estimated. When requirements change the functionalities of a system must be adjusted which is easier when interfaces are in place.

Multiple teams make coordination more difficult which is why interfaces are necessary to reduce coordination effort.

Time constraints make it necessary to start a project at different points. With interfaces agreed upon the implementation can work form different sides and meet in the middle rather than work sequentially.

Building from scratch requires definition of interaction protocols. One might as well agree on one rather than let different standards evolve. Different protocols mostly appear in systems that have evolved over time.

Domain Boundary: Here one often needs to provide interoperability between different systems and one might as well do it in a proper way. Within one homogeneous domain interfaces add too much overhead but they provide valuable interoperability between domains.

Different Systems require interfaces and therefore doing it properly doesn't cost a lot.

Scarce expert knowledge is a problem when one needs people who understand two different systems. Experts for one system alone are easier to find and if interoperability is guaranteed via interfaces the experts spend less time integrating their solutions and more time on actually implementing their parts.

Shared Services: Reuse is especially likely if one service is determined to be available to different users.

Multi Channel Scenarios are a special case of shared services. The different channels use the same data and just provide it differently to the user. One service providing the data in a generic format can then be reused in every channel and only needs some adaption to the needs of the specific presentation device.

The Criteria as stated above are rather soft and one might want them to be more strictly specified in order to allow a consistent application of the criteria in several different projects. Complexity for example could be measured in function points [9]. This method allows a project manager to estimate the time needed to implement a project depending on its complexity which is measured by different factors like programming effort or number of people involved. In addition to defining metrics for measuring the criteria one would also need to define when exactly (from which number of function points for example) a project starts being complex or when a project has a high risk of changing requirements and when not. However, how important is a consistent application of the criteria above? We are not trying to establish a method to compare different projects in a consistent way. Our goal is to provide a tool that supports the decision for or against the adoption of SOA ideas in the specific context of a project. In the real world such a decision will always depend on subjective judgment rather than on highly sophisticated measuring frameworks. Therefore we consider the derivation of the criteria as the most important aspect of the given problem and the specification of strict metrics a second order goal which might be addressed in further research.

IV. Application

Following [10], part of design science is the evaluation of ones research effort regarding utility, quality and efficacy. In an attempt to live up to those standards we will now demonstrate the application of our criteria to five projects from the chemicals industry. The projects were analyzed as part of an attempt to evaluate the benefits of SOA in the chemicals industry. The selection process was aimed at finding projects which would benefit from SOA. Therefore one might assume that the project selection is somewhat biased towards SOA. This is true with respect to the fact that the projects would all benefit from SOA. However, their suitability as first time SOA projects was not a selection criterion. However, the chosen selection should be considered as an illustration of the above criteria and their intended use and not as an attempt to empirically validate this work. To show the application of the criteria we will firstly give short descriptions of the projects we analyzed and then apply the criteria. The first column of the tables indicates whether the criterion gives and indication in favor or against SOA. The bottom row simply adds up the that column. This adding up should not be over emphasized and we will discuss the question of numerical evaluation of the criteria further down.

A. The Case of Achem

At Achem we evaluated a project that analyzed the possible extension of an existing web portal. The portal is used to provide manufacturers with information and let them capture information that they need to give to Achem. In order to provide this functionality the web portal is integrated with the backend SAP Systems. Data is taken from special tables and displayed in the front-end. If data needs to be stored it is written to certain tables using SAP proprietary functions. For the users of the web portal there are several ways to retrieve and store data. The standard way is manually reading and capturing information. For larger volumes it is also possible to transmit files in predefined formats. Both approaches work in either direction. With the goal to further automate the process of retrieving and capturing data a connection to the ERP systems of the partners is being considered. However, this task is not simple because on the side of the partners the systems are very heterogeneous. Everything from spreadsheets to sophisticated ERP systems can be expected. Furthermore the possible partners change frequently.

B. The Case of Bchem

At Bchem a complete redesign of the logistics systems was attempted in a project because the existing solution was terminal based and could not be adapted to serve the changing needs. The core of the system is a modified off the shelf logistics tool in combination with a large data base that is used to consolidate all relevant information. This information is drawn from ERP systems of the several business units. Upon consolidation the data is used for transport organization which involves several tasks like scheduling, giving a transport order to a carrier or managing in plant traffic. These tasks involve several rounds of bidirectional cross company communication.

C. The Case of Cchem

At Cchem we analyzed a project that had been started to integrate a manufacturing system with the ERP. In the chemicals industry in general there is little interaction between ERP systems and the productive systems and there are few standard solutions to improve this situation. The specific problem at Cchem was that production quantities were manually entered into the ERP system after production. This is a constant source of errors, is expensive because a trained worker has to enter the information and most importantly causes delays because other processes can only be started once the production data has been entered. In order to improve this Cchem had decided to use a program that took data from the production system and triggered the necessary ERP actions. After rollout in one plant, the solution will be adapted for use at up to twelve different sites.

D. The Case of Dchem

The question Dchem wanted to answer was whether or not SOA could help IT to perform better in the early phase of post merger integration. During a merger time is a crucial factor for success. Therefore Dchem's objective was to prepare the IT in order to be able to react very fast in case a merger happened. In the post merger phase one can differentiate between three types of data that must be integrated into the buying firms systems. The highest priority is given to core financial data on a very high level. This information must be available within weeks after the merger. Secondly Dchem needs more fine grained information on the success of each business units which is why data on a per customer per article base must be available to management within two months. Then in a third wave all other systems of lower priority are integrated with Dchem' systems.

E. The Case of Echem

Echem uses a rather old self developed application to track changes in its SAP System. The application is used to document the whole workflow from first change request to the final transportation of the changes from the test system to the productive system. The application has evolved for several years and is perceived to be rather hard to modify. Changing business rules for example need to be hard coded which is expensive and tedious since business rules are subject to frequent changes. The present project has to construct an application that fulfills the need for flexibility and ease of use.

F. Summary and Discussion of the Examples

From the short descriptions of the projects alone it is probably rather hard to tell, whether or not the evaluation in the tables really gives an indication of the right decision. Therefore we now discuss whether or not our personal judgment corresponds with the results in the tables.

Achem received six points in the evaluation and in fact we believe that SOA with web services is the only feasible way to

TABLE III
EVALUATING THE ACHEM PROJECT

+	Building from scratch	The connectors to the backend systems are already in place but the connectors to the partners are completely new
+	Size	Development teams of Achem and its partners are involved
-	Time	Not relevant
+	Domain boundary	Cross firm Interaction
+	Complexity	Heterogeneity of partner systems
-	Risk of changing requirements	Stable data requirements
+	Multiple teams	Implementations at both ends are necessary
-	Changing environment	The present system has been running for several years without changes
+	Different systems involved	Heterogeneous partner systems
+	Scarce expert knowledge	Experts for either system are in house but don't know the other systems respectively
+	Shared services	Once implemented a service could be used by all partners
+	Multi channel scenarios	Visual data retrieval stays necessary but automation is desired
+6		

TABLE IV
EVALUATING THE BCHEM PROJECT

-	Building from scratch	Backend and partner systems stay but the logistics system is built from scratch
+	Size	A multi million Euro project with several departments involved
+/-	Time	Not critical
+	Domain boundary	Cross firm interaction
+	Complexity	Heterogeneity of partner and backend systems
+	Risk of changing requirements	Logistics are a dynamic field at the moment
+	Multiple teams	Implementations at both ends are necessary
+	Changing environment	Logistics are a dynamic field at the moment
+	Different systems involved	Heterogeneous partner and backend systems
+	Scarce expert knowledge	Experts for each system are in house but don't always know the other systems respectively
+	Shared services	Once implemented a service could be used by all partners
+	Multi channel scenarios	Mobile devices are envisioned to be integrated
+9		

TABLE V
EVALUATING THE CCHEM PROJECT

-	Building from scratch	All systems exist already
-	Size	A small project with few people involved
-	Time	Of little relevance
+	Domain boundary	Cross systems interaction
-	Complexity	Few involved systems
-	Risk of changing requirements	No changes likely during implementation
+	Multiple teams	Implementations at both ends are necessary
-	Changing environment	All involved systems stay in place for decades
+	Different systems involved	Communication between different systems
+	Scarce expert knowledge	Experts for either system are in house but don't know the other system very well
+	Shared services	Once implemented a service could be used several times
-	Multi channel scenarios	Not relevant
-2		

automate the portal. The reason for this is that a 1:1 connection between Achem and the different partners is too expensive and therefore the portal was put in place as a platform for manual interaction.

Bchem currently conducts a very big project that would probably benefit considerably from SOA. Interestingly the project uses many aspects of SOA without calling it SOA. The communication among the different elements of the transportation management solution for example all communicate message based via one middleware.

Cchem is a very good example why SOA is an idea that is worth consideration. The whole project is only necessary because the production system is shipped with the wrong SAP connectors. Proper interfaces would have saved a lot of money in this context. However, the project itself is not very well suited for a proof of concept because it would require a lot of preconditions like the said interfaces in order to profit from SOA.

Dchem is a debatable case because on the one hand SOA offers notable benefits like the possibility for reuse of different components of an integration solution, on the other hand "quick and dirty" should be much faster. Therefore Dchem

TABLE VI
EVALUATING THE DCHEM PROJECT

+	Building from scratch	All source systems exist already but the interconnection part is new
-	Size	Rather small project with few involved members
+	Time	The crucial factor
+	Domain boundary	Cross company interaction
-	Complexity	Few systems with well defined requirements
-	Risk of changing requirements	No changes likely during implementation
+	Multiple teams	Implementations at both ends are necessary
-	Changing environment	Stable during implementation
+	Different systems involved	Communication between different systems
+	Scarce expert knowledge	There can be no experts who know the systems at Dchem and at a possible take over candidate
+	Shared services	Services could be reused in many mergers
-	Multi channel scenarios	Not relevant
+2		

TABLE VII
EVALUATING THE ECHEM PROJECT

+/-	Building from scratch	System is completely new but the target systems stay in place.
-	Size	Rather small project with few involved members
-	Time	Not important
+	Domain boundary	Interaction between the new application and SAP
-	Complexity	Few systems with well defined requirements
-	Risk of changing requirements	Rather small
-	Multiple teams	One team does the implementation
-	Changing environment	Target system will stay
+	Different systems involved	Communication between different systems
+	Scarce expert knowledge	Experts for application development and for SAP transports are required
+	Shared services	Once implemented the program can possibly used to track changes in other systems as well if the architecture is flexible enough
-	Multi channel scenarios	Not relevant
-3		

would surely not be a prime example for a proof of concepts SOA project because it would not benefit a lot in the short run.

Echem is a case that is similar to Dchem. The project itself would probably not benefit from SOA in the short run. Therefore this project would not be very well suited for a first SOA implementation.

Even though all the projects would benefit from SOA in one way or the other, only two of them promise to be successful as a proof of concepts for SOA. Achem would profit becuase SOA takes away complexity and Bchem would profit because SOA helps to cope with the complexity. Cchem shows very clearly how ignoring the SOA paradigms can become expensive in the long run while Dchem and Echem would benefit little from SOA in the short run. Therefore we conclude that the criteria listed in the tables serve well as indicators for suitable first SOA projects. They do not allow a precise measurement but they give an indication which is probably all that can be achieved without an overly sophisticated approach.

V. CONCLUSIONS AND FURTHER RESEARCH

SOA in the authors' opinion is an interesting approach to designing and managing large system landscapes. When ignoring the marketing hype, well established ideas like modularity and loose coupling build the foundation of the concept which is then enriched with business oriented concepts that make SOA more tangible and facilitate transfer to real systems. Since we consider SOA more than a hype the question how this idea can be brought into the companies is an important one because even good ideas need good marketing to gain acceptance. One facet of this is to use it where it is most likely to provide immediate benefits. This is exactly what this work is trying to support by providing criteria for the selection process.

Possible extensions of the presented framework for the selection of SOA projects would be a weighing of the presented criteria. Complexity and project size for example are stronger drivers towards SOA than the number of teams involved in the development process. Therefore one could weigh the different factors and get a more differentiated picture. However, this step doesn't really provide any further insight form the authors' point of view. If a project is evaluated using the given criteria one will intuitively weigh them and the presented framework will serve as a checklist. On the other hand, if projects were evaluated solely based on the number in the last row, weighing the factors would be worth further consideration.

Extending the idea of weighing the factors one could use more formal tools like cost benefit analysis [11] or its extension by [12]. Practitioners apparently consider the AHP method [13] especially promising. Less so because it forces a ranking through pairwise comparisons but because it appears to be more sophisticated. This work didn't apply any of those

approaches because the focus was on criteria for SOA and not on the application of such a method.

In the presented way our criteria are little more than a checklist.An rather important next step would be the definition of proper metrics to measure the different criteria in a more transparent and reproducible manner. This would also enhance the applicability of the framework because comparisons across projects evaluated by different people would become feasible.

The paper used theoretical sources to deduce criteria for projects that promise to benefit from SOA in the short run. These criteria were then evaluated by applying them to five projects the authors conducted at major German chemicals firms. The result of this work is a set of criteria that should serve as a guideline for the selection of early SOA projects. Having such criteria is important because they make it more likely that successful projects can be conducted which in turn provides momentum for the move towards SOA.

REFERENCES

[1] C. Legner and R. Heutschi, "Soa adoption in practice - findings from early soa implementations," in *Proceedings of the 15th European Conference on Information Systems (ECIS 2007), St. Gallen*, 2007.

[2] J. Siedersleben, "SOA revisited: Komponentenorientierung bei Systemlandschaften," *Wirtschaftsinformatik*, vol. 49, pp. 110–117, 2007.

[3] H. Kerzner, *Project Management: A Systems Approach to Planning, Scheduling, and Controlling*. John Wiley and Sons, 2003.

[4] R. Atkinson, "Project management: cost, time and quality, two best guesses and a phenomenon, its time to accept other success criteria," *International Journal of Project Management*, vol. 17, no. 6, pp. 337–342, 1999.

[5] A. Shenhar, O. Levy, and D. Dvir, "Mapping the dimensions of project success," *Project Management Journal*, vol. 28, no. 2, pp. 5–13, 1997.

[6] M. Cook, "Building enterprise information architectures: reengineering information systems," 1996.

[7] D. Krafzig, K. Banke, and D. Slama, *Enterprise SOA: Service-oriented Architecture Best Practices*. Prentice Hall Ptr, 2004.

[8] R. Heutschi, "Serviceorientierte Architektur, Architekturmodell und Umsetzung in der Praxis," Master's thesis, IWI-HSG, 2006.

[9] G. Low and D. Jeffery, "Function points in the estimation and evaluation of the softwareprocess," *Software Engineering, IEEE Transactions on*, vol. 16, no. 1, pp. 64–71, 1990.

[10] A. Hevner, S. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[11] C. Zangemeister, *Nutzwertanalyse in der Systemtechnik*. Wittemann, 1970.

[12] H. Grob, "Investitionsrechnung für Informations-und Kommunikationssysteme auf der Grundlage von Preis-Leistungs-Modellen," *Integration und Flexibilität: Eine Herausforderung für die Allgemeine Betriebswirtschaftslehre, Hrsg.: D. Adam ua*, pp. 335–352, 1989.

[13] T. Saaty, *What is the analytic hierarchy process?* Springer-Verlag New York, Inc. New York, NY, USA, 1988.

[14] G. Alonso, F. Casati, H. Kuno, and V. Machiraju, "Web Services: Concepts, Architectures and Applications," 2004.

[15] D. Dvir, T. Raz, and A. Shenhar, "An empirical analysis of the relationship between project planning and project success," *International Journal of Project Management*, vol. 21, no. 2, pp. 89–95, 2003.

[16] T. Erl, *Service Oriented Architecture: Concepts*. Technology, and Design (Upper Saddle River, NJ: Prentice Hall, 2005).

[17] M. Freeman and P. Beale, "Measuring project success," *Project Management Journal*, vol. 23, no. 1, pp. 8–17, 1992.

[18] R. Heutschi, C. Legner, B. Nr, B. HSG, C. BN, A. Back, W. Brenner, H. Österle, and R. Winter, "Serviceorientierte Architekturen: Vom Konzept zum Einsatz in der Praxis," *Integration, Informationslogistik und Architektur–Proceedings der DW2006, September 21-22, Friedrichshafen, Germany*, pp. 361–382.

[19] K. Nagel, *Nutzen der Informationsverarbeitung: Methoden zur Bewertung von strategischen Wettbewerbsvorteilen, Produktivitätsverbesserungen und Kosteneinsparungen*. R. Oldenbourg Verlag, 1990.

[20] E. Newcomer and G. Lomow, *Understanding SOA with Web Services (Independent Technology Guides)*. Addison-Wesley Professional, 2004.

[21] J. Schemm, R. Heutschi, T. Vogel, K. Wende, C. Legner, B. Nr, B. HSG, C. BN, A. Back, W. Brenner *et al.*, "Serviceorientierte Architekturen: Einordnung im Business Engineering," *Universit of St. Gallen*, 2006.

[22] D. Woods and T. Mattern, "Enterprise SOA: Designing IT for Business Innovation," 2006.

[23] R. Yin, *Case Study Research: design and methods*. Sage Publications Inc, 2003.

[24] R. Zarnekow, W. Brenner, and U. Pilgram, *Integriertes Informationsmanagement: Strategien und Lösungen Für das Management von IT-Dienstleistungen*. Springer, 2005.

Design Rules for User-Oriented IT Service Descriptions

Henrik Brocke, Thorsten Hau, Alexander Vogedes, Bernhard Schindlholzer,
Falk Uebernickel, Walter Brenner

University of St.Gallen

{henrik.brocke, thorsten.hau, alexander.vogedes, bernhard.schindlholzer,
falk.uebernickel, walter.brenner}@unisg.ch

Abstract

Customers of complex IT-services increasingly demand integrated value bundles that fit their individual needs. At the same time, IT service providers are facing commoditization of their products and need to standardize their portfolios to realize economies of scale. While approaches to coping with the gap between individual customer demand and the economic necessity of standardization have a long standing tradition in mature manufacturing industries, IT-service providers still struggle with translating their standardized portfolio into a form that is understandable and relevant for their customers. This paper proposes a way of describing IT-services that follows the paradigm of a "service dominant logic". We therefore transfer "service dominant logic" to the realm of IT and propose guidelines to create customer oriented service descriptions. An excerpt of a prototype description serves as an example, how the technical, inside view on IT-services can be translated into a customer-oriented outside view.

1. Introduction

This paper focuses on business-to-business IT-services such as hosting an application or providing desktop services. Providers of that sort of service (IT-service providers henceforth) are increasingly facing challenges of cost pressure and higher quality requirements. This development is largely due to increasing standardization and commoditization of information technology [1, 2]. While competition is thus driving down prices for IT services, customers question the value of IT-services and wonder whether they get their money's worth.

The combination of commoditization of their services and growing customer demands poses a significant threat for the business models of companies offering IT-services [1, 3, 4]. To cope with the cost pressure IT-service providers are facing, they have started to standardize their services. From producing specialized offerings for individual customers, IT-providers have moved to producing standard components at low cost and bundling them into customized service bundles. This approach promises economies of scale in delivering services to customers [5].

While this shift from creating highly individual solutions towards standardized offerings, which can be assembled into service bundles, has been taken up quickly in the industry (customizable standard software such as SAP or blade servers are good examples), a second related evolution has not yet taken place. When markets are shifting from seller markets to highly competitive buyer markets, as is happening in the IT-service market, companies have to focus on the value propositions they are offering to win and retain customers [4, 6, 7]. Other industries that have been confronted with commoditization have learnt to deal with this situation by moving away from a purely product centered perspective towards a service centered perspective, where the tangible product is only one component in a whole bundle. Even manufacturers of such tangible goods as automobiles have turned towards offering "utility" instead of products and today do not simply sell 'cars' but 'mobility' (take GM with services like "GMAC" and "OnStar" for example). The creation of integrated value bundles, that include maintenance, insurance and mobility services together with the base product automobile, is a manifestation of this shift from simple products towards value propositions.

Until today, many IT-service providers have not focused on making value propositions to their customers, but still focus on offering disintegrated commodity products and services. This gives rise to

978-0-7695-3450-3/09 $25.00 © 2009 IEEE

two considerations: Firstly, IT-service providers must focus on making value propositions (in the sense of [8, 9]) to address the customers' needs. Secondly, the value propositions have to be communicated to the customer in an appropriate way, as we will detail below.

Customers therefore should be offered integrated solutions that solve their problems instead of separate products and services, which they have to integrate themselves into usable services. In this article the word service is used in the sense of [8]. Thus a computer screen proposes a value by providing the service of displaying information. Only by looking at it, the customer can realize that value and use it in its process of making value propositions to other entities. Throughout we consider this proposed value to be the relevant aspect of all instances of IT-services. The most relevant distinguishing factor in this work is to whom value is proposed. The IT-service provider is an integrator of different services and bundles them according to the customer's needs. Thus the challenge for the provider is to propose the right service- and value- bundles to the customer to support his/her business process. In the screen example above the right service would be "displaying business data" and not the mere potential to display anything. Thus, one key aspect of the discussion below would be how to identify the customer need and then describe the corresponding integrated service solution in a customer oriented way.

A key aspect for the definition of customer focused services is their description. Nowadays, service descriptions of IT-services exhibit a resource-oriented, inside perspective, since they typically contain an enumeration of all input factors used to create a service and list descriptions and quality parameters of each input factor. Customer-oriented services on the other hand need to focus on the value proposition for the customer. A separation between the inside view and the outside view is necessary. This work does not solve the problem of translating a set of resources into marketable products, since we believe this to be a genuine engineering task. Much rather this work focuses on proposing design rules for developing customer-oriented descriptions.

The remainder of the paper is structured as follows. First we analyze current service descriptions with a focus on SLAs in order to support the thesis, that IT service provider have not yet realized the difference between input resources and outcome value and that current service descriptions are not appropriate means of communicating value to customers. We also analyze some recent research in the realm of SLA design and service descriptions and explain the difference to our work. Then we lay the theoretical foundation for our work by explaining what "service dominant logic" is and by giving a short overview over the concept of "end-to-end" offerings. Following this part, we describe our research process and methodology that led to the rules we created. The main part is dedicated to describing our design rules for customer oriented service descriptions. We finish by drawing some conclusions from our work.

2. State of the Art in Describing Services

This section establishes, that today's service descriptions are not the best way to address customers of IT- services. As shown above, IT services are becoming commodities. Therefore, providers of such services need to focus on creating value propositions and on communicating them attractively to potential customers. Unfortunately this is not the case. Nowadays descriptions are merely a statement of the characteristics of the resources that are involved to provide the service, filled with technical terms and therefore difficult to understand [10]. Due to their binding character, today Service Level Agreements (SLAs) are one of the most crucial description components in business-to-business relationships. But just like the whole service description, their focus commonly is on defining service-level parameters of IT components or input factors and not on the actual end-to-end services that affect the customer, i.e. the outcomes. The definition of service-levels is mostly limited to single components that interact to deliver the IT services. Yet, from the perspective of the customer, the individual service-level agreements are irrelevant, because the service-level is defined through all components that are involved in the service delivery chain [11].

The screening of today's offerings within the IT industry shows that SLAs for different service building blocks are put together without translating them into customer relevant value propositions. Imagine a hosting service including the delivery of data over a network. Instead of defining an SLA for availability at the customer's site, two figures, one for network availability, one for server availability are commonly given. Techniques like quality function deployment [12], a long standing engineering method to map product functionality to customer needs to develop products/services, do not enhance this situation, because they solidify the separation of customer needs into separate solution building blocks. Filling the matrix and adding up all the components does not lead to a complete end-to-end service description. It merely gives the requirements list for the engineers who build the product or compose the service.

To illustrate these shortcomings, consider the following examples taken from real world SLAs.

1. Giving details on "internal space management" and "Database administration" is commonplace in today's SLAs. However, this information is not helpful to any prospective consumer of those service components. "We do all necessary maintenance and administrative work" would be a commitment that suffices to underline that the customer will not be bothered with technical issues of application management.
2. Consider the following original text from an SLA: "This service module includes provision, operation, rollout and updating of Lotus Domino databases and Lotus Domino mail databases." The user might wonder what Lotus Notes can be used for. He/she should rather be told that "the email program will be Lotus Notes (which has a different look and feel than Microsoft products)." Constraints like restricted database size should be mentioned and explained in a user-friendly way i.e. "Storage space is restricted to 100 MB which is equivalent to about ten thousand emails without attachments."
3. The SLA-phrase "Service transfer point is the server system concerned." declares that the provider does not care how the service is delivered to the customer. Instead, delivery of the service is a different service that is described elsewhere within the description. This connection must be transparent for the customer. Better yet it should not be necessary because the transfer point is the customer's screen.
4. "LAN availability 99.97%" is another very commonplace quality parameter. It is very important that the local network is available almost always, but how much is 0.03% downtime? Is that an hour per year or per week? And how does this affect the availability of the application I need to use? Can I use SAP and Word during that time?
5. "Response time for SAP data record read <4sec". This SLA does not tell the user that he will experience an average of 4 seconds of response time by the system but that the server will need that long to look up a record in the database. How much time lies between a click and the visible response on the screen is not specified.

These examples show that firstly that SLA-authors rarely consider whom they are addressing with what kind of information and secondly all information is inherently focused on single resources. Of the five SLAs we have analyzed, not a single one attempted to bridge the gap between the resource-oriented, inside perspective and outcome-oriented, outside perspective. The impression that SLAs are resource focused is further supported by sources such as [13] which try to give business relevant advice on how to construct reliable SLAs. The customer as the target audience of an SLA is never considered in those expositions. Other literature that is interested in describing services is not focused on the service consumer either. The literature on web services for example explores ways of describing such services that are targeted at programmers or even machines [14].

The exposition in this section has focused on the weaknesses that service descriptions exhibit as means of communicating with the customer. We acknowledge that certainly there must also be a technical, resource oriented description of each service delivered as an inside definition of the service portfolio building block. However, an outside perspective must be defined and inside and outside perspective must never be mixed. A crucial factor to achieve user-oriented service descriptions is the strict separation between the manufacturing or input factor perspective and the value delivery or customer perspective. In this work we specify how one aspect of the separation between inside and outside view of manufacturing IT services can be done. We show how the description for the customer should be constructed. For the inside perspective see for example [15].

Arguing for customer oriented SLAs is quite easy if one takes it as given, that businesses want to make their offering transparent to the customers. While better communication is one possible way of differentiation in the market, there are also more subtle advantages of using descriptions of value propositions instead of descriptions of input factors to communicate with the customer. Firstly, customers have a higher willingness to pay for a service than for a commodity input due to the transfer of risk from the customer to the provider. Secondly, an integrated service offering decouples input and output and thus price and cost. The seller can therefore optimize both sides independently with possibly positive impact on the bottom line.

3. Service Dominant Logic for IT service providers

'Service dominant logic' as a new paradigm was first put forward by [9]. They promote a shift of focus from value creation through the exchange of goods to value co-creation through interaction between organizations and customers. Several key principles that mark the difference between goods and service focused thinking were identified by the authors. Below, we transfer

these principles to the realm of IT-services. Some principles from the original paper focus rather on economics than on business issues and are therefore omitted.

The application of specialized skills and knowledge is the fundamental unit of exchange. Traditionally IT service providers have provided hard- and software that has been used within an organization. In a service dominant logic, the units of exchange are not hardware and software but skills and knowledge (related to information technology), that help customers to reach their goals. IT service providers have to focus on skills and knowledge and use hard- and software as means of delivering these services.

Goods are distribution mechanisms for service provisioning. Even though hard- and software are not at the center of attention anymore in a service dominant logic, they are still of some importance. While hard- and software are not the fundamental unit of value exchange, they are the distribution mechanism for services and are therefore essential for delivery from the service provider's perspective. From the customer's perspective, hard- and software are of little interest since users care only for the delivery of certain services.

The customer is always a co-producer. When shifting from the perspective of creating value through exchange of goods to a perspective where value is created by applying certain skills and knowledge provided through a service for the customer, the customer becomes a co-producer of value. With customers as co-producers, knowledge about the customer's processes becomes extremely important in order to provide services that can be applied within the customer's processes.

The enterprise can only make value propositions. With customers being co-producers, it is obvious that without the customers' interactions no value can be created. This also leads to the realization that organizations can only make value propositions. It depends on the customers to determine the value of the service and realize it in a process of co-creation.

A service centered view is customer oriented and relational. The shift from a goods dominant logic to a service dominant logic also affects the understanding of business relationships for IT service providers. Traditionally, goods-dominant logic is transaction-oriented while service-dominant logic requires customer- and relation orientation. This requires service providers to identify and define necessary processes, that facilitate this change in the business relationship (i.e. sophisticated controlling systems to implement pay-per-use models).

These principles form the basis for a mind-change within IT-service providers towards service orientation. Table 1 is based on [9] and shows more clearly the change in perspective when moving from a goods-dominant logic to a service-dominant logic.

The change in perspective and the shift to a new dominant logic are essential steps towards customer-oriented service descriptions. By focusing solely on services and how they are delivered to the customer, the focus shifts from single instances of hardware and software towards the total combination of hardware and software that is used to deliver a service. This also leads to a change in understanding of service delivery responsibility which we call the end-to-end perspective of the IT service delivery chain.

Table 1: Comparison of goods based versus service based IT providers.

	Traditional perspective of goods-centered IT service providers	**Emerging perspective of service-centered IT service providers**
Primary unit of exchange	Hardware Products and Software Licenses.	Knowledge and skills (possible embedded in hardware, software) are applied in the form of services.
Role of goods	Hardware and Software are developed, configured and installed.	Hardware and Software are used to deliver services but provide no value in themselves.
Role of customer / user	Customers are the receivers and users of hardware and software.	Customers are co-producers who create value by using the service in their processes.
Determination and meaning of value	Value is created through the exchange of hardware, software and project deliverables.	Value is co-created in a co-production process between IT-service provider and customer. Value can only be determined by customers and realized in their process.
Firm-customer interaction	Customers are acted upon. Hardware and Software is delivered and support is given to the customer.	The services are continuously provided to enable the customer's processes. Customers are operant; they participate in delivery of the service through its use.

4. End-to-End View on IT Services

End-to-end design is the design principle that has its roots in the design of distributed computer systems [16]. The principle suggests that the definition of requirements for a communication system requires an understanding of the end-points of this communication system. The requirements on the communication system for a tele-surgery application for example are different from those for a video-conferencing application. They require different control mechanisms at different levels of the communication or application system, respectively.

The end-to-end argument is an important part in the design of modern communication networks but it is not just limited to this context. The essential tasks when considering end-to-end design is the identification of the individual end-points. Applying the end-to-end principle to the IT service delivery chain, the end-points are the IT service provider on one side and the customer on the other side. With this perspective it becomes obvious that different components are involved when delivering IT services. All these components have to be included in the specification of service delivery parameters to ensure the delivery of value propositions according to the agreed upon service levels. Figure 1 depicts this end-to-end perspective of IT services which is not limited to the infrastructure at the IT service provider, but also has the different elements between IT service provider and end-user (i.e. WAN/LAN components, personal computer of the customer) in view.

The service levels for individual hard- or software components become irrelevant for customers, since from their perspective only the service level of the complete service is relevant. The essential change in perspective here is not only to recognize that there are different components involved in the service delivery chain, but rather that it is necessary to be in control of them to deliver a certain quality degree of value proposition to the customer. In that case, SLAs could still govern inter-provider relationships along the service value chain whilst the customer shall only be confronted with the end-to-end description.

These two concepts, the service-dominant logic and the end to end perspective on the IT-service delivery chain form the basis for the understanding of the changing role of IT-service providers and the need to adopt service descriptions accordingly. In the following sections we will outline a way of describing services that is more appropriate for communicating with the customer than nowadays service descriptions without forfeiting the detail and binding character of common SLAs.

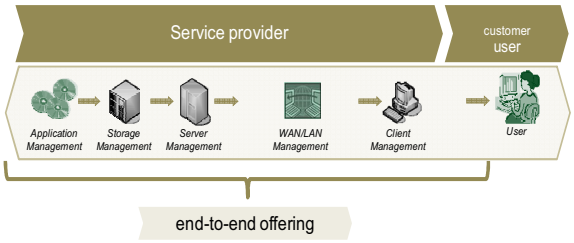


Figure 1: End-to-end offerings include the whole service creation process.

5. Research Process and Methodology

The research described in this paper follows "design research" guidelines as promoted (formerly as "design science") by [17] or [18]. In this section we are going to describe the overall research process as well as the design part of this research project which was carried out in cooperation with a large German IT-service provider (ITS henceforth).

One of the project's goals within the development of a prototype ERP System for a provider of IT-services was a suitable description of end-to-end services. As our analysis showed, current service descriptions with their SLAs are not able to provide this functionality due to their technical character, their use of complex terminology and their focus on input factors into IT-services rather than the business relevant effects.

The service to be offered within the prototype application was an end-to-end accounting service consisting of hosting, bandwidth, a desktop computer and all necessary supporting services as shown in figure 1. From our SLA analysis and multiple in-depth interviews with an IT Manager we had identified the need to create IT product offerings that focus on user needs and integrate several elements of the value chain. Assuming the ability to create the building blocks of such products as given (since they are being sold today), we identified our research question: Which rules should govern the design of user oriented descriptions of IT service bundles?

For theoretical grounding we reviewed literature on service level agreements [11], service orientation [19, 20] and service-dominant logic [9]. Additionally we analyzed the documentation of a major ERP system in order to complement our SLA analysis with a product manual that is explicitly focused on the user.

In order to understand the structural aspects of descriptions we built prototypes with a professional authoring tool for product descriptions which is being used to create manuals for complex machinery and analyzed the design specification of BMEcat [21], an

industry standard for structuring XML Product descriptions. We also considered the available web-service specifications [14, 22, 23].

The next step involved the first cycle of an iterative description-engineering process. We built a model description by creating a document that described the IT-Service "End-to-End Accounting" which involved all necessary elements from application hosting to a desktop computer. After several iterations of design and prototyping, we informally evaluated and improved that description in a workshop with six IT professionals. After numerous improvements and internal review sessions we created a questionnaire for a field test, this time with IT-professionals who had not been exposed to our service description before. We presented the description together with a questionnaire to four accountants and one product manager of major European companies. The feedback was very positive and especially the product manager pointed out several aspects of our description that he found to be superior to the average service description he received as a basis for his buying decision.

After minor changes due to the interview results we presented our prototype and the rules to two IT product managers. Together with these managers we then created a new product description that was to be used within an ITS project. We started with the prototype and extended it to fit in the project context. After we had completed that task we had a real world product description that contained a considerable amount of details and was conforming to the design rules we had established with our prototype. The timelines of the research activities are shown in Table 2.

Table 2: Interviews and Workshops conducted to improve and test our artifact.

June 2007	First draft of product description, 2 day workshop with one IT professional from ITS
August 2007	Workshop with 6 professionals from IT organizations of 3 large firms
September 2007	Interview with two professional accountants from ITS
September 2007	Interview with two professional accountants from an ITS customer
October 2007	Interview with a procurement manager of a SME
January 2008	Workshop with two ITS product managers of IT services

6. The Design Rules

In the following we present the conceptualization of the findings within our research project. We do this by giving a set of design rules that should govern the creation of customer-oriented SLAs and service descriptions. The principles laid out in Table 1 can be considered as the transfer agents that help to make the transition from a set of ordinary SLAs to the type of customer oriented service description we propose in this paper.

Throughout the following text we will use an example of an end-to-end service for commercial managers. We will suppose that one provider is responsible for all elements of the service delivery chain. This approach eliminates some complications like the coordination of a chain of interdependent providers. One way to picture this "all in one" IT service provider is to think of it as the service provider of a large company. It has to buy all input factors and then delivers the integrated service to its customers. The rules, that should govern the design of a customer oriented description, are the following:

Rule 1. Outcomes define content. This most important rule for the creation of user oriented service description states that the leading principle for describing a service offering is outcome orientation. How the IT-provider produces the value propositions e.g. the potential to work on a workstation, is of no importance from the users point of view. Not the provisioning process of hard- and software is in focus, but the value that can be co-created. For example, the phrase "The software lets you create accounts and manage them" is acceptable while the phrase "on our blade servers we host SAP R/3 for you" is not.

Rule 2. Outcomes define the structure. Different from input-oriented SLA structure, service descriptions should be structured along dimensions of outcomes. The reader does not seek information on the building blocks needed to deliver a service but needs to know in which of his/her processes the service can be used. One should not describe accounting software and a database as separate entities but focus on accounting and controlling as tasks that are supported. Consequently, the table of contents of a written service description will not contain the basic elements like "ERP software", "hosting" and "Internet connection", but much rather "accounting", "controlling" and "annual accounts".

Besides these two very general rules there are more specific rules that provide a guideline for the creation of good service descriptions.

Rule 3. Quality parameters are defined end-to-end. End-to-end orientation of service descriptions requires the definition of customer oriented quality parameters. The parameter service setup time includes all tasks that have to be performed before the user can actually use the service. The setup time for a single component is irrelevant. With all quality parameters being focused on the user, effects of errors have to be observed from the user's point of view. Data loss probability and backup interval are therefore translated to 'hours of lost work'. Response times are measured at the client computer and not for some resource along the delivery chain. Technical parameters that have an influence on the response time line LAN latency need not be communicated to the customer.

Rule 4. Changes to the service potential are services themselves. Similar to goods that are delivered, services need to be set up. To start, change or stop a service, additional services, which carry out the appropriate functions, are necessary. These services need descriptions just as the core services. Stopping a service consists of picking up your hardware, returning all stored data and deleting your personal information. All these service elements create value for the user. Picking up the desktop frees space and returning data enables the customer to abide by the law that expects you to store accounting data for several years**.**

Rule 5. The description is binding and declared in a user / process oriented language. Services are to be described in a customer-oriented way that helps the customer to understand how the service creates value for him/her. However, the statements within a service description must also have the binding character of commitments. Short precise statements about the process support are desirable with the employed language coming from the targeted realm of application. While an accountant knows what 'SAP-GUI' is and a computer artist is familiar to a 'tile cache', both do not necessarily understand the meaning of "OSI layer 7" within a service description.

Rule 6. Cooperation is a defined process. As value is created in a cooperative process, customer and provider have to define how they work together. Instances of this rule are the regulation of service provider access to resources like PCs or printers that are located on the customer's premises. Furthermore, it must be made clear to the customer which actions the service can perform on its own and for which tasks it can only provide support. Closing all accounts is an automatic process, while ensuring their correctness cannot be guaranteed by the service beyond the correct use of math.

Rule 7. Information requirements are defined. All data needed for order processing should be requested during the ordering process. Therefore, a service description has to specify all data that needs to be known by the supplier in order to be able to deliver the service. However, this data must be within the knowledge domain of the purchaser: Data like the IP-address, application version numbers or system specifications are certainly not. This rule very simply avoids going back to the customer and asking the same question several times. Imagine situations where first the user management asks the customer for the exact location of his desk within the building and then the hardware delivery department calls to get exactly the same information. This would certainly reduce the perceived value of the service since the co-creation process of setting up the service involves more customer actions.

Rule 8. Dependencies are explicit. As a consequence of defining services in line with business processes, the possibility to order a service is dependent on the existence of other services, as explained above. Therefore the description of a service has to specify the dependencies on other services. In contrast to the order process of goods it is not only relevant which services can be ordered but also which services are already in use at the customer's site. As an easy example consider a cancellation service that can only be ordered once per corresponding service and cannot be ordered before the corresponding service has been ordered. This is different in an input oriented SLA. Nobody keeps you from ordering software without having a PC even though you cannot derive any value from it.

Rule 9. Structure goes from general to detail. As explained above the service offered is end to end. This means the scope includes all elements that are necessary to deliver a certain service to the customer. This includes a mouse as well as a LAN, hosting etc. In order to be able to describe such a service one could, as current day SLAs do, add up all elements that contribute to the service. Then the description would start by describing the mouse, go on to describe the LAN and end with the hosting service. For customers a different approach going from general to detail is better suited. On the highest level of abstraction, "IT-assistance for commercial managers" is offered. This service includes sub services like "IT-assistance for accounting", "IT-assistance for controlling" and "IT-assistance for office tasks". On each level of abstraction the value proposition is clear. On the topmost level the IT-service supports commercial managers. If more specific information is needed, the next hierarchy level tells the reader that accounting is part of that service. Within accounting the reader then

gets some specifics about the accounting service. Its end-user oriented availability, the software deployed and the response times when using the user interface.

There are two notable aspects in the explanation above. Firstly, the description has a hierarchy that is as deep as desired by the writer. One can start at a very abstract level that puts the service description in a nutshell and get down to so much detail at the bottom of the description that one could describe every single transaction possible in an accounting system. Secondly, the information that is contained within the structure always pertains to the utility that the user can create by using the service. On the topmost level the value is 'getting work done'. At the most detailed level this utility could be 'generate printouts of transactions. On each level the description of the value creating service is accompanied by crucial information such as availability or quality parameters such as screen size or response time.

The hierarchical structure is a key element of making the description accessible to human readers. We tried different structures such as first describing all the functions before describing all the quality parameters or describing the service by adding up all elements in consecutive order. The accessibility of the information is severely reduced by such approaches.

Rule 10. Each service element is well defined. Describing all services in the same structure helps the reader to find information and ensures completeness of information. Therefore each service element should be described by the same categories of information:

- Overview: a short text describes as an abstract the main content of the service.
- Order Data: This includes the ordering number and name of the service, the minimum purchase quantity and the runtime of the service.
- Prices: They might consist of separate elements like setup charge and usage fee.
- Order information: As explained above all necessary data for processing an order has to be declared during the order process. Which data is needed is stated for each service element.
- Obligation to co-operate: An exact definition of the customer's tasks in value creation.
- Quality parameters: Are defined for every service element and always targeted at the end user.
- Dependencies: Specification which services have to be ordered together and which services have to be ordered, before another service.
- Additional information like target group of the offering, current technical solutions, information about resources and other trust building information.

Rule 11. Modularity is crucial. Having described each service in a user oriented manner with each service element containing all necessary declarations and being located in the right spot in the hierarchy, the description of the whole service portfolio is easily accessible to the reader. The structure can even be used to generate role-specific views of the description. Depending on the intended audience, specific elements of the descriptions can be omitted. User with low access rights for example can be excluded from seeing the price-elements of the descriptions while a manager would only see the overviews and price elements. For this to work it is necessary that all information is in the right place and marked correctly as specified under rule 10.

Figure 2 shows an example of an end-to-end product description. It is structured according to the presented rules and contains the appropriate information. The rules are implemented in the following way:

Rule 1 The outcome "functioning PC" is described rather than the process of performing tasks that are aimed at that outcome.

Rule 2 From the description it is clear, that not the delivered hardware is described but the functionalities that are available, such as a shared drive.

Rule 3 The quality parameter declares that within five days from ordering, one can use the computer and access all promise functionalities.

Rule 4 Not given in example.

Rule 5 The language is clear, concise and service levels such as "five days" are made explicit.

Rule 6 The customer's tasks are clearly stated.

Rule 7 The data the customer has to provide is clearly stated.

Rule 8 The dependencies are separately stated. A frame contract must be established, before this product can be ordered.

Rule 9 Not shown. Picture the whole description as a tree. Below the element shown here there are more detailed descriptions available. The shown description is supplemented with additional details like technical information.

Rule 10 The structure of table 2 is repeated in every element of the description.

Rule 11 Table 2 only shows a very high level aggregation of the service provided. More detail is given further down in the description hierarchy (not shown here).

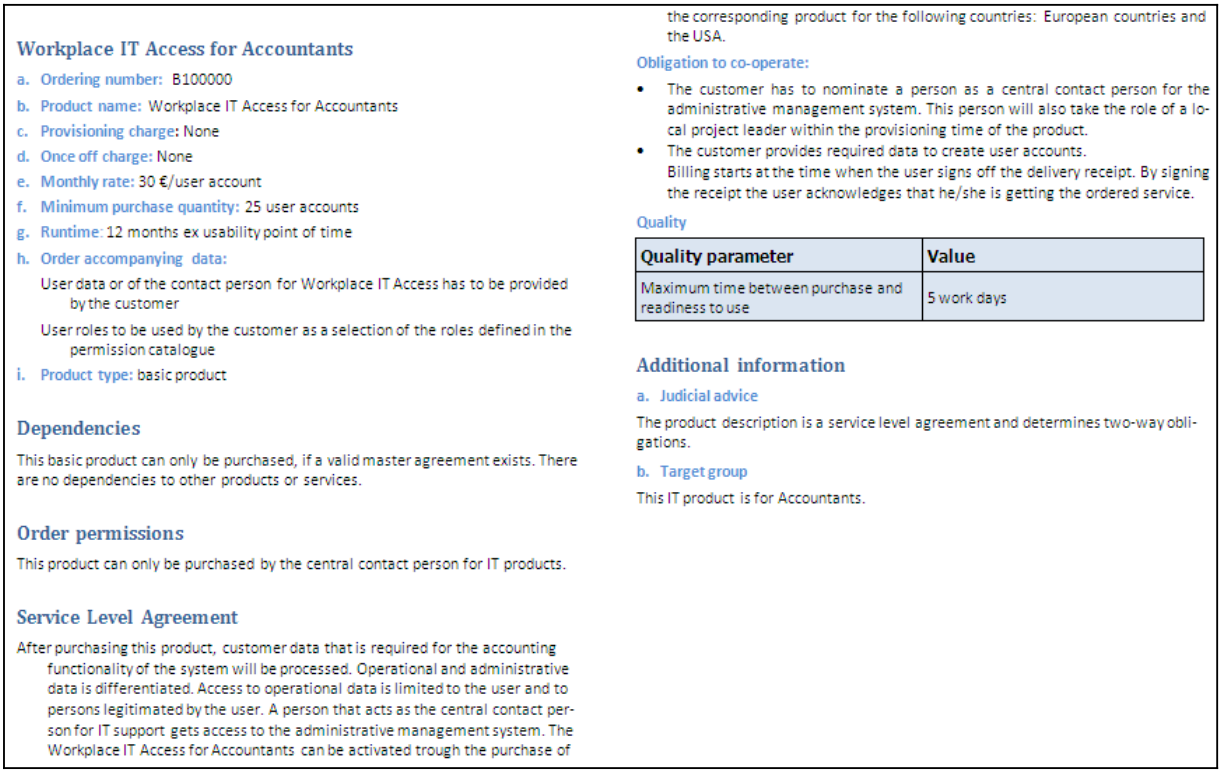
7. Conclusions

In the beginning of this paper we motivated our work by illustrating the lack of customer orientation in the IT industry. We showed this through the analysis of several SLAs that were resource focused instead of customer focused. Then we explained the sources we used for the construction of our prototype description. We drew from service science, from marketing, from the literature on service descriptions and SLAs as well as from works on industrialization. The description of the research process is quite extensive so as to show that theoretical as well as practical considerations have led to the artifact and the rules proposed above.

The results of our research process are design rules to describe customer focused IT services, that are tested through applying it in a description prototype. The most important achievement of this prototype is that due to the rules the whole document is focused on customer needs and not on technical inputs.

The advantage of such an approach lies in the communication to the customer. Firstly the customer of the service directly sees the value proposition of the service and does not have to compose different products and services into his/her personal value proposition. Thus we improve the value for the customer by lowering his transaction costs. At the same time, however, we do not forfeit the binding character of the service descriptions. All information necessary to judge the business relevant availability and quality of the service is declared. Furthermore this value-focused way to describe services offers an important advantage in a market for commodities. With only few possibilities to differentiate through features of input factors, the composition and communication of a service become more important. If a service provider succeeds in understanding the customer's needs and can create a spot on solution for him, the description will reflect this and thus offer an advantage over competitors.

Commoditization of IT services will force service providers to differentiate on other aspects than on price if they want to stay in the market. The proposed design rules to declare end-user oriented service descriptions concretize an unprecedented way to archive such an advantage.



Workplace IT Access for Accountants

a. Ordering number: B100000
b. Product name: Workplace IT Access for Accountants
c. Provisioning charge: None
d. Once off charge: None
e. Monthly rate: 30 €/user account
f. Minimum purchase quantity: 25 user accounts
g. Runtime: 12 months ex usability point of time
h. Order accompanying data:
User data or of the contact person for Workplace IT Access has to be provided by the customer
User roles to be used by the customer as a selection of the roles defined in the permission catalogue
i. Product type: basic product

Dependencies

This basic product can only be purchased, if a valid master agreement exists. There are no dependencies to other products or services.

Order permissions

This product can only be purchased by the central contact person for IT products.

Service Level Agreement

After purchasing this product, customer data that is required for the accounting functionality of the system will be processed. Operational and administrative data is differentiated. Access to operational data is limited to the user and to persons legitimated by the user. A person that acts as the central contact person for IT support gets access to the administrative management system. The Workplace IT Access for Accountants can be activated trough the purchase of the corresponding product for the following countries: European countries and the USA.

Obligation to co-operate:

- The customer has to nominate a person as a central contact person for the administrative management system. This person will also take the role of a local project leader within the provisioning time of the product.
- The customer provides required data to create user accounts.
Billing starts at the time when the user signs off the delivery receipt. By signing the receipt the user acknowledges that he/she is getting the ordered service.

Quality

Quality parameter	Value
Maximum time between purchase and readiness to use	5 work days

Additional information

a. Judicial advice

The product description is a service level agreement and determines two-way obligations.

b. Target group

This IT product is for Accountants.

Figure 2: Excerpt of the prototype description.

References

[1] Carr, N., *IT doesn't matter.* IEEE Engineering Management Review Online, 2004. 32(1): p. 24-24.

[2] Carr, N.G., *The End of Corporate Computing.* Mit Sloan Management Review, 2005. 46(3): p. 67-73.

[3] Böhmann, T., M. Junginger, and H. Krcmar. *Modular Service Architectures: A Concept and Method for Engineering IT Services*. in *Proceedings of the 36th Hawaii international Conference on System Sciences (HICCS'03)*. 2003.

[4] Zarnekow, R., W. Brenner, and U. Pilgram, *Integrated Information Management. Applying Successful Industrial Concepts in IT.* 1 ed. 2006, Berlin: Springer.

[5] Zarnekow, R., *Produktionsmanagement von IT-Dienstleistungen. Grundlagen, Aufgaben und Prozesse*. 2007, Berlin: Springer Verlag.

[6] Peppard, J., *Managing IT as a Portfolio of Services.* European Management Journal, 2003. 21(4).

[7] Trienekens, J.M., J.J. Bouman, and M. van der Zwan, *Specification of Service Level Agreements: Problems, Principles and Practices.* Software Quality Journal, 2004. 12(1).

[8] Vargo, S.L. and R.F. Lusch, *Service-dominant logic: continuing the evolution.* Journal of the Academy of Marketing Science, 2008. 36(1): p. 1-10.

[9] Vargo, S. and R. Lusch, *Evolving to a New Dominant Logic for Marketing.* Journal of Marketing, 2004. 68(1): p. 1-17.

[10] Sturm, R., W. Morris, and M. Jander, *Foundations of Service Level Management*. 2000: Sams.

[11] Pietsch, W., *Kundenorientierte Ausgestaltung von IT Service Level Agreements.* Software Process Improvement: 12th European Conference, EuroSPI 2005, Budapest, Hungary, November 9-11, 2005: Proceedings, 2005.

[12] Sullivan, L., *Quality Function Deployment.* 1986.

[13] Soebbing, T., *Handbuch IT-Outsourcing.* Rechtliche, strategische und steuerliche Fragen. Redline Wirtschaft bei Ueberreuter, Frankfurt Wien, 2002. 19.

[14] W3C, *Web Services Glossary.* www.w3.org/TR/ws-gloss/.

[15] Ebert, N.U., Falk; Hochstein, Axel; Brenner, Walter, *A Service Model for the Development of Management Systems for IT-enabled Services.* Proceedings of the Thirteenth Americas Conference on Information Systems (AMCIS 2007), 2007.

[16] Saltzer, J., D. Reed, and D. Clark, *End-To-End Arguments in System Design.* Technology, 1984. 100: p. 0661.

[17] Hevner, A., et al., *Design Science in Information Systems Research.* MIS Quarterly, 2004. 28(1): p. 75-105.

[18] Peffers, K., et al. *The Design Science Research Process: A Model for Producing and Presenting Information Systems Research*. in *Proceedings of the First International Conference on Design Science Research in Information Systems and Technology (DESRIST 2006)*. 2006.

[19] Maglio, P.P., et al., *Service systems, service scientists, SSME, and innovation.* Communications of the ACM, 2006. 49(7): p. 81-85.

[20] Papazoglou, M.P., *Service-oriented computing: concepts, characteristics and directions.* Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on, 2003: p. 3-12.

[21] Schmitz, V., O. Kelkar, and T. Pastoors, *Specification BMEcat, Version 1.2.* URL: http://www. bmecat. org, 2001.

[22] Alonso, G., *Web Services: Concepts, Architectures and Applications*. 2004: Springer.

[23] Casati, F., et al., *Business-oriented management of Web services.* Commun. ACM, 2003. 46(10): p. 55-60.

Economic Effects of Multi-Homing and Content Delivery Networks

Thorsten Hau*, Jochen Wulf†, Rüdiger Zarnekow† and Walter Brenner*
*University of St. Gallen, Switzerland
†Technical University of Berlin, Germany

***Abstract*—The structure of the Internet serves as a big "commoditizer" of all traffic. Therefore data transport is a commodity business. However, recent trends in the internet are changing this structure. The practices of multi-homing and using content delivery networks reduce the commodity nature of data being transported and put terminating internet service providers in a position to price discriminate against specific providers or types of traffic. We firstly formalize multi homing and content delivery networks, we then derive end user prices for paid content and lastly show consequences of the modeled situation. We thus show how the two technologies to bypass crowded peerings change the internet business model. Traffic which is sensitive to transport quality will be paying higher fees to terminating ISPs.**

I. Introduction

Internet pricing is a vast field of research that has its roots in telecommunications. Early work on pricing of voice communications established some of the corner stones of our thinking about how communications pricing works and which topics are relevant. This paper departs somewhat from the "classical" literature on communications pricing by considering a special pricing problem present in todays internet that has not been covered by the extant literature.

We focus on content oriented pricing of network services. In such a scenario internet service providers (ISPs) with access to end users (EUs) can discriminate against content providers and charge higher prices for termination. This scenario departs from the idealized bill and keep regime that is often used as a basis for analysis of network pricing decisions. However, our scenario is quite realistic. It is commonplace that content providers (CPs) directly buy transit from terminating ISPs, thus effectively paying them for preferential access to end users. This is a viable strategy because by bypassing crowded peerings used by the CP's original ISP, the CP gets higher quality access to EUs. This is due to the fact that peering points are a major bottleneck on the internet. While there are usually few capacity problems present inside a single providers network, peerings are often at their capacity limit [1], [2]. For the purpose of this paper we call this practice multi-homing (MH).

Secondly, content delivery networks (CDNs) are a popular way to enhance the flow of information on the web. A CDN uses local caches to keep distributed images of content close to EUs thus bypassing peerings. The CDN effectively buys transit from the ISP that terminates the traffic. We analyze these two settings providing a new perspective on the question of quality of service (QoS) on the internet.

The paper is structured as follows: First we explain the relevant entities of the current Internet that we need for a formal model. Then we present a formalized treatment of four scenarios that show how CDN and MH affect ISPs incentives to price traffic. Lastly we discuss consequences of our model and sketch out an agenda for further research.

II. Status Quo: The Market for Internet Connectivity

In this Section we present a schematic model of the internet and briefly discuss the different entities of the model.

A. Overview

Figures 1 and 2 show in an idealized manner the structure of the internet which are commonly used to model the internet [3], [4], [5]. In figure 2 at the bottom customers and content providers receive and send traffic, respectively. ISPs interconnect to exchange traffic to and from their customers (EUs or CPs). In figure 2 traffic could for example originate at the content provider and be passed on to its ISP 3. From there it is routed to ISP 1 to be terminated at end user (EU) one. The ISPs are fully meshed, each one interconnecting with all the others. It is a common approximation [4] that CPs (web sites) only send traffic and EUs only receive traffic. This approximation is justified by the real traffic patterns on the internet which show that downstream data transmission volume to EUs is much bigger than that upstream.

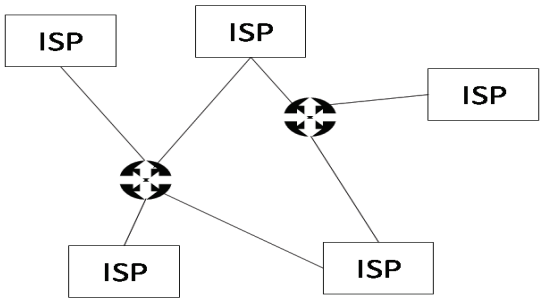


Fig. 1. Simplified model of the internet.

We now discuss the different model entities in detail:

B. Internet Service Providers

ISPs provide connectivity to internet users who are either end users or content providers. The ISPs pay the market price

p_w to other ISPs for termination of traffic. Typically ISPs have no lack of bandwidth on their backbones and can provide quality assurance to traffic either through excess capacity or network management techniques commonly employed only within a single providers network [6]. Bottlenecks[1] We ignore possible problems due to constrained access bandwidth and concentrate on the peering points. Note that all of the following also ignores dynamic considerations of integration and foreclosure through the ISP targeted at the CPs.

C. Points of Interconnection

In figure 2 the cicle with arrows in upper middle represents the point of interconnection where the three ISPs interconnect their networks. This is also called a "peering point". There are two dominant modes of interconnection: Peering and transit. Peering is a settlement free agreement to exchange traffic while transit involves payment for exchanged data typically on a bit-rate per time scale. Typically peering agreements are used between ISPs of similar size while transit is paid from small ISPs to larger ISPs.

Peering points with peering agreements are among the major bottlenecks of the internet. This is due to the fact that it always takes both parties to agree to extend the capacity of a peering point in order to increase its usable capacity. Since the telecommunications firms who own the internet infrastructure tend to be not overly cooperative among themselves these peering points represent excellent opportunities for power-play or generally uncooperative behavior. This results in many peerings being overloaded at peak times because one of the two parties involved has no interest in paying for extending its capacity [9], [10], [11], [7]. Ways for CPs to circumvent frequently overloaded (and therefore slow and packet dropping) peerings are multi-homing and the use of CDN services. Transit on the other hand involves a payment from one ISP to the other for the delivery of traffic. With such an agreement a guaranteed bandwidth is bought. Due to strategic considerations the biggest networks (so called Tier 1 networks) only peer among themselves and charge smaller networks for sending traffic to them. Since small ISPs have to pay for sending traffic to larger networks which is necessary to reach the whole internet they optimize their out-payments for transit fees by buying the least amount of bandwidth their users will tolerate. EUs receiving more and more traffic due to the internet becoming multimedia heavy and not paying for extra traffic due to flat-rate pricing adds to the problem of crowded peerings.

D. Content Providers

Content Providers are for example websites that buy connectivity to the Internet from an ISP (possibly mediated through a hosting provider). Content providers are able to multi-home which means they can buy connectivity for one fraction of their traffic from IPS1 and the rest from ISP 2. Furthermore they can host their service with any provider anywhere in the world giving them a very large set of ISPs to choose from. This creates competition amongst ISPs for CPs' business and therefore CPs face a market price for internet connectivity which is based on perfect competition. This price only includes un-prioritized traffic transportation across peering points.

Canonical analysis [4], [12], [13], [14] usually assumes the following model of Internet payments:

$$EU \rightarrow ISP_t \leftrightarrow_a ISP_o \leftarrow CP$$

(t=terminating, o=originating), ignoring where the CP gets funding from and emphasizing the analysis of the inter ISP settlement a.

Focusing on payments for content viewing, we model the payments flows according to the following scheme:

$$ISP_t \leftarrow_a ISP_o \leftarrow_{p_w} CP \leftarrow_p EU.$$

We ignore payments from the EU to the terminating ISP for access to the Internet. Payments from the EU to the CP might be paid through watching advertisements or direct money exchange. The arrow with the a stands for the access charge that is paid from one ISP to the other for sending traffic to the receiving ISP [15]. p is the final price paid by the EU for viewing some content. p_w is the price paid from the CP to the ISP for reaching the EU. If the ISP receiving p_w cannot terminate the traffic it has to pay an access charge to another ISP able to terminate the traffic. If a CDN is involved in content delivery, the CP has to pay the cost of the CDN, too. Modifying the way CPs get access to the EUs and modifying payments accordingly will be the key part of this paper. The two variations we will consider are: multi homing and CDN delivery. With MH, the terminating ISP is directly connected with the CP, while with CDNs a neutral third party mediates between CP and ISP_t. Under MH payment flows are

$$ISP_t \leftarrow_{p_w} CP \leftarrow_p EU$$

and the originating ISP is cut out of the equation. With CDN deliver the payments are:

$$ISP_t \leftarrow_{p_w} CDN \leftarrow_{p_w + c_{cdn}} CP \leftarrow_p EU.$$

We model CDNs as fully competitive entities only passing on costs.

E. End Users

Unlike CPs, EUs cannot divide their traffic amongst several ISPs and are immobile in the sense that they cannot chose their provider globally but need to chose among a small number of local ISPs. Therefore we consider a static scenario in which EUs are bound to their ISP, providing the ISP with a monopoly over terminating traffic to those EUs. Neither the competition among local ISPs for market share nor the dynamic considerations of consumers with switching costs [16] are considered here.

[1]Not meant in the sense of bottleneck facility as in [7], [8] may be present in the peerings points and in the access network.

F. Multi Homing

Multi homing is the common practice that CPs interconnect with several different ISPs. There are several instances of big content providers that are present in public peering points in order to gain direct access to ISPs with many EUs. Google for example is not just with one ISP in the USA but has its own transatlantic fiber capacity and is present in DE-CIX [17]. Besides the reliability aspect this practice allows the CPs to get direct access to an ISPs network and thus be more in control over bandwidth agreements with that ISP. This situation is apparently preferable to relying on one single ISP making transit and peering agreements with all other ISPs.

G. Content Delivery Networks

CDNs consist of a network of servers that are distributed around the internet within many ISPs infrastructures. A CDN takes content from a CP and caches it on those distributed servers which has two effects: Firstly content is brought closer to the EU without passing through inter ISP peerings thus making its delivery faster. Secondly the CDN buys transit from the ISP where it needs to terminate traffic and thus gets a guaranteed bandwidth it can extend as needed. The CDN then delivers the content it is caching from the mirror site to the EU. By using the services of a CDN a CP does not need to multi-home with every possible network. The CDN does this for its customers. We assume that CDNs do not make any profit and charge efficient prices. Multi homing and CDN are technologies to bypass crowded inter ISP peerings and thus achieve a higher quality of service as perceived by the EU loading a website.

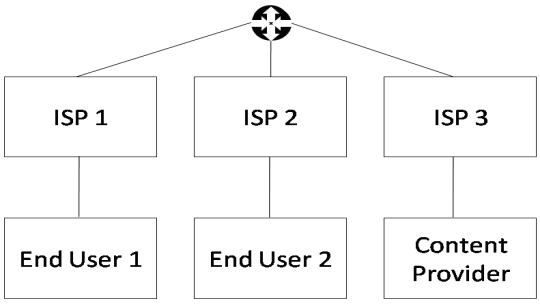


Fig. 2. A tiered view of the Internet without MH or CDN.

III. Modeling the Effects of Multi-Homing and CDNs

A key feature of the Internet is the fact that all traffic is being treated (more or less) equally. Methods for discriminating against specific types of traffic are quite expensive and have not been used widely so far. All traffic transport is a commodity. While this makes the Internet very egalitarian, it is a problem for content which needs assured performance parameters. Video streaming for example can only tolerate a certain package loss and Internet gaming requires assured low delays. Since the Internet cannot assure constant quality

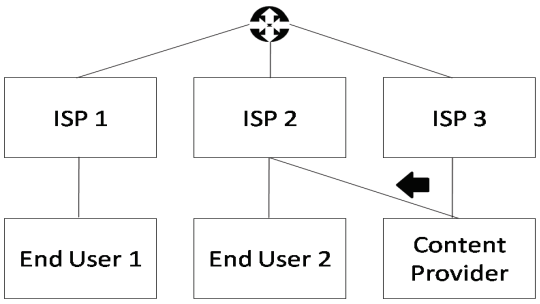


Fig. 3. The Internet with Multi-Homing

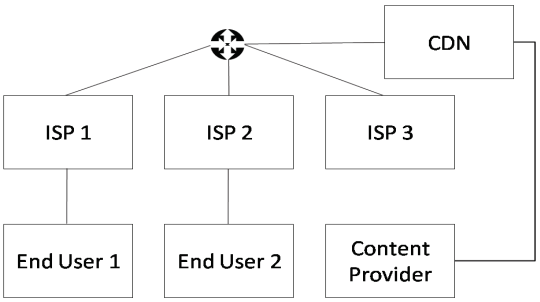


Fig. 4. The Internet with CDNs

levels methods to bypass the main bottlenecks are used by commercial CPs. Two relevant methods are multi-homing and CDN caching. An interesting aspect of both technologies is that they "decommoditize" Internet traffic because with multi-homing, the source of traffic is known to the terminating ISP and with the use of CDN it is still possible to infer the increased quality requirements of traffic originating from a CDN.

In the following we will analyze the four combinations of market form faced by the CP and strategic behavior of the ISPs shown in table I that result from the use of MH or CDN. The columns of table I represent the possible actions the terminating ISP can take: With MH it can perfectly discriminate against the known source of traffic and extract profits; with CDNs it can at least discriminate against all traffic that requires higher quality and has demonstrated this through the use of a CDN (self selection). The rows represent the market situation the CP is facing. Under competition many CPs are serving a market; with a monopoly, only one CP serves a specific market with one product. Cross product/market elasticities are assumed to be zero. The cells of the table show what the ISP will do to extract the maximum revenue from the CP for terminating traffic with EUs. All results are derived based on the assumption that price elasticities between quality traffic (CDN or MH) and normal traffic across peering points are zero: $\epsilon_{i,j} = 0$.

A. Scenario 0: No Discrimination by ISP Possible

The situation when the ISP has no way of differentiating different traffic and therefore cannot set different prices for

	Perfect / 1^{st} degree price discrimination (MH)	Class of Service / 2^{nd} degree price discrimination (CDN)
Content Competition	1: ISP sets access price until monopoly price for content is reached	2: ISP sets monopoly price for termination, inefficient since ISP sets avg. price
Content Monopoly	3: ISP extracts all revenue from CP through a two part tariff	4: ISP sets monopoly access price but cannot capture all rents from CPs (dbl. marginalization)

TABLE I
THE FOUR POSSIBLE SITUATIONS WITH AN ACCESS MONOPOLIST

different traffic sources and different quality classes (figure 2) is not shown in the table. It corresponds to the "normal" way traffic is exchanged on the internet. Since traffic transport is a commodity it is priced at marginal cost: Price for terminating traffic p_w is equal to the ISP's cost c. The charge p_w which is levied by the terminating ISP feeds through to the originating CP as a rise in connection costs to be included in the charge to the EU. The CP therefore does not have to play any strategic game but can simply involve in (depending on its market position) monopoly pricing or consider the market price as given under perfect competition. The whole structure is as efficient as the market between CP and EUs permits. This treatment ignores the considerations put forward by [4] about the access charge one ISP charges to another.

B. Scenario 1: Perfect Discrimination by ISP and Content Competition

The first scenario assumes that the terminating ISP can perfectly discriminate against CPs while the CPs face a competitive market which does not allow for profits above marginal cost. Assuming perfect competition only efficient firms are active in each differentiated market and the ISP can perfectly discriminate against each market segment (i.e. treat a whole efficient segment of CPs like one single CP). More fine grained segmentation based on individual CPs is not necessary. By setting its price $p_{w,i}$ charged to the CPs of segment i equal to the monopoly price the ISP can extract monopoly profits while leaving no profits to the CPs.

Each CP sets the competitive price p_i in segment i equal to its marginal costs. For simplicity we assume that marginal costs of production of the CPs are zero. Therefore $p_i = p_{w,i}$. The ISP with marginal costs c_i per market segment determines $p_{w,i}$ by solving the following problem:

$$\max_{p_{w,i}} \left[(p_{w,i} - c_i) D_i(p_{w,i}) \right]. \tag{1}$$

For example suppose that the demand function in one of the market segments i is given by $D_i(p_{w,i}) = 1 - p_{w,i}$. Then the target function of the ISP is:

$$\max_{p_{w,i}} \left[(p_{w,i} - c_i)(1 - p_{w,i}) \right]. \tag{2}$$

Solving this gives for price and output quantity

$$p_{w,i} = \frac{1 + c_i}{2}, \tag{3}$$

$$D_i(p_{w,i}) = q_i = \frac{1 - c_i}{2} \quad \text{and} \tag{4}$$

$$\Pi = p_{w,i} q_i - c_i = \left(\frac{1 - c_i}{2} \right)^2. \tag{5}$$

Interpretation: The ISP simply converts each competitive market segment i into perfect monopoly and extracts the maximum profit. The output quantity is reduced and the CPs make no profits. Note that this result would not hold with users being able to switch between ISPs. Switching users without switching costs would restore competitiveness of the market. Due to long term contracts, however, the ISP has some market power.

C. Scenario 2: Market Segmentation by ISP and Content Competition

This situation is similar to scenario 1 with a reduction in the ability to discriminate against CPs. With a CDN as mediator only CDN (i.e. quality sensitive) and non-CDN traffic can be distinguished. CDNs buy transit from the ISP who can thus differentiate between quality sensitive traffic from CDNs and ordinary traffic through normal peerings with other ISPs.

The price taken by the competitive CPs in this scenario depends on the price p_w set by the ISP which acts like a monopolist. The ISP faces an aggregate demand for quality traffic. Assuming that the CDN segment is competitive, the ISP can set monopoly prices for CDN traffic which raises marginal costs of the CDN providers and thus costs for CPs. In the following treatment we assume the CDN costs to be fully absorbed into the marginal cost of the CP. For simplicity both are assumed to be constant and zero. Again the CPs are competitive in each market segment. In each of n market segment the CPs have the target function

$$\max_{p_i} [(p_i - p_w) D_i(p_i)] \tag{6}$$

where p_i is the price charged to the EUs and p_w is the uniform price for CDN traffic paid to the ISP. Different from scenario one, now the ISP can not set different $p_{w,i}$ per market segment but just one averaged price for CDN traffic. Under perfect competition price equals marginal costs:

$$p_i = p_w : \tag{7}$$

Interpretation: It is obvious form equation 6 that the ISP influences the pricing decisions of the CP's through its price p_w. However, since the ISP can only set a a uniform price for all CND traffic, it can only optimize across an aggregated demand function $D = \sum_{i=1}^{n} D_i$. Therefore prices in the EU market need not be those an integrated monopolist would have chosen. In this situation inefficiencies are present due to the fact that the ISP charges an average monopoly price to the CPs. Some segments profit from this by paying a price below the monopoly level, thus improving efficiency, while other

segments pay more than the optimal monopoly price which might result in some markets not being served.

D. Scenario 3: Perfect Discrimination by ISP and Content Monopoly

In this case, the ISP has great power over the CPs since it can target them individually and set individually profit maximizing prices. At first sight this situation seems prone to double marginalization with inefficiencies induced by two cascaded monopoly decision. However, the ISP can anticipate this sort of inefficiency and avoid it by setting an efficient usage price p_w for access to its network and then extracting all profits of the monopoly CP through a fixed fee A. This way of pricing is known in the literature as franchising [18]. The total price set by the ISP is

$$T(q) = A + p_w q. \tag{8}$$

Assuming the ISP sets $p_w = c$ efficiently at the marginal cost of providing the service (the variable q is the output quantity of the ISP), the CP choses the optimal end user price p:

$$\max_p \left[(p-c)D(p) - A\right]. \tag{9}$$

This results in the monopoly EU price

$$p = \frac{1+c}{2}. \tag{10}$$

By setting the fixed fee A equal to the profit of the CP the ISP can now extract all profit without distorting the target function of the CP.

The same result could have been achieved with a "tax" $\tau\Pi$ levied by the ISP. Information requirements to enforce such a "tax" are lower than for the fixed fee.

Interpretation: An interesting aspect of the profit based charge A is that it does not alter the price paid by consumers but shifts profits from the CP to the ISP. This can be a problem since monopoly profits could be the reward for innovation and if those profits are taken away, innovation might not be profitable any more.

This situation is good for static efficiency since the output decision of the CP is not changed by the ISPs behavior. However, this becomes a problem when considering the development over time. No matter how much of the CP's revenue is extracted by the ISP, the EU price for the content stays the same and thus there is no competitive pressure from EUs on the ISP.

E. Scenario 4: Market Segmentation by ISP and Content Monopoly

In this case each of n CPs serves a monopoly market as in case three but the ISP can not differentiate between those markets and can only optimize its revenue based on gross demand (as in case two) by all fully differentiated CPs.

This case corresponds to a situation know as double marginalization [18]. In this situation the CPs (with constant marginal cost set to zero) set their price for content as a standard monopolist. Their target functions are:

$$\max_{p_i}[(p_i - p_w)D_i(p_i)] \tag{11}$$

where p_i is the price charged to the EU and p_w is the price paid to the ISP.

Since the ISP can only 2^{nd} degree price discriminate we will analyze the ISP's problem based on average figures for price and demand from all CPs demanding quality (using CDN):

$$p = \frac{1}{n}\sum_{1=1}^{n} p_i \qquad \text{and} \tag{12}$$

$$D = \frac{1}{n}\sum_{1=1}^{n} D_i(p_i). \tag{13}$$

Assuming for simplicity that aggregate demand is linear: $D(p) = 1 - p$, the solution to this problem is identical to the treatment in section III-B and yields optimal prices of

$$p = \frac{1+p_w}{2} \tag{14}$$

which corresponds to an average output quantity per CP of

$$q = \frac{1-p_w}{2}. \tag{15}$$

Using this quantity, the optimization problem of the ISP with marginal output cost c for CND traffic then is:

$$\max_{p_w}\left[(p_w - c)\left(\frac{1-p_w}{2}\right)\right] \tag{16}$$

resulting in

$$p_w = \frac{1+c}{2}. \tag{17}$$

Now we can go back to formula 14 and paste p_w from equation 17 and get

$$p = \frac{3+c}{4} \tag{18}$$

for the final consumer price. Comparing the sum of the profits of ISP Π_{isp} and CP Π_{cp} we see that it is smaller than the profit Π_{int} of an integrated monopoly provider:

$$\Pi_{isp} + \Pi_{cp} < \Pi_{int} \tag{19}$$

$$\frac{(1-c)^2}{8} + \frac{(1-c)^2}{16} < \frac{(1-c)^2}{4}. \tag{20}$$

The end user price $p = \frac{1+c}{2}$ in monopoly is lower than that in equation 18 (as long as c<1 which makes sense because with $D = 1 - p$, $1 > p \geq p_w \geq c$ must be true if no losses are made and there is a positive demand).

Interpretation: On average across EUs' demand for the perfectly differentiated markets this situation is suboptimal since prices could be lower and revenues could be higher.

In addition to this double marginalization problem there exists an problem due to the averaging of price and demand of the different CPs. Since the price set by the ISP is targeted at the average CP, it will typically be either too high or too low. Thus there is a second source of inefficiency. ISPs could also opt for a different pricing model and charge (as in scenario

3) a fixed fee plus an efficiently set usage fee p_w. Thus inefficiencies due to the ISPs behavior would be removed. However, since the ISP can only set an average A and thus it is impossible to extract all rents from the CPs. A would even make some market segments with low profits unattractive to serve since profits are too low to cover the fixed fee. Thus one has to chose between double marginalization reducing output in some segments and franchising resulting in some markets not being served at all.

F. Synopsis of the Four Scenarios

The first conclusion from the above analysis is that in all four scenarios welfare is below optimal and prices are above the competitive level. It is even possible as shown in case four that welfare is lower than in monopoly. The first general result therefore is that CNDs and multi-homing reduce welfare due to reducing the efficiency of price setting for data transport. This result is true if one ignores the welfare gains of being able to deliver higher QoS by the use of those technologies which might enhance welfare.

The second result which is common to all cases is that ISPs' price setting reduces CPs' profits. In some cases all profits may be extracted. While this in itself does not need have a negative effect on welfare in a static environment it can be detrimental when considering monopoly revenues of CPs as the reward for innovation. If ISPs extract these profits innovation might become unprofitable for CPs.

Furthermore ISPs are able to exploit their access monopoly and create monopolies from otherwise competitive markets. The only precondition for this is a quality requirement of the service that does not allow the use of ordinary peerings. For the result to be true in the presented pure form it is necessary that price elasticities between multi homing / CDN and ordinary peering are zero, i.e. ordinary peerings are no substitute for multi homing / CDN. Nonzero elasticities will soften the results.

IV. Conclusions and Further Research

This work provides a new view on access pricing and quality of service on the internet. Assuming that CND and multi-homing are used to improve the quality of service provided to the EU we have shown how a "decommoditization" of traffic enables the terminating ISP to charge more for termination than it would do under competition.[2]

Our analysis shows that there exist incentives for ISPs to further degrade peering quality to attract more traffic to the more profitable segments in which price discrimination is possible. Already peerings are problematic bottlenecks on the Internet and since they can only be set up cooperatively by two ISPs they are usually too small to accommodate all traffic during peak times. For the reasons just presented this situation is unlikely to change.

[2]The termination fee without MH/CND can in general not be considered to be set at the efficient level (our scenario zero). Consider for example [13], [12] for an extensive discussion of this topic.

On the positive side, understanding multi-homing and CDNs as quality mechanisms opens up a whole new view on the quality of service debate. Standard approaches of QoS always require global carrier collaboration. All carriers have to agree on service classes and forward each others traffic with the appropriate service level. With agreements on peerings being so hard today, it is rather unlikely that carriers will agree on such ground braking fields as changing the Internet protocol. With multi-homing and CDN, an edge based solution to QoS is available that can (in combination with network oriented changes that only need to take place inside each carriers network) deliver QoS for many applications on the Internet.

The presented work leaves open and poses many further research question that need to be addressed. Firstly the four presented scenarios rely on economic theory that was created for one sided markets with one type of buyers and sellers. With non commodity traffic and pricing power over CPs, ISPs now have to base their considerations on a two sided market scheme with two types of customers. Such an analysis would move the analysis of Internet ISPs' decision problems closer to the classical analysis of voice telecommunications providers [12], [13], [19]. This aspect could provide further insights into the pricing decisions of the ISP since it will possibly engage in exploiting one group of customers and subsidize the other in order to maximize its revenue.

Furthermore the whole setup chosen in this work relies on EUs that are trapped with their provider. With switching consumers the results would probably be softened. The same is true for allowing nonzero cross elasticities between the MH / CDN segment and the best effort / ordinary peering segment.

Besides economic questions that need further clarification there is also a more technical side. Can CND and MH fully replace inter carrier agreements on quality parameters of traffic? Which quality mechanisms are necessary inside one carriers network to complement the peering bypass capability of CDN and MH with the ability to deliver to the EUs workstation?

There are many questions such as multi cast communication and peer to peer that we have not addressed. However, the text presents a new perspective on the QoS debate and on economic aspects of CDN and MH. We believe this paper will add an important building block to our understanding of QoS on the Internet and spark ideas to QoS enable the web.

References

[1] G. Shrimali and S. Kumar, "Bill-and-keep peering," *Telecommunications Policy*, vol. 32, pp. 19–32, Feb. 2008.

[2] M. Cave, S. K. Majumdar, and I. Vogelsang, *Handbook of telecommunications economics*. Elsevier Boston, Mass, 2002.

[3] S. Shakkottai and R. Srikant, "Economics of network pricing with multiple isps," *IEEE/ACM Trans. Netw.*, vol. 14, pp. 1233–1245, 2006.

[4] J.-J. Laffont, S. Marcus, P. Rey, and J. Tirole, "Internet interconnection and the off-net-cost pricing principle," *The RAND Journal of Economics*, vol. 34, no. 2, pp. 370–390, 2003.

[5] S. Uludag, K. Lui, K. Nahrstedt, and G. Brewster, "Analysis of topology aggregation techniques for qos routing," *ACM Computing Surveys*, vol. 39, no. 3, 2007.

[6] Z. Wang, *Internet QoS: Architectures and Mechanisms for Quality of Service*. Morgan Kaufmann, 2001.

[7] M. Armstrong, "Network interconnection in telecommunications," *The Economic Journal*, vol. 108, no. 448, pp. 545–564, May 1998.
[8] ——, "Competition in two-sided markets," *RAND Journal of Economics*, vol. 37, no. 3, pp. 668–691, 2006.
[9] J. Cremer, P. Rey, and J. Tirole, "Connectivity in the commercial internet," *The Journal of Industrial Economics*, vol. 48, pp. 433–472, Dec 2000.
[10] N. Badasyan and S. Chakrabarti, "A simple game-theoretic analysis of peering and transit contracting among internet service providers," *Telecommunications Policy*, vol. 32, pp. 4–18, Feb. 2008.
[11] O. Foros, H. J. Kind, and J. Y. Sand, "Do internet incumbents choose low interconnection quality?" *Information Economics and Policy*, vol. 17, no. 2, pp. 149–164, Mar. 2005.
[12] J.-J. Laffont, P. Rey, and J. Tirole, "Network competition: II. price discrimination," *The RAND Journal of Economics*, vol. 29, no. 1, pp. 38–56, 1998.
[13] J. J. Laffont, P. Rey, and J. Tirole, "Network competition: I. overview and nondiscriminatory pricing," *The RAND Journal of Economics*, vol. 29, no. 1, pp. 1–37, 1998.
[14] J. J. Laffont and J. Tirole, *Competition in Telecommunications*. MIT Press, 2000.
[15] J. J. Laffont, S. Marcus, P. Rey, and J. Tirole, "Internet peering," *The American Economic Review*, vol. 91, no. 2, pp. 287–291, 2001.
[16] P. Klemperer, "Competition when consumers have switching costs: An overview with applications to industrial organization, macroeconomics, and international trade," *Review of Economic Studies*, vol. 62, pp. 515–539, 1995.
[17] DE-CIX, "http://www.de-cix.net/content/clients.html," 2008.
[18] J. Tirole, *The Theory of Industrial Organization*. MIT Press, 1988.
[19] S. C. Littlechild, "Mobile termination charges: Calling party pays versus receiving party pays," *Telecommunications Policy*, vol. 30, no. 5-6, pp. 242–277, 2006.

QUALITY OF SERVICE DELIVERY: ECONOMIC EFFECTS OF MULTI-HOMING AND CONTENT DELIVERY NETWORKS

Thorsten Hau[1], Jochen Wulf, Rüdiger Zarnekow[2], Walter Brenner[1]

Abstract
The structure of the Internet serves as a big "commoditizer" of all traffic. Therefore all data, be it time critical or not is transported at the same speed. However, recent trends in the internet are changing this structure. The practices of multi-homing and using content delivery networks reduce the commodity nature of data being transported and put terminating Internet service providers in a position to price discriminate against specific providers or types of traffic. We firstly formalize multi-homing and content delivery networks, we then derive end user prices for paid content and lastly show consequences of the modeled situation. We thus show how the two technologies to bypass crowded peerings change the Internet business model. Traffic which is sensitive to transport quality, such as business critical or delay sensitive traffic, will be paying higher fees to terminating ISPs.

1. INTRODUCTION

It is a trivial thought that content (a service, music or text) needs to be delivered to its consumer. However, apparently this problem is by and large being ignored in the web service hype. Implicitly in all service visions it is assumed that the Internet is there and available in appropriate quality. There is for example no work that the authors are aware of, which connects web service availability with the use of the underlying network infrastructure. The available literature on web service standards does consider quality of service of web services but only as a question of setting a protocol standard to exchange quality information between two nodes in a network [7, 23]. A first tentative investigation into the matter of availability of web services is provided by [16] and [17]. Even though [16] do not attribute the differences in download speed to connection quality, such an interpretation does not seem far-fetched, especially when considering the geographical disparities in the measured data.
The authors of this work believe that one key factor for the slow uptake of services (SaaS for example) by businesses is a result of the performance and reliability problems [15]; and these are in large part due to today's Internet. This work focuses on one specific aspect of quality of service of

[1] Institut für Wirtschaftsinformatik, Universität St. Gallen
[2] Institut für Technologie und Management, Technische Universität Berlin

the Internet and analyses the economic impact of two quality assurance technologies on today's pricing regime of the Internet.
Internet pricing is a vast field of research that has its roots in telecommunications. This paper departs somewhat from the "classical" literature on communications pricing by considering a special pricing problem present in today's Internet that has not been covered by the extant literature. We focus on content oriented pricing of network services. In such a scenario Internet service providers (ISPs) with access to end users (EUs) can discriminate against content providers and charge higher prices for termination. This scenario departs from the idealized bill and keep regime that is often used as a basis for analysis of network pricing decisions. However, our scenario is quite realistic. It is commonplace that content providers (CPs) directly buy transit from terminating ISPs, thus effectively paying them for preferential access to end users. This is a viable strategy because by bypassing crowded peerings used by the CP's original ISP, the CP gets higher quality access to EUs. This is due to the fact that peering points are a major bottleneck on the Internet. While there are usually few capacity problems present inside a single provider's network, peerings are often at their capacity limit [19], [4]. For the purpose of this paper we call this practice multi-homing (MH). Content delivery networks (CDNs) are also a popular way to enhance the flow of information on the web. A CDN uses local caches to keep distributed images of content close to EUs thus bypassing peerings. The CDN effectively buys transit from the ISP that terminates the traffic.
The paper is structured as follows: First we explain the relevant entities of the current Internet that we need for a formal model. Then we present a formalized treatment of four scenarios that show how CDN and MH affect ISPs incentives to price traffic. Lastly we discuss consequences of our model and sketch out an agenda for further research.

2. STATUS QUO: THE MARKET FOR INTERNET CONNECTIVITY

2.1. Overview

Figures 1 and 2 show the structure of the Internet as commonly used to model the Internet [18], [12] and [21]. In figure 2 at the bottom, customers and content providers receive and send traffic, respectively. ISPs interconnect to exchange traffic to and from their customers (EUs or CPs). In figure 2 traffic could for example originate at the content provider and be passed on to its ISP 3. From there it is routed to ISP 1 to be terminated at end user (EU) one. The ISPs are fully meshed, each one interconnecting with all the others. It is a common approximation [12] that CPs (web sites) only send traffic and EUs only receive traffic.

2.2. Internet Service Providers

ISPs provide connectivity to Internet users who are either end users or content providers. The ISPs pay the market price pw to the ISPs for termination of traffic. Typically ISPs have no lack of bandwidth on their backbones and can provide quality assurance to traffic either through excess capacity or network management techniques commonly employed only within a single provider's network [22]. Bottlenecks (Not meant in the sense of bottleneck facility as in [1, 2] may be present in the peering points and in the access network. We ignore possible problems due to constrained access bandwidth and concentrate on the peerings. All of the following ignores dynamic considerations of integration and foreclosure through the ISP.

2.3. Points of Interconnection

In figure 2 the circle with arrows in upper middle represents the point of interconnection where the three ISPs interconnect their networks. This is also called a "peering point". There are two dominant modes of interconnection: Peering and transit. Peering is a settlement free agreement to exchange traffic while transit involves payment for exchanged data typically on a bit-rate per time scale. Typically peering agreements are used between ISPs of similar size while transit is paid from small ISPs to larger ISPs.

Peering points are among the major bottlenecks of the Internet because it always takes both parties to agree to extend the capacity of a peering point. Since the telecommunications firms who own the Internet infrastructure tend to be not overly cooperative these peering points represent excellent opportunities for power play or generally uncooperative behavior. Peerings are frequently overloaded at peak times because one of the two parties involved has no interest in paying for extending its capacity [1] [5], [3], [8]. Ways for CPs to circumvent frequently overloaded (and therefore slow and packet dropping) peerings are multi-homing and the use of CDN services.

Transit in contrast to peering involves a payment from one ISP to the other for the delivery of traffic. With such an agreement a guaranteed bandwidth is bought. Due to strategic considerations the biggest networks only peer among themselves and charge smaller networks for sending traffic to them. Since small ISPs have to pay for sending traffic to larger networks which is necessary to reach the whole Internet they optimize their out-payments for transit fees by buying the least amount of bandwidth their users will tolerate.

2.4. Content Providers

Content Providers are for example websites that buy connectivity to the Internet from an ISP (possibly mediated through a hosting provider).Content providers are able to multi-home. They can buy connectivity for one fraction of their traffic from IPS1 and the rest from ISP 2. Furthermore they can host their service with any provider anywhere in the world giving them a very large set of ISPs to choose from. This creates competition amongst ISPs for CPs' business and therefore CPs face a market price for Internet connectivity based on perfect competition. This price only includes un-prioritized traffic transportation across peering points. Canonical analysis [10-12, 14] usually assumes the following model of Internet payments: $EU \rightarrow ISP_t \leftrightarrow_a ISP_o \leftarrow CP$ (t=terminating, o=originating), ignoring where the CP gets funding from and emphasizing the analysis of the inter ISP settlement *a*. Focusing on payments for content viewing, we model the payments flows according to the following scheme: $ISP_t \leftarrow_a ISP_o \leftarrow_{p_w} CP \leftarrow_p EU$. We ignore payments from the EU to the terminating ISP for access to the Internet. Payments from the EU to the CP might be paid through watching advertisements or direct money exchange. The arrow with the a stands for the access charge that is paid from one ISP to the other for sending traffic to the receiving ISP[13]. p is the final price paid by the EU for viewing some content. p_w is the price paid from the CP to the ISP for reaching the EU. If the ISP receiving pw cannot terminate the traffic it has to pay an access charge to another ISP able to terminate the traffic. If a CDN is involved in content delivery, the CP has to pay the cost of the CDN, too. Modifying the way CPs get access to the EUs and modifying payments accordingly will be the key part of this paper. The two variations we will consider are: multi-homing and CDN delivery. With MH, the terminating ISP is directly connected with the CP, while with CDNs a neutral third party mediates between CP and ISP_t. Under MH payment flows are $ISP_t \leftarrow_{p_w} CP \leftarrow_p EU$ and the originating ISP is cut out of the equation. With CDN delivery, the payments are: $ISP_t \leftarrow_{p_w} CDN \leftarrow_{p_w+c_{cdn}} CP \leftarrow_p EU$ We model CDNs as fully competitive entities only passing on costs.

2.5. End Users

Unlike CPs, EUs cannot divide their traffic amongst several ISPs and are immobile in the sense that they cannot chose their provider globally but need to chose among a small number of local ISPs. Therefore we consider a static scenario in which EUs are bound to their ISP, providing the ISP with a monopoly over terminating traffic to those EUs. Neither the competition among local ISPs for market share nor the dynamic considerations of consumers with switching costs [9] are considered here.

2.6. Multi-homing

Multi-homing is the common practice that CPs interconnect with several different ISPs. There are several instances of big content providers that are present in public peering points in order to gain direct access to ISPs with many EUs. Google for example is not just with one ISP in the USA but has its own transatlantic fiber capacity and is present in DE-CIX [6]. Besides the reliability aspect this practice allows the CPs to get direct access to an ISP's network and thus be more in control over bandwidth agreements with that ISP.

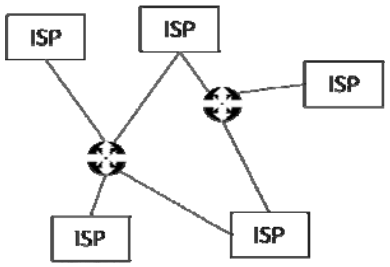


Figure 1: Simplified model of the Internet.

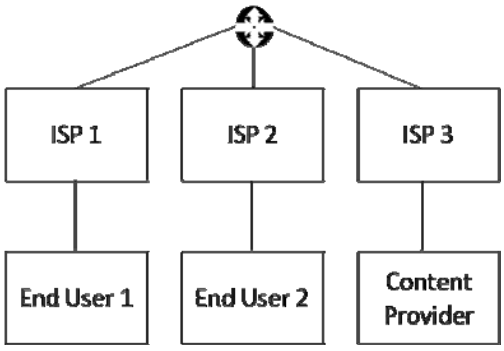


Figure 2: A tiered view of the Internet.

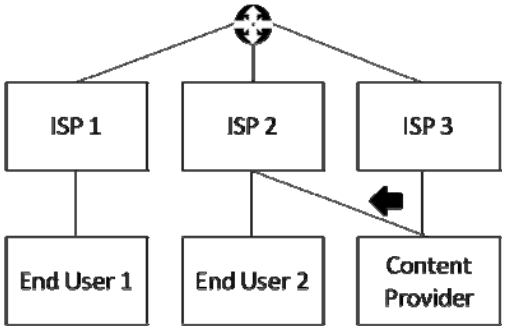


Figure 3: The Internet with MH.

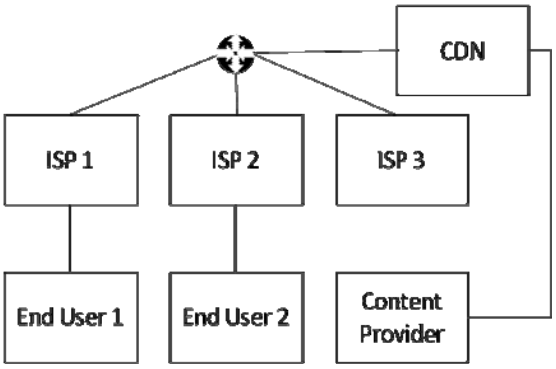


Figure 4: The Internet with CDNs.

2.7. Content Delivery Networks

CDNs consist of a network of servers that are distributed around the Internet within many ISPs' infrastructures. A CDN takes content from a CP and caches it on those distributed servers, which has the effect that content is brought closer to the EU without passing through peerings. The CDN then delivers the content it is caching from the mirror site to the EU. By using the services of a CDN a CP does not need to multi-home with every possible network. We assume that CDNs do not make any profit and charge efficient prices.

3. MODELING THE EFFECTS OF MULTI-HOMING AND CDN

A key feature of the Internet is that all traffic is being treated (more or less) equally. Methods for discriminating against specific types of traffic are expensive and are not used aggressively. While this makes the Internet very egalitarian, it is a problem for content which needs assured performance parameters. Video streaming for example can only tolerate a certain package loss and Internet gaming requires assured low delays. Since the Internet cannot assure constant quality levels, methods to bypass the main bottlenecks are used by commercial CPs: CDN and MH. Both technologies "de-commoditize" Internet traffic because the source of traffic becomes a business partner.

In the following we will analyze the four combinations of market form faced by the CP and strategic behavior of the ISPs. The columns of table 1 represent the possible actions the terminating ISP can take: With MH it can perfectly discriminate against the known source of traffic and extract profits; with CDNs it can discriminate against all traffic that requires higher quality and has demonstrated this through the use of a CDN (self selection). The rows represent the market situation the CP is facing. Under competition many CPs are serving a market; with a monopoly, only one CP serves a specific market with one product. Cross product/market elasticities are assumed to be zero. The cells of the table show what the ISP will do to extract the maximum revenue from the CP.

	Perfect / 1st degree price discrimination (MH)	Class of Service / 2nd degree price discrimination (CDN)
Content Competition	1: ISP sets access price until monopoly price for content is reached	2: ISP sets monopoly price for termination, inefficient since ISP sets avg. price
Content Monopoly	3: ISP extracts all revenue from CP through two part tariff	4: ISP sets monopoly access price but cannot capture all rents from CPs (dbl. marginalization)

Table 1: The four possible situations with an access monopolist

3.1. Scenario 0: No Discrimination by ISP Possible

The situation when the ISP has no way of differentiating different traffic, and therefore cannot set different prices for different traffic sources and different quality classes, is not shown in table 1. It corresponds to the "normal" way traffic is exchanged on the Internet. The price for terminating traffic pw is equal to the ISP's cost c. The charge pw which is levied by the terminating ISP feeds through to the originating CP as a rise in connection costs to be included in the charge to the EU. The CP therefore can consider the market price as given. This treatment ignores the considerations put forward by [12].

3.2. Scenario 1: Perfect Discrimination by ISP and Content Competition

Here the terminating ISP can perfectly discriminate against CPs while the CPs face a competitive market. Assuming perfect competition only efficient firms are active in each differentiated market and the ISP can perfectly discriminate against each market segment (i.e. treat a whole segment like one single CP). By setting its price $p_{w,\mathrm{i}}$ charged to the CPs of segment i equal to the monopoly price the ISP can extract monopoly profits while leaving no profits to the CPs. Each CP sets the competitive price p_{i} in segment i equal to its marginal costs. For simplicity we assume that

marginal costs of production of the CPs are zero. Therefore $p_i = p_{w,i}$. The ISP with marginal costs c_i per market segment determines $p_{w,i}$ by solving the following problem:

$$\max_{p_{w,i}} \left[(p_{w,i} - c_i) D_i(p_{w,i}) \right]. \quad (1)$$

For example suppose that the demand function in one of the market segments i is given by $D_i\left(p_{w,i}\right) = 1 - p_{w,i}$. Then the target function of the ISP is:

$$\max_{p_{w,i}} \left[(p_{w,i} - c_i)(1 - p_{w,i}) \right]. \quad (2)$$

Solving this gives for price and output quantity yields

$$p_{w,i} = \frac{1 + c_i}{2}, \quad (3)$$

$$D_i(p_{w,i}) = q_i = \frac{1 - c_i}{2} \text{ and} \quad (4)$$

$$\Pi = p_{w,i} q_i - c_i q_i = \left(\frac{1 - c_i}{2} \right)^2. \quad (5)$$

Interpretation: The ISP converts each competitive market-segment *i* into perfect monopoly and extracts the maximum profit. The output quantity is reduced and the CPs make no profits. Note that this result would not hold with users being able to switch between ISPs. Switching users without switching costs would restore competitiveness of the market. Due to long term contracts, however, the ISP has some market power.

3.3. Scenario 2: Market Segmentation by ISP and Content Competition

This situation is similar to scenario one with a reduction in the ability to discriminate against CPs. With a CDN as mediator only CDN (i.e. quality sensitive) and non-CDN traffic can be distinguished. CDNs buy transit from the ISP who can thus differentiate between quality sensitive traffic from CDNs and ordinary traffic through normal peerings with other ISPs.

The price taken by the competitive CPs depends on the price pw set by the ISP which acts like a monopolist. The ISP faces an aggregate demand for quality traffic. Assuming that the CDN segment is competitive, the ISP can set monopoly prices for CDN traffic which raises marginal costs of the CDN providers and thus costs for CPs. In the following treatment we assume the CDN costs to be fully absorbed into the marginal cost of the CP. For simplicity both are assumed to be constant and zero. Again the CPs are competitive in each market segment. In each of n market segment the CPs have the target function

$$\max_{p_i} [(p_i - p_w) D_i(p_i)] \quad (6)$$

where pi is the price charged to the EUs and pw is the uniform price for CDN traffic paid to the ISP. Different from scenario one, now the ISP cannot set different $p_{w,i}$ per market segment but just one averaged price for CDN traffic. Under perfect competition price equals marginal costs:

$$p_i = p_w. \quad (7)$$

Interpretation: It is obvious form equation (6) that the ISP influences the pricing decisions of the CP's through its price pw. However, since the ISP can only set a a uniform price for all CDN traffic, it can only optimize across an aggregated n demand function $D = \sum D_i$. Therefore prices in the EU market need not be those an integrated monopolist would have chosen. In this situation inefficiencies are present since the ISP charges an average price. Some segments profit from this by paying a price below the monopoly level, thus improving efficiency, while other segments pay more than the optimal monopoly price which might result in some markets not being served.

3.4. Scenario 3: Perfect Discrimination by ISP and Content Monopoly

The ISP can target the CPs individually and set individually profit maximizing prices. This situation seems prone to double marginalization with two cascaded monopolies. However, the ISP can anticipate this sort of inefficiency and avoid it by setting a usage price p_w and then extracting all profits of the monopoly CPs through a fixed fee A. This way of pricing is known as franchising [20]. The total price set by the ISP is

$$T(q) = A + p_w q. \tag{8}$$

Assuming that the ISP sets $p_w = c$ efficiently at the marginal cost of providing the service (the variable q is the output quantity of the ISP), the CP chooses the optimal end user price p:

$$\max_p \left[(p-c)D(p) - A\right]. \tag{9}$$

This results in the monopoly EU price

$$p = \frac{1+c}{2}. \tag{10}$$

By setting the fixed fee A equal to the profit of the CP the ISP can now extract all profit without distorting the target function of the CP. The same result could have been achieved with a "tax" $\tau\,\Pi$ levied by the ISP.

Interpretation: The profit based charge A does not alter the price paid by consumers but shifts profits from the CP to the ISP. This can be a problem since monopoly profits could be the reward for innovation and if those profits are taken away, innovation might not be profitable any more. This situation is good for static efficiency since the output decision of the CP is not changed by the ISP's behavior. However, when considering the development over time, no matter how much of the CPs' revenue is extracted by the ISP, the EU price for the content stays the same and thus there is no competitive pressure from EUs on the ISP.

3.5. Scenario 4: Market Segmentation by ISP and Content Monopoly

Each of n CPs serves a monopoly market as in case three but the ISP cannot differentiate between those markets and can only optimize its revenue based on gross demand (as in case two) by all fully differentiated CPs. This corresponds to a situation know as double marginalization. Now the CPs (with constant marginal cost set to zero) set their price for content as a standard monopolist. Their target functions are:

$$\max_{p_i} [(p_i - p_w) D_i(p_i)] \tag{11}$$

where p_i is the price charged to the EU and p_w is the price paid to the ISP.

Since the ISP can only 2^{nd} degree price discriminate we will analyze the ISP's problem based on average figures for price and demand from all CPs demanding quality (using CDN):

$$p = \frac{1}{n}\sum_{i=1}^{n} p_i, \text{ and} \tag{12}$$

$$D = \frac{1}{n}\sum_{1=1}^{n} D_i(p_i). \tag{13}$$

Assuming for simplicity that aggregate demand is linear: $D(p) = 1 - p$, the solution to this problem is similar to the treatment in section 3.2 with the c_i replaced here by p_w and yields optimal prices of $p = \frac{1+p_w}{2}$ and an optimal output $q = \frac{1-p_w}{2}$. Using this quantity, the optimization problem of the ISP with marginal output cost c for CDN traffic is:

$$\max_{p_w} \left[(p_w - c)\left(\frac{1-p_w}{2}\right)\right] \tag{14}$$

resulting in $p_w = \frac{1+c}{2}$ and $p = \frac{3+c}{4}$ for the final consumer price. Comparing the sum of the profits of ISP Π_{isp} and CP Π_{cp} we see that it is smaller than the profit Π_{int} of an integrated monopoly provider:

$$\Pi_{isp} + \Pi_{cp} < \Pi_{int} \Leftrightarrow \frac{(1-c)^2}{8} + \frac{(1-c)^2}{16} < \frac{(1-c)^2}{4}. \quad (15)$$

The end user price $\mathrm{p} = \frac{1+c}{2}$ in monopoly is lower than the $p = \frac{3+c}{4}$ above (as long as $\mathrm{c} < 1$ which makes sense because with $\mathrm{D} = 1 - p,\ 1 > \mathrm{p} \geq p_w \geq \mathrm{c}$ must be true if no losses are made and there is a positive demand).

Interpretation: On average across EUs' demand for the perfectly differentiated markets this situation is suboptimal since prices could be lower and revenues could be higher. In addition to this double marginalization problem, there exists a problem due to the averaging of price and demand of the different CPs. Since the price set by the ISP is targeted at the average CP, it will typically be either too high or too low. Thus there is a second source of inefficiency. ISPs could also opt for a different pricing model and charge (as in scenario 3) a fixed fee plus an efficiently set usage fee p_w. Thus inefficiencies due to the ISPs behavior would be removed. However, since the ISP can only set an average A it is impossible to extract all rents. A would even make some market segments with low profits unattractive to serve since profits are too low to cover the fixed fee. Thus one has to chose between double marginalization reducing output in some segments and franchising resulting in some markets not being served at all.

3.6. Synopsis of the Four Scenarios

In all four scenarios welfare is below optimal and prices are above the competitive level. It is even possible as shown in case four that welfare is lower than in monopoly. The first general result therefore is that CDNs and multi-homing reduce welfare due to reducing the efficiency of price setting for data transport. This result is true if one ignores the welfare gains of being able to deliver higher QoS by the use of those technologies.

The second result which is common to all cases is that ISPs' price setting reduces CPs' profits. While this in itself does not need have a negative effect on welfare in a static environment it can be detrimental when considering monopoly revenues of CPs as the reward for innovation. If ISPs extract these profits, innovation might become unprofitable for CPs. Furthermore ISPs are able to exploit their access monopoly and create monopolies from otherwise competitive markets. The only precondition for this is a quality requirement of the service that does not allow the use of ordinary peerings.

4. CONCLUSIONS AND FURTHER RESEARCH

This work provides a new view on access pricing and quality of service on the Internet. Assuming that CDN and multi-homing are used to improve the quality of service provided to the EU we have shown how a "de-commoditization" of traffic enables the terminating ISP to charge more for termination. Our analysis shows that there exist incentives for ISPs to further degrade peering quality to attract more traffic to the more profitable segments. On the positive side, understanding multi-homing and CDNs as quality mechanisms opens up a whole new view on the quality of service debate. Standard approaches of QoS always require global carrier collaboration. All carriers have to agree on service classes and forward each other's traffic with the appropriate service level. With multi-homing and CDN, an edge based solution to QoS is available that can deliver QoS for many applications on the Internet.

The presented work leaves open and poses many further research question that need to be addressed. On the technical side, it would be interesting to know whether CDN and MH can fully replace inter carrier agreements on quality parameters of traffic. Which quality mechanisms are necessary inside one carrier's network to complement the peering bypass capability of CDN and MH with the ability to deliver to the EUs workstation? On the service oriented side, more work needs to be done to better understand the (non-) adoption of services and its connection with (un-) reliability and issues of the Internet. In an experimental set-up with one service hosted by CDN or MH and one service hosted within another network one could empirically validate the influence of peerings on user satisfaction.
There are many questions that we have not addressed. However, the text presents a new perspective on the QoS debate and on economic aspects of CDN and MH. Services hosted on the Internet are a case in point for the quality sensitivity of consumers and the lack of quality in the internet. We believe this paper will add an important building block to our understanding of QoS on the Internet and spark ideas to QoS enable the web.

REFERENCES

[1] ARMSTRONG, M., "Network Interconnection in Telecommunications," The Economic Journal, vol. 108, pp. 545-564, 1998.

[2] ARMSTRONG, M., "Competition in two-sided markets.," RAND Journal of Economics, vol. 37, pp. 668-691, 2006.

[3] BADASYAN, N.C., SUBHADIP, "A simple game-theoretic analysis of peering and transit contracting among Internet service providers," Telecommunications Policy, vol. 32, pp. 4-18, 2008.

[4] CAVE, M.M., S. K. & VOGELSANG, I., Handbook of telecommunications economics: Elsevier Boston, Mass, 2002.

[5] CREMER, J.R., PATRICK & TIROLE, JEAN, "Connectivity in the Commercial Internet," The Journal of Industrial Economics, vol. 48, pp. 433-472, 2000.

[6] DE-CIX, "http://www.de-cix.net/content/clients.html," 2008.

[7] ERRADI, A. and MAHESHWARI, P., "A broker-based approach for improving Web services reliability," presented at Web Services, 2005. ICWS 2005. Proceedings. 2005 IEEE International Conference on, 2005.

[8] FOROS, O.K., HANS JARLE & SAND, JAN YNGVE, "Do internet incumbents choose low interconnection quality?," Information Economics and Policy, vol. 17, pp. 149-164, 2005.

[9] KLEMPERER, P., "Competition when Consumers have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade," Review of Economic Studies, vol. 62, pp. 515-539, 1995.

[10] LAFFONT, J.-J., REY, P., and TIROLE, J., "Network Competition: II. Price Discrimination," The RAND Journal of Economics, vol. 29, pp. 38-56, 1998.

[11] LAFFONT, J.-J., REY, P., and TIROLE, J., "Network Competition: I. Overview and Nondiscriminatory Pricing," The RAND Journal of Economics, vol. 29, pp. 1-37, 1998.

[12] LAFFONT, J.-J., MARCUS, S., REY, P., and TIROLE, J., "Internet Interconnection and the Off-Net-Cost Pricing Principle," The RAND Journal of Economics, vol. 34, pp. 370-390, 2003.

[13] LAFFONT, J.J.M., S.; REY, P. & TIROLE, J., "Internet Peering," The American Economic Review, vol. 91, pp. 287-291, 2001.

[14] LAFFONT, J.J.T., J., Competition in Telecommunications: MIT Press, 2000.

[15] MA, Q., PEARSON, J., and TADISINA, S., "An exploratory study into factors of service quality for application service providers," Information & Management, vol. 42, pp. 1067-1080, 2005.
[16] ONIBOKUN, A. and PALANKAR, M., "Amazon S3: Black-Box Performance Evaluation."´

[17] PALANKAR, M.R., IAMNITCHI, A., RIPEANU, M., and GARFINKEL, S., "Amazon S3 for science grids: a viable solution?," in Proceedings of the 2008 international workshop on Data-aware distributed computing. Boston, MA, USA: ACM, 2008, pp. 55-64.

[18] SHAKKOTTAI, S.S., R., "Economics of network pricing with multiple ISPs," IEEE/ACM Trans. Netw., vol. 14, pp. 1233-1245, 2006.

[19] SHRIMALI, G.K., SUNIL, "Bill-and-Keep peering," Telecommunications Policy, vol. 32, pp. 19-32, 2008.

[20] TIROLE, J., The Theory of Industrial Organization: MIT Press, 1988.

[21] ULUDAG, S.L., K.S.; NAHRSTEDT, K. & BREWSTER, G., "Analysis of Topology Aggregation techniques for QoS routing," ACM Computing Surveys, vol. 39, 2007.

[22] WANG, Z., Internet QoS: Architectures and Mechanisms for Quality of Service: Morgan Kaufmann, 2001.

[23] ZENG, L.B., B.; NGU, A.H.H.; DUMAS, M.; KALAGNANAM, J. & CHANG, H., "QoS-aware middleware for Web services composition," IEEE Transactions on Software Engineering, vol. 30, pp. 311-327, 2004.

Price Setting in Two-sided Markets for Internet Connectivity

Thorsten Hau[1] and Walter Brenner[1]

[1]Institute of Information Management, University of St. Gallen
Müller-Friedberg-Strasse 8, CH-9000 St. Gallen, Switzerland
Phone +41 (0)71 224 3803 Fax +41 (0)71 224 3296
thorsten.hau@unisg.ch

Abstract.

Due to a lack of incentives, Internet peerings are a notorious bandwidth bottleneck. Through the use of direct interconnection and content delivery networks, content providers are able to provide better services to their customers. These technologies have a profound impact on the business models of internet service providers. Instead of competing for consumers and keeping uplink connection costs low, ISPs face a two-sided market in which they compete for EUs and generate revenues on the CP side of the market. This work presents a formal model for the providers' pricing decision towards content providers and discusses consequences for the Internet.

Keywords: Quality of Service, Network Economics, Peering, Internet

Introduction

The Internet is made up of many independent sub-networks – so called "autonomous systems" (AS). Generally speaking these ASs correspond to different carriers or Internet service providers (ISPs): firms that own and operate the infrastructure (cables and routers) that make up the Internet. These ISPs have customers who are either content providers (CP) with mostly outgoing traffic or end-users (EU) with mostly incoming traffic. To form the Internet, each ISP offers all of its customers connectivity to all of the other ISPs' customers. In order to uphold this universal connectivity, the ISPs have to exchange user traffic, an activity that is governed by contractual agreements between the ISPs and physically enabled by infrastructure that interconnects their networks. Thus, even though the Internet consists of many independent subnetworks, each user can reach every website on the web. However, it is also well known, that ISPs are not very cooperative in their peering behavior [7]. The decision to interconnect usually is more expensive for one party than for the other and therefore peerings tend to have smaller capacity than what would be optimal.

The rules regulating the exchange of traffic between ISPs have been subject to extensive treatment in the literature. Issues like hot potato routing [22, 24] and determination of access charges have been extensively studied and are quite well understood [8, 10, 17]. However, the available literature studies an idealized Internet in which there are EUs, CPs and ISPs that have relations as depicted in fig. 1. The Internet is modeled as a strictly hierarchical system in which traffic flows from a CP to its ISP, is exchanged with the requesting user's ISP and is then sent to the EU.

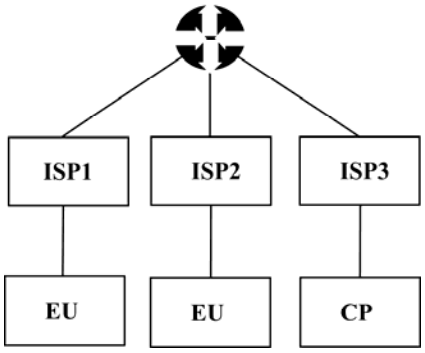


Fig. 1.

A key characteristic of this setup is that traffic is exchanged through a peering point. Depending on the contract between two ISPs this traffic exchange may happen in exchange for a payment or "for free". Due to lacking incentives to extend peering capacity sufficiently, these peerings represent major traffic bottlenecks [1, 7, 15].

In contrast to the available work [17, 18, 23], this paper focuses on two important variations of this idealized model of Internet infrastructure as shown in figures 2 and 3. The existing literature has ignored the possibility that content providers and terminating internet service providers interconnect directly.

There are two modes of "direct interconnection" that we will consider. Firstly, a content provider can directly buy transit from the terminating ISPs, thus effectively paying them for preferential access to end-users. This practice shown in fig. 2 is called mulit-homing (MH) and contributes to exponential growth of routing tables [5]. Secondly, content delivery networks (CDNs) shown in fig. 3 are a popular way to enhance the flow of information on the Internet. A CDN uses local caches to keep distributed images of content close to EUs without the need to traverse several ISPs' networks [26]. Both technologies provide viable means to improve the speed and reliability of data transport from a CP's website to EUs. They allow bypassing peerings and gaining more direct access to the EUs, thus increasing the probability of timely delivery of data to the end-user. The motivation to use CDN or MH is to provide better quality of service (QoS) with the following chain of causality: Traversing peerings degrades user experience by creating delays → QoS access to EUs through CDN and MH creates better user experience → more visitors to a website → higher revenues from selling ad space on the website.

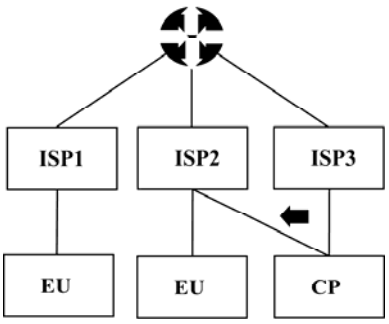


Fig. 2.

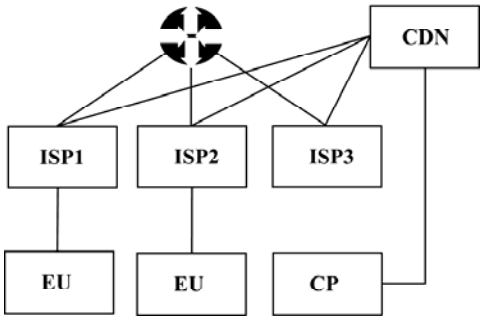


Fig. 3.

This paper uses the economic theory of two-sided markets to understand the pricing decision an ISP has to make with respect to the charges levied on the CP side of the market in settings such as those in figs. 2 and 3. Neither the Internet as a whole, nor individual internet service providers (ISPs) can straightforwardly be considered platforms that optimize their revenue from two sides of a market. With the standard Internet business model, each terminating ISP lacks the power to charge content providers that are signed up with another ISP. There are technical as well as contractual barriers to charge some remote content provider for single data packets it sends to an ISP's network. The access charge (the interconnection fee) exchanged between two ISPs is only an imperfect tool to exploit an access monopoly on the Internet due to the fact that it is often reciprocal or zero for external reasons [2, 13, 17]. This is a key difference between the Internet and telephone services (PSTN), where for each call, sender and receiver can be identified and billed per unit of time and a per unit settlement between providers is possible [4, 6, 10]. With CNDs or MH, the property of PSTN that the participating parties can be identified (and are billable customers) is recreated.

The CDN is a third party mediating between CP and ISP but the ISP can charge the CDN for delivery of traffic which will pass this cost on to its CPs. This situation is

different from the case of access charges between different providers (as analyzed by [17] with a focus on the Internet or [9] with a focus on PSTN) because in a CDN relationship there is no reciprocity or two way access which is an important condition for that model to be applicable. For the rest of this article we simplify the role that CDNs play on the Internet by treating them as pure mediators between atomic CPs and ISPs. They aggregate CP demand but do not engage in strategic actions. This simplification allows us to model the situation of fig.1 and fig. 2 in the same way. In the last section of this work we sketch a path to relaxing this rather strong assumption.

The paper is structured as follows: Firstly we review the relevant literature on two-sided markets and related topics in telecommunications pricing. Then we explain the abstracted situation we wish to understand and motivate our use of a two-sided market model. Thirdly, we present a formalized model for an ISP facing a two-sided market, deriving results from the market setup. We derive optimal prices charged by ISPs to CPs and CDNs that wish to directly interconnect with them. Lastly we summarize our findings and discuss implications and future research topics.

Literature Review

Armstrong's discussion of competition in two-sided markets [3] provides much of the foundation for this work. Two-sided markets are markets where a platform optimizes profit across two distinct sets of customers instead of just one. In the credit card industry, the card issuing company would be the platform and the merchants accepting the card constitute one group of customers while the buyers using the card to pay form the other. Armstrong analyzes three distinct settings with different customer behaviors and levels of platform competition. The situation relevant for this work is termed "competitive bottleneck": One group of customers can use any number of platform providers simultaneously, while the other group chooses only one of the competing platforms. In our problem, this situation corresponds to EUs being subscribed to only one single ISP while CPs can deal with any number of ISPs at the same time.

Rochet et al. [21] provide a comprehensive overview of the current literature on two-sided markets. They define two-sided markets as markets in which not only the total price but also the price structure influences the number of transactions on the market. For the case at hand, the ISP provides the platform on which transactions between EUs and CPs can take place. They also provide definitions for membership and usage externalities. In the first case one party profits from the sheer presence of the other, while a usage externality is a network effect that arises in an transaction between members of the two sides. They also discuss the effects of fixed and variable prices on the platform. Since variable prices reduce the externality exerted by one group of customers on the other, participation incentives are reduced.

Laffont et al. [17] are not directly concerned with two-sided markets. This work analyzes the access charge paid from one ISP to another for passing traffic on to that ISP's network. In their model the ISP optimizes the prices it charges to CPs and EUs subject to the access charges it pays (for sending traffic to an EU on another ISP's network) and receives (for terminating traffic with its own EUs). In their model the

access charge turns out to be a pure tool for reallocating termination costs between EUs and CPs. In the common case of zero access charges all termination costs are born by the EU which corresponds to a subsidy to CPs.

Musacchio et al. [19] compare the effects of single and multi-homing of CPs. They provide explicit formulations of welfare under both regimes and offers results for an economy with many ISPs. However, they do not model EU and CP demand separately but base their model on the assumption of click rates of EUs as a measure of demand for both customer groups and only differentiate CPs from EUs via the per-click price.

Problem Description

This work uses the theory on two-sided markets to explore two special cases of interconnection that are different from the symmetric and reciprocal case studied by [17]. The standard model of Internet traffic exchange as shown in fig. 1 follows the pattern $CP \rightarrow ISP_o \leftrightarrow_a ISP_t \leftarrow EU$ (t=terminating, o=originating, a=access charge) as shown in fig. 4. CP and EU pay their respective ISP and the ISPs exchange traffic for a fee *a*. This scheme ignores the source of the CP's funding and emphasizes the analysis of the inter-ISP settlement *a*, which has an influence on the prices paid to the ISPs. By contrast, this work focuses on the setup $EU \rightarrow ISP_t \leftarrow CP \leftarrow Adv.$ as shown in fig. 5. EUs derive utility from high quality of service (QoS) access to CPs' websites while the CPs generate profits from selling ad space to third parties. There is no monetary flow between CPs and EUs. Both, however, may exchange money with the ISP, which acts as a profit maximizing platform. This situation corresponds to the majority of today's Internet business models. CPs create websites that appeal to many EUs, thus generating page views that translate into value of ad-space on that site. (Figs. 1, 2 and 3 focus on physical interconnection while figs. 4 and 5 depict the business view on connection relationships.)

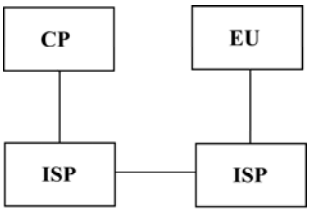


Fig. 4.

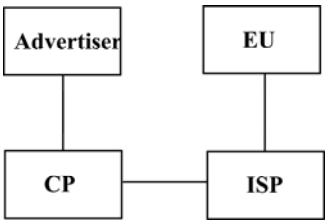


Fig. 5.

This CP business model has received wide attention in the two-sided markets literature as it corresponds to the business model of newspapers [3, 20, 21]. This work, however, does not consider the business model of the CP but that of the ISP. CPs pay a transaction-independent price for direct connection to EUs through buying bandwidth from the terminating ISP. EUs on the other hand pay a flat rate fee to the ISP to be connected to the Internet and no transaction based fee for viewing content.

There is no payment between EU and CP. The case with payments between the CPs and the EUs has been analyzed in [12].

In the sense of the two-sided market literature we have the following setup: Platform = ISP, single homing side = EU, multi-homing side = CP. The platform charges both sides a lump sum fee for facilitating transactions. This is more reasonable than a linear fee since for EUs, flat rates are the common pricing model and CPs commonly buy a certain bandwidth or a fixed traffic volume. Furthermore the price for Internet services delivered by an ISP might depend on the data volume but rarely on the value of a transaction. Therefore we assume that there is no linear payment that reduces the size of the externalities exerted on the other side of the market, respectively.

ISPs as Platforms in Two-sided Markets

The analysis in this section is related to the competitive bottleneck case of [3]. Competitive bottlenecks arise, when one firm has a monopoly over access to certain customers. Suppose there are two ISPs in a region denoted $\text{ISP}_i, i\epsilon\{1;2\}$. There are also two groups of agents. Group one agents are called end-users (EUs) while group two members are called content providers (CPs) or websites. There is a fraction n_j^i of agents of Group j participating on platform i . In other words, ISP_i has n_1^i subscribed EUs and n_2^i directly interconnecting customers from the CP side.

The setup is such that two ISPs are present in a market and serve two distinct groups of EUs with Internet connectivity. EUs are single-homing with their ISP. This means that they are only subscribed with one ISP at a given time. CPs on the other hand multi-home. They may be connected to zero, one or two ISPs in order to reach potential customers (EUs).

To analyze this situation we start by modeling the target function of two ISPs that compete in a market for EUs. The ISPs maximize their respective profits. Symbolically,

$$\pi^i = n_1^i p_1^i + n_2^i p_2^i - C^i(n_1^i, n_2^i), i \in \{1;2\} \tag{1}$$

which is a function of the number of EUs times the price they have to pay, plus the number of CPs times the price they have to pay minus the cost for connecting the two types of customers. The fraction $n_1^i \in [0,1]$ of EUs that are customers of ISP_i is given as a function of the utilities offered by the two ISPs:

$$n_1^i = \phi^i(u_{1,}^i u_1^j) = \frac{1}{2} + \left(\frac{u_1^i - u_1^j}{2t} \right), \forall i \neq j \, . \tag{2}$$

The function ϕ^i is thus increasing in the first argument and decreasing in the second. Note that $n_1^i + n_1^j = 1$ holds since EUs do not multi-home. To specification of EU-demand in equation (2), i.e. the fraction of EUs that are signed up with either ISP is described in a Hotelling [14, 25] way. This implies that the two ISPs share the market equally if they are undifferentiated from the consumers' point of view. If, however, one ISP offers superior utility, it can capture more than half of the market.

The utility EUs get from subscribing to ISP_i is given by

$$u_1^i = U^i(n_2^i) - p_1^i = \alpha_1 n_2^i - p_1^i \,. \tag{3}$$

It equals the gross utility they get from being connected with superior QoS to n_2^i directly interconnected CPs minus the price they have to pay for that connection. The function U^i is increasing in n_2^i since more content in better quality is always better than less. The parameter α_1 can be interpreted as the utility an EU derives from being able to reach one high QoS CP. The EUs perceive the ISP with more CPs connected with QoS as providing a better connection to the Internet.

The fraction $n_2^i \in [0,1]$ of CPs that is connected to ISP_i is given by

$$n_2^i = \phi_2^i(n_1^i, p_2^i) = 1 - \boldsymbol{F}(\gamma^i) \text{ with } \gamma^i = \frac{p_2^i}{n_1^i} \,. \tag{4}$$

It is a function of the number of EUs that can be reached through ISP_i and the price charged. The number of CPs the ISP can persuade to directly interconnect depends on the parameter $\gamma^i = p_2^i / n_1^i$. This parameter is calculated as the fraction of the fixed price for connectivity over number of reachable EUs. Thus it can be interpreted as the perceived price per EU. The distribution $\boldsymbol{F}(\gamma^i)$ then yields the fraction of CPs that are not willing to pay that price and $\boldsymbol{1 - F}$ yields the fraction of CPs that are willing to pay that price because their expected revenue per EU covers the expense.

The CPs do not deal exclusively with a single ISP but may be connected to zero, one or two ISPs, depending on their participation constraint being fulfilled. Therefore in general $n_2^i + n_2^j \neq 1$.

While equation (4) only depends on factors under control of ISP_i, equation (2) also depends on factors controlled by the other ISP. This reflects the fact that there is competition for EUs, but none for CPs.

Costs for interconnection are defined as

$$C^i(n_1^i, n_2^i) = c n_2^i \,. \tag{5}$$

This implicitly includes the assumption that the cost of the access network is not part of the considerations for interconnecting with CPs. This assumption is justified by the fact that access networks largely represent sunk costs.

Now, in order to solve the ISPs' optimization problem $\max \Pi^i$, assume that the platforms have reached an equilibrium and offer utility $\hat{u}_1^i$ to their $\hat{n}_1^i$ EUs, respectively. That is, we keep these values fixed while varying the others. This corresponds to today's situation in many markets for DSL or cable. There is some churn, but by and large networks operate in saturated markets with stable customer numbers. Since (4) defines n_2^i as a function of p_2^i, we can eliminate p_2^i and only have n_2^i left as a dependent variable. Thus, given an equilibrium $(\hat{u}_1^i, \hat{n}_1^i)$, we can solve for the optimal number of CPs n_2^i.

Rewriting equation (3) as $p_1^i = U^i(n_2^i) - u_1^i$ we can insert this expression into (1) to get

$$\begin{aligned} \Pi^i &= \hat{n}_1^i (U^i(n_2^i) - u_1^i) + p_2^i \overbrace{(1 - F(p_2^i / \hat{n}_1^i))}^{n_2^i} - C(\hat{n}_1^i, n_2^i) \\ &= \hat{n}_1^i (\alpha_1 n_2^i - \hat{u}_1^i) + \left(p_2^i - c\right) n_2^i \ . \end{aligned} \tag{6}$$

This expression shows that given an arbitrary equilibrium we can explicitly write the profit of the platform as a function of the price charged to its group two customers (i.e. CPs). The platform can thus easily calculate the optimal price and the resulting number of CPs, given its current competitive situation on the EU side of the market.

To give a concrete example, we define the distribution $\boldsymbol{F}$ and explicitly calculate the profit maximizing price p_2^i. Let the distribution function $\boldsymbol{F}$ be given by the probability density function $f(\gamma) = 1/\tau, \forall \gamma \in [0;\tau]$ of the uniform distribution. γ represents the expected revenue from ad-clicks per EU and τ represents the maximum price a CP is willing to pay for access to such an EU. The corresponding cumulated distribution function is

$$\boldsymbol{F} = \gamma / \tau = p_2^i / n_1^i \tau \ . \tag{7}$$

Any other distribution function would work as well. However, the normal distribution for example is not easily manipulated and thus would only allow a numerical solution to the problem at hand.

Now we insert (4) and (7) into (6)

$$\begin{aligned} \Pi^i &= \hat{n}_1^i (\alpha_1 n_2^i - \hat{u}_1^i) + (p_2^i - c) n_2^i \\ &= \hat{n}_1^i (\alpha_1 n_2^i - \hat{u}_1^i) + \left[(1 - n_2^i) \tau \hat{n}_1^i - c \right] n_2^i \end{aligned} \tag{8}$$

and find the maximizer of the resulting expression:

$$\frac{\partial \Pi^i}{\partial n_2^i} = \hat{n}_1^i \alpha_1 + (1 - 2n_2^i)\tau \hat{n}_1^i - c \overset{!}{=} 0$$
$$n_2^i = (1 - \frac{c - \hat{n}_1^i \alpha_1}{\tau \hat{n}_1^i}) \frac{1}{2} \ . \tag{9}$$

This is the optimal number of CPs the ISP should allow on its platform (since the 2nd order condition for a maximum holds). Together with (4) and (7) this yields the optimal price to CPs

$$p_2^i = \frac{1}{2}(c + \hat{n}_1^i \tau - \hat{n}_1^i \alpha_1) \ . \tag{10}$$

Therefore, CPs pay a price that is calculated on the basis of the cost they cause, increased by a factor relating to their per-EU-valuation and decreased by the externality they exert on the EUs. The factor $1/2$ should not be over-interpreted since it is an artifact of the definition of the distribution function in (7).

We thus have calculated the optimal number of CPs and the optimal price that an ISP should charge atomistic and ex-ante identical CPs for quality interconnection.

Conclusions and Further Research

To sum up, we have firstly explained two phenomena of the Internet that fundamentally change the way CPs and EUs are interconnected. CDNs and MH foster more direct links between these two user groups with only one mediating ISP instead of many. Employing the theory of two-sided markets we then went on to show how direct interconnection puts the ISP into a position to charge CPs directly. In the main section we showed how the optimal price p_2^i can be calculated for any given equilibrium on the EU side of the market.

While today it is uncommon to explicitly charge content providers for delivering traffic to their customers, there are clearly developments in the marketplace that can be understood in the above context. Google's effort to provide free W-Lan to customers in the US is only one example. Google wants to control the platform over which its content is delivered so that the profits it makes on the advertisement side cannot be extracted by ISPs.

To interpret the results obtained, lets first compare the predicted price to today's bill and keep regime. In today's peering agreements between ISPs, the fee for carrying traffic is very often zero. As [17] point out, this corresponds to a subsidy to CPs, since EUs carry most of the transmission cost. In our two-sided market framework on the other side, the CPs have to bear the cost they cause. They may be furthermore charged by the ISP, depending on their willingness to pay. This charging

is balanced by a "bonus" for the externality they exert on the ISP's EUs. Since the difference between being subsidized and paying bottleneck prices can be quite large, there will probably be a transitional period before ISPs can leverage their whole power in charging CPs. However, the presence of charges to content providers in itself does not represent a market failure. As long as ISPs are competing in the market for EUs, the profits they make on the CPs are used to compete in the EU market [3]. A waterbed effect might occur, but would merely be a sign of imperfect competition in the EU market [11].

Secondly, the last term in (10) illustrates a very interesting result. Imagine that the ISP could perfectly discriminate between two different groups of CPs. The group that exerts a higher externality on the customers through its presence would pay a lower price than the group with the lower externality effect. Thus, CPs that are very important to EUs will pay a low price to the bottleneck ISP, while those CPs, the presence of which is less valued by EUs, will pay a high price for access to EUs. Thus, a power balance could develop, in which CPs are charged by the network if they have low market power; or charge the network, if their content is highly desired by EUs.

Lastly, look again at the externality exerted by CPs. Here might lie an interesting option for future ISP business models. The ISP could try to capture some of the externality. This could happen for example through transaction dependent charges. Aside from contractual problems this would fulfill many ISPs' long standing vision to capture some of the profits of the content business. This development can already be witnessed in the mobile sector where Vodafone provides high quality ad-financed content to its customers.

An important aspect of this work that requires further research is the effect of the two-sided markets phenomena on the quality of standard peerings. As it stands today, peerings do not generate revenue for ISPs but only costs. With revenues from direct interconnection there is obviously a strong incentive for ISPs to move as many CPs as possible to a paying interconnection model. The ultimate consequence of this would be that, in order to foster a self selection process, standard peering quality would be considerably degraded to make sure that all customers with a willingness to pay are in the paying group. While such price discrimination is welfare enhancing, it is crucially important the market for EUs is competitive since otherwise, ISPs are in a position to appropriate rents.

This paper demonstrates the use of two-sided market theory to analyze the decision problems faced by Internet service providers in more complex setups then the standard peering scenario examined in earlier works. A first analysis demonstrates that new business models such as content delivery networks and multi-homing can fundamentally change the rules for interconnection pricing. This work thus extends the work on Internet interconnection [17] and the work on voice interconnection such as [16] or [2] (as well as the references cited therein).

As this is only a first step towards a thorough understanding of the new rules of interconnection pricing brought about by new interconnection regimes, there remains considerable work to be done:

Firstly, the presented analysis cuts short some more in depth equilibrium analysis by assuming a market equilibrium as given.

Furthermore, a more thorough analysis of the effects of the ISPs actions on the secondary markets for advertisements would be interesting. How do two vertically dependent two-sided markets interact?

In a similar line of thought, the aspect that CDNs are intermediaries between ISPs and CPs has been used as a starting point of the analysis but is then abstracted from in the further analysis. This can be justified by assuming that CDNs only pass on costs but their role certainly deserves more attention, especially since CDNs are potent players in the Internet market. A further topic to be analyzed is the role of peer to peer traffic.

The paper has shown an aspect of the quality of service debate that has been under-researched. The market for Internet interconnection has a considerable influence on the deliverable quality of Internet services. Understanding these markets (the contribution of this work) and "engineering" them to function better (future research) propose challenging research topics that might shape the next generation of networks.

References

1. Akella, A., Seshan, S. and Shaikh, A., "An empirical evaluation of wide-area Internet bottlenecks," presented at Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement, Miami Beach, FL, USA (2003)
2. Armstrong, M., "Network Interconnection in Telecommunications," The Economic Journal, vol. 108, pp. 545-564 (1998)
3. Armstrong, M., "Competition in two-sided markets.," RAND Journal of Economics, vol. 37, pp. 668-691 (2006)
4. Berger, U., "Bill-and-keep vs. cost-based access pricing revisited," Economics Letters, vol. 86, pp. 107-112 (2005)
5. Bu, T., Gao, L. and Towsley, D., "On characterizing BGP routing table growth," Computer Networks, vol. 45, pp. 45-54 (2004)
6. Cambini, C. and Valletti, T. M., "Network competition with price discrimination: `bill-and-keep' is not so bad after all," Economics Letters, vol. 81, pp. 205-213 (2003)
7. Cremer, J. R., Patrick & Tirole, Jean, "Connectivity in the Commercial Internet," The Journal of Industrial Economics, vol. 48, pp. 433-472 (2000)
8. DeGraba, P., "Reconciling the off-net cost pricing principle with efficient network utilization," Information Economics and Policy, vol. 16, pp. 475-494 (2004)
9. Economides, N., Lopomo, G. and Woroch, G., "Strategic Commitments and the Principle of Reciprocity in Interconnection Pricing," Stern School Business, NYU (2008)
10. Gans, J. S. and King, S. P., "Using bill and keep interconnect arrangements to soften network competition," Economics Letters, vol. 71, pp. 413-420 (2001)
11. Genakos, C. and Valletti, T., "Testing the'Waterbed'Effect in Mobile Telephony," CEIS Working Paper No. 110 (2008)
12. Hau, T., Wulf, J., Zarnekow, R. and Brenner, W., "Economic Effects of Mult Homing and Content Delivery Networks on the Internet," Proceedings of the 19th ITS European Regional Conference in Rome (2008)
13. Hermalin, B. E. and Katz, M. L., "Sender or Receiver: Who Should Pay to Exchange an Electronic Message?," The RAND Journal of Economics, vol. 35, pp. 423-448 (2004)
14. Hotelling, H., "Stability in Competition," Economic Journal, vol. 39, pp. 41-57 (1929)
15. Kushman, N. K., Srikanth & Katabi, Dina, "Can you hear me now?!: it must be BGP," SIGCOMM Comput. Commun. Rev., vol. 37, pp. 75-84 (2007)

16. Laffont, J.-J., Rey, P. and Tirole, J., "Network Competition: I. Overview and Nondiscriminatory Pricing," The RAND Journal of Economics, vol. 29, pp. 1-37 (1998)
17. Laffont, J.-J., Marcus, S., Rey, P. and Tirole, J., "Internet Interconnection and the Off-Net-Cost Pricing Principle," The RAND Journal of Economics, vol. 34, pp. 370-390 (2003)
18. Laffont, J. J. R., P. & Tirole, J., "Competition between telecommunications operators," European Economic Review, vol. 41, pp. 701-711 (1997)
19. Musacchio, J., Walrand, J. and Schwartz, G., "Network Neutrality and Provider Investment Incentives," presented at Signals, Systems and Computers, 2007. ACSSC 2007. Conference Record of the Forty-First Asilomar Conference on (2007)
20. Rochet, J. C. T., J., "Platform Competition in Two-Sided Markets," Journal of the European Economic Association, vol. 1, pp. 990-1029 (2003)
21. Rochet, J.-C. T., Jean, "Two-sided markets: a progress report," The RAND Journal of Economics, vol. 37, pp. 645-667 (2006)
22. Roughgarden, T., Selfish Routing and the Price of Anarchy: MIT Press, (2005)
23. Shakkottai, S. S., R., "Economics of network pricing with multiple ISPs," IEEE/ACM Trans. Netw., vol. 14, pp. 1233-1245 (2006)
24. Teixeira, R., Shaikh, A., Griffin, T. and Rexford, J., "Dynamics of hot-potato routing in IP networks," presented at Proceedings of the joint international conference on Measurement and modeling of computer systems (2004)
25. Tirole, J., The Theory of Industrial Organization: MIT Press, (1988)
26. Vakali, A. and Pallis, G., "Content delivery networks: status and trends," Internet Computing, IEEE, vol. 7, pp. 68-74 (2003)

Lebenslauf

Persönliche Angaben

05. Dezember 1980	Geboren in Gießen, Deutschland

Ausbildung

1991 – 2000	*Kepler-Gymnasium Freudenstadt* Abitur
1997 – 1998	*St. John De Brebeuf Secondary School, Kanada* Austauschjahr
2001 – 2006	*Universität Karlsruhe* Diplom Wirtschaftsingenieur
2004 – 2005	*Ecole Polytechnique Fédérale de Lausanne* Erasmus Jahr
2007 – 2009	*Universität St. Gallen* Doktoratsstudium in Wirtschaftswissenschaften

Praktische Tätigkeiten

2001	*Daimler Chrysler Sindelfingen* Praktikum (Balanced Scorecard im CKD)
2002	*Intarsys Consulting* Praktikum (Softwareentwicklung Java)
2005	*Daimler-Chrysler Südafrika* Praktikum (Prozessdesign, Change Management)