Universität Hamburg

Fachbereich Informatik

Arbeitsbereich Angewandte und Sozialorientierte Informatik

Diplomarbeit
im Fach Wirtschaftsinformatik

Systemantwortzeiten als Aspekt der Software-Ergonomie und der Wirtschaftsinformatik

Marco Glier

Datum der Abgabe: 19. November 2005

Erstbetreuer: Prof. Dr. Horst Oberquelle

Zweitbetreuer: Dr. Guido Gryczan

Danksagung

Das Schreiben ist meist ein einsamer Prozess, aber es gibt vielen Menschen, die ihn begleitet und unterstützt haben und denen ich hier danken möchte.

Allen voran danke ich Prof. Dr. Oberquelle für die Betreuung meiner Diplomarbeit sowie Dr. Guido Gryczan für die Übernahme der Zweitbetreuung.

Ich danke Petra Vogt und Dr. Herbert A. Meyer für die vielen Diskussionen und den Gedankenaustausch, aus denen parallel zu dieser Arbeit ein gemeinsamer Konferenzbeitrag auf der *Mensch und Computer 2005* und ein Beitrag für die Zeitschrift *icom* entstand.

Ich bedanke mich bei Yvonne Runge, dass sie die, für sie völlig fachfremde Arbeit Korrektur gelesen hat und kritische Bemerkungen beisteuerte.

Bei meiner Freundin Jacqueline möchte ich mich für ihre Geduld, Unterstützung, Diskussionen und den Korrekturen zu dieser Arbeit bedanken.

Zu guter Letzt möchte ich mich bei meinen Eltern und Großmutter bedanken, die mich während meines Studiums unterstützten, in mich vertrauten und mein Studium dadurch ermöglicht haben.

Inhaltsverzeichnis

Abbildungsverzeichnis VI

Tabellenverzeichnis VII

Abkürzungsverzeichnis VIII

1 Einleitung 1

2 Entwicklung der Computer- und Informationstechnologie 5

2.1 Computertechnologie 5

2.2 Netzwerke 7

3 Wissenschaftsstandpunkte 8

3.1 Software-Ergonomie 8

3.2 Wirtschaftsinformatik 10

3.3 Vergleich beider Disziplinen 12

4 Zeitverhalten interaktiver Systeme 15

4.1 Definition der Systemantwortzeiten 15

4.2 Psychologische und physiologische Aspekte 20

4.2.1 Kognitive Leistung 20

4.2.2 Beanspruchung und Belastung 22

4.2.3 Arbeitsplatztypen 24

4.2.4 Stress am Bildschirmarbeitsplatz 26

4.3 Soziographische Aspekte 29

4.3.1 Benutzergruppen 29

4.3.2 Alter 29

4.3.3 Besondere Anforderungen 30

4.3.4 Erfahrungen der Benutzer 31

4.4 Normen ... 32
4.4.1 Grundlage ... 32
4.4.2 DIN EN ISO 9241-11 ... 33
4.4.3 DIN EN ISO 9241-10 ... 34
4.4.4 ISO/IEC 9126 ... 36
4.4.5 Weitere Normen ... 37

5 Systemantwortzeiten von Anwendungssystemen ... 39

5.1 Grundlagen und Modellierungsaspekte ... 39
5.1.1 Definition ... 39
5.1.2 Qualitätskriterien ... 41
5.1.3 Leistungskenngrößen und physikalische Eigenschaften ... 45
5.1.4 Leistungen der Netzknoten ... 49
5.1.5 Netzauslastung ... 52

5.2 Einzelsystem ... 53
5.2.1 Hardware ... 53
5.2.2 Software ... 56

5.3 Verteilte Systeme ... 58
5.3.1 Client-Server-Architektur ... 58
5.3.2 Lokales Netz ... 59
5.3.3 Lokale Funknetzwerke ... 60
5.3.4 Weitverkehrsnetze ... 61
5.3.5 Mobile Systeme ... 62

5.4 Internet ... 66
5.4.1 Bedeutung des Webs ... 66
5.4.2 Technische Infrastruktur ... 67
5.4.3 Dienstgüte bei Webservices ... 70
5.4.4 Ansätze zur technische Optimierung ... 71
5.4.5 Benutzersicht ... 73

6 Wirtschaftliche Aspekte der Systemantwortzeiten 75

6.1 Wirtschaftlichkeit von Informationssystemen 75

6.1.1 Bedeutung der Informationssysteme 75

6.1.2 Produktivitätsparadoxon 75

6.1.3 Wirtschaftlichkeitsvergleich 76

6.2 Verfahren der Investitionsrechnung 79

6.2.1 Investitionsrechnung als Entscheidungsgrundlage 79

6.2.2 Statische Verfahren 79

6.2.3 Dynamische Verfahren 81

6.2.4 Beschränkung der Investitionsverfahren 82

6.3 Bewertungsmethoden 83

6.3.1 Return of Investment (ROI) 83

6.3.2 Total cost of ownership (TCO) 86

6.3.3 Implikation für Systemantwortzeiten 89

7 Schlussbetrachtung 90

7.1 Zusammenfassung 90

7.2 Fazit 91

7.3 Ausblick 93

Literaturverzeichnis 96

Erklärung 114

Abbildungsverzeichnis

Abbildung 1 Aspekte der Software-Ergonomie 8
Abbildung 2 Modifizierte Leavitt-Raute 9
Abbildung 3 Stellung der Wirtschaftsinformatik 11
Abbildung 4 Schwerpunkte der Wirtschaftsinformatik 14
Abbildung 5 Standardabweichung der Antwortzeiten 16
Abbildung 6 Einfaches Antwortzeitmodel 17
Abbildung 7 Erweitertes Antwortzeitmodel 17
Abbildung 8 Zeitmodell der Mensch-Computer-Interaktion 18
Abbildung 9 Gedächtnisse und Prozessoren 20
Abbildung 10 Erinnerungskurve des Kurzzeitgedächtnisses 21
Abbildung 11 Belastungs-Beanspruchungsmodell 22
Abbildung 12 Beanspruchte Fähigkeiten 23
Abbildung 13 Aufbau der Vorschriften zur Software-Ergonomie 32
Abbildung 14 Netzklassen 40
Abbildung 15 Skalierbarkeit von Anwendungssystemen 43
Abbildung 16 Verzögerungszusammensetzung 47
Abbildung 17 Grundlegendes Bedienmodell 50
Abbildung 18 Rechnergrundstruktur 53
Abbildung 19 CPU Auslastung und Antwortzeit 54
Abbildung 20 Systemaufbau 56
Abbildung 21 Allgemeine Client-Server-Kommunikation 58
Abbildung 22 Systemantwortzeit im Verhältnis zur Benutzerantwortzeit 60
Abbildung 23 Entwicklung der Mobilfunkstandards 62
Abbildung 24 Wachstumsentwicklung des Internets 66
Abbildung 25 Einfaches Web-Modell 67
Abbildung 26 Aufschlüsselung der Systemantwortzeit 67
Abbildung 27 HTTP Transaktion 68
Abbildung 28 Systemantwortzeitkomponenten von Webseiten 69
Abbildung 29 Kosten-Nutzen-Vergleich 77
Abbildung 30 Systemantwortzeiten und Fehleranfälligkeit der Benutzer 85
Abbildung 31 Lebenszyklusphasen einer IT-Anwendung 86

Tabellenverzeichnis

Tabelle 1 Gegenüberstellung Software-Ergonomie und Wirtschaftsinformatik 13
Tabelle 2 Beschreibung von Stressbedingungen 28
Tabelle 3 Verfügbarkeit und resultierende Ausfallzeiten 41
Tabelle 4 Benutzerbezogene Leistungskenngrößen 45
Tabelle 5 Eigenschaften von Rechnernetzen 46
Tabelle 6 Modelle der parallelen Verarbeitung 55
Tabelle 7 Systemantwortzeiten bei GPRS und EGPRS 64
Tabelle 8 Leistungserwartung der Endbenutzer 65
Tabelle 9 Dienstgüteparameter von Webservices 70
Tabelle 10 Systemantwortzeiten von Webseiten 74
Tabelle 11 TCO Model Distributed Computing Chart of Accounts 87

Abkürzungsverzeichnis

3GPP	3rd Generation Partnership Project
ABC	Anwender Benutzer Computer
AMR	Adaptive Multi-Rate
ANSI	American National Standards Institute
ARPA	Advanced Research Projects Agency
ASP	Application Service Provider
ATM	Asynchronous Transfer Mode
bps	bit per second
CDMA	Code Division Multiple Access
CEN	Comité Européen de Normalisation
CICS	Customer Information Control System
CPU	Central Processing Unit
CSMA/CD	Carrier Sense Multiple Access / Collision Detection
DIN	Deutsches Institut für Normen
DV	Datenverarbeitung
EDGE	Enhanced Date Rates for GSM Evolution
EGPRS	Enhanced General Packet Radio System
EN	Europäische Norm
FCFS	First Come First Serve
FTP	File Transfer Protocol
Gbps	Gigabit per second
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications
HDSPA	High Speed Downlinks Packet Access
HTTP	Hypertext Transfer Protocol
IAS	International Accounting Standards
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
Internet	Interconnected Networld
IP	Internet Protocol

IrDA	Infrared Data Association
ISC	Internet Systems Consortium
ISO	International Standardisation Organisation
ISP	Internet Service Provider
IT	Informationstechnologie
IuK	Information und Kommunikation
LAN	Local Area Network
MAN	Metropolitan Area Network
Mbps	Megabit per second
MCI	Mensch-Computer-Interaktion
MMS	Multimedia Messaging Service
NPV	Net Present Value (Nettobarwert)
OSI	Open Systems Interconnect
PC	Personal Computer
PDA	Personal Digital Assistant
QoS	Quality of Service
ROI	Return of Investment
SAZ	Systemantwortzeit
SLA	Service Level Agreement
SMTP	Simple Mail Transfer Protocol
SPE	Software Performance Engineering
TCO	Total Cost of Ownership
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunikation System
WAM	Werkzeug Automat Material
WAN	Wide Area Network
WAP	Wireless Application Protocol
WiFi	Wireless Fidelity
WiMax	Worldwide Interoperability for Microwave Access
WKWI	Wissenschaftliche Kommission Wirtschaftsinformatik
WWW	World Wide Web

1 Einleitung

Die Entwicklung der Computertechnik hat in den letzten sechzig Jahren einen stetigen Fortschritt erfahren. Waren es in den Anfängen um 1945 elektromechanische Rechengeräte, so entwickelten sich Großrechenanlagen zu den heute bekannten Personal Computern (PC) und den Rechnernetzen.[1] Insbesondere durch „die zunehmende Bedeutung weltweiter Rechnernetze (wie des Internets) [...] erwachsen neue Möglichkeiten“ (Fink, Schneidereit & Voss, 2000, S. 8), so dass wir heute von *Informationssystemen*[2] sprechen. Informationssysteme werden als soziotechnische Systeme charakterisiert, die sowohl die menschlichen als auch die maschinellen Komponenten umfassen, „die voneinander abhängig sind, ineinandergreifen und/oder zusammenwirken“ (WKWI[3], 1994, S. 80). Dieses Zusammenwirken muss durch Schnittstellen realisiert werden.

Die technologischen Entwicklungen der Informationssysteme zeichnen sich durch eine stetig zunehmende Miniaturisierung und gleichzeitige Leistungssteigerung der Systeme aus, die sich auch auf die Benutzungsschnittstellen auswirken. Die ursprüngliche Fokussierung der Software-Ergonomie auf die Unterstützung der menschlichen Arbeit durch die Computer weitet sich mittlerweile zu einem allumfassenden und ubiquitären Bereich aus. Als Gründe hierfür sind, wie Herczeg (2005, S. V) bemerkt, die Veränderungen in der Gesellschaft und die starke Überlappung von Arbeit, Bildung und Freizeit zu sehen. Die Informationssysteme wurden immer komplexer und heterogener. Allerdings wurden bei der Entwicklung leistungsstarker Systeme die Erforschung des Systemantwortzeitverhaltens der Informationssysteme und ihre Auswirkungen auf die Benutzer als ein Randaspekt weitestgehend vernachlässigt. Das Systemantwortzeitverhalten ist allerdings von entscheidender Bedeutung, da es den Benutzer in seinem Interaktionsverhalten direkt beeinflusst und sich damit auch auf den betriebswirtschaftlichen Bereich auswirkt.

---

1 Eine umfassende Übersicht bietet Ceruzzi (2003)

2 Informationssysteme werden hier als Synonym zu Informations- und Kommunikationssystemen betrachtet.

3 Wissenschaftliche Kommission Wirtschaftsinformatik

Erste Forschungsansätze zu den Systemantwortzeiten lassen sich bei Miller (1968) finden, die von Shneiderman (1984) aufgegriffen und vertieft wurden. Sie liegen allerdings schon weit zurück und beschäftigen sich eher allgemein mit den Antwortzeiten in Bezug auf Einzelsysteme. Durch die technologische Entwicklung und die schon erwähnte verstärkte Nutzung verteilter Systeme in weltweiten Rechnernetzen im wissenschaftlichen und betrieblichen Bereich muss der Fokus hinsichtlich der Systemantwortzeiten verstärkt auf die der vernetzten Systeme gerichtet werden.

Lange Systemantwortzeiten behindern den Arbeitsfortschritt in der Mensch-Computer-Interaktion (MCI) und sorgen für Frustration und Verärgerung bei den Nutzern (vgl. Shneiderman, 1998, S. 411). Dies führt zu einer Beeinträchtigung in der Effektivität und Effizienz der Arbeit und wirkt sich damit negativ auf die Zufriedenheit der Benutzer aus. Effektivität, Effizienz und Zufriedenstellung sind klassische Ansatzpunkte der Software-Ergonomie, sich mit der Zusammenwirkung von Reaktionszeiten zwischen Benutzern und Systemen zu beschäftigen.

Ferner gilt es zu beachten, dass durch die Beeinträchtigungen, die durch zu lange Systemantwortzeiten erzeugt werden, Opportunitätskosten entstehen, die es im betriebswirtschaftlichen Sinne zu minimieren gilt. Somit sollten die Kosten der Unbenutzbarkeit ein Thema der Wirtschaftsinformatik sein (vgl. Oberquelle, 2000, S. 4 ff.). Gegenstand der Wirtschaftsinformatik sind die Informationssysteme in Wirtschaft und Verwaltung (vgl. WKWI, 1994, S. 80). Diese sieht sich interdisziplinär „an der Schnittstelle zwischen der Betriebswirtschaftslehre und der (angewandten) Informatik“ (Fink et al., 2000, S. 1). Es gilt die Systemantwortzeiten der Informationssysteme dahingehend zu optimieren, dass sie dem Benutzer in seiner Arbeit nicht negativ beeinflussen und keine unnötigen Kosten verursachen. Durch eine höhere Zufriedenstellung der Benutzer wird eine bessere Effektivität realisiert, die die Kosten minimiert und dadurch höhere Umsätze und Gewinne in der Unternehmung ermöglicht. Dies führt zu einer höheren Effizienz und damit zu Wettbewerbsvorteilen. Es stellt sich somit die wichtige Frage, was ein optimales Systemantwortzeitverhalten sowohl im Sinne der Software-Ergonomie als auch der Wirtschaftsinformatik charakterisiert.

Das Systemantwortzeitverhalten an sich hängt von verschiedenen Faktoren ab. Es sind zum einen die technischen Systeme zu betrachten, die in ihrem heterogenen Aufbau Ansatzpunkte zur Optimierung bieten. Leistungskennzahlen der Anwendungsprogramme, der Prozessorleistung und der Netzleistungskapazität sind hier exemplarisch aufzuzählen. Zum anderen sind die Benutzer zu betrachten, die mit unterschiedlichen Erwartungshaltungen die Systeme nutzen wollen. Hier gilt es zu bedenken, welches Vorwissen entgegengebracht wird, in welchem Kontext die Systeme genutzt werden und wie die persönlichen Empfindungen einzuordnen sind.

Würde der technischen Leistungssteigerung der Vergangenheit entsprochen, müsste man eine gewagte These aufstellen und fragen, ob die Benutzer damals – vor zehn oder zwanzig Jahren – ineffektiver gearbeitet hätten als heute. Machen schnellere Systeme die menschliche Arbeit wirklich schneller? Ist ein anzustrebendes zeitliches Minimum mit dem Optimum gleichzusetzen? Wie weit lässt sich noch an der Zeit- und Geschwindigkeitsschraube drehen? Wann ist ein System zu schnell oder zu langsam und lässt den Benutzern keine Handlungsmöglichkeiten bzw. erhöht die Fehleranfälligkeit und damit die Kosten der (Un-)Benutzbarkeit?

Ziel dieser Arbeit ist es, die Systemantwortzeiten als Aspekt der Software-Ergonomie und der Wirtschaftsinformatik im Bereich der vernetzten Systeme – speziell des Internets – darzustellen und Ansatzpunkte zur Optimierung aufzuzeigen.

In den nachfolgenden beiden Kapiteln wird die technische Entwicklung aufgezeigt und die wissenschaftliche Ausgangsbasis geschaffen, um sich in den folgenden Kapiteln detailliert mit dem Zeitverhalten interaktiver Systeme – den so genannten Systemantwortzeiten – zu beschäftigen. Hierzu wird im zweiten Kapitel ein kurzer historischer Abriss über die Entwicklung der Computer- und Informationssysteme gegeben, um dem Leser deren rasante Entwicklung in den letzten sechzig Jahren vor Augen zu führen. Da die Systemantwortzeit sowohl ein Aspekt der Software-Ergonomie als auch der Wirtschaftsinformatik darstellt, werden im dritten Kapitel die jeweiligen

Wissenschaftsstandpunkte und die Verknüpfungspunkte zwischen den beiden Disziplinen aufgezeigt.
Die Kapitel vier und fünf beschäftigen sich mit den Aspekten der Systemantwortzeiten. Dieses sind die Verhaltensauswirkungen auf die Benutzer und die technologische Gestaltung der Systeme. Zunächst wird im vierten Kapitel auf die grundlegenden psychologischen, physiologischen und soziographischen Aspekte des Zeitverhaltens mit interaktiven Systemen eingegangen. Abgerundet wird es mit einer Betrachtung von relevanten Normen und Empfehlungen und deren Bezug zu den Systemantwortzeiten. Im fünften Kapitel werden die verschiedenen Informationssysteme mit ihren Komponenten und Ansatzpunkte zur Optimierung von Systemantwortzeiten detailliert beschrieben. Hierzu erfolgt zunächst eine Analyse von Einzelsystemen, die dann auf verteilte Systeme mit den verschiedenen Netzwerktypen und den mobilen Systemen erweitert wird. Das Kapitel wird abgerundet mit einer Betrachtung des World Wide Webs (WWW) und der technischen Infrastruktur des Internets.

Das sechste Kapitel beschäftigt sich mit den wirtschaftlichen Aspekten der Systemantwortzeiten bei der Nutzung von Informationssystemen. Dies im Hinblick auf die Effektivität und Effizienz der Unternehmungen und die dadurch entstehenden Kosten, die es zu minimieren gilt. Hierzu werden investitionstheoretische Methoden untersucht. Zwei Ansätzen von Berechnungsverfahren für die Systemantwortzeiten werden vertieft vorgestellt.

Die Arbeit schließt im siebten Kapitel mit einer Schlussbetrachtung. Diese besteht aus einer Zusammenfassung der Arbeit, einem Fazit mit der Darstellung der gewonnenen Erkenntnisse, sowie einem Ausblick für weitere Forschungsmöglichkeiten.

2 Entwicklung der Computer- und Informationstechnologie

2.1 Computertechnologie

Die Entwicklung der Computertechnologie lässt sich sicher in ihrer Urform auf die Unterstützung beim Rechnen zurückführen. In diesem Kapitel soll ein Überblick über die moderne Entwicklung der Informations- und Kommunikationstechnik und deren Motivation gegeben werden, um Ansatzpunkte für die Thematik dieser Arbeit aufzuzeigen, die in späteren Kapiteln wieder aufgegriffen werden.

Die eigentliche Erfindung des *Computers* fand Mitte bis Ende der 1940er Jahre statt. Weizenbaum (1984, S. 16 f.) nennt es einen Simultanfall, weil der Computer unabhängig voneinander an drei Orten entwickelt wurde: In Deutschland entwickelte Zuse und in Amerika Eckert und Mauchly einen Computer. Dagegen waren es in England mehrere Erfinder, unter anderem auch Alan Turing. Während Zuse von der persönlichen Motivation getrieben wurde, statische Berechnungen schneller durch eine Maschine selbständig durchführen zu lassen (vgl. Mons, 2000, S. 22 ff.), waren es bei den anderen beiden Teams das Militär, das Forschungsgelder bereitstellte. Dieser Beginn der informationstechnischen Entwicklung wurde von zwei wesentlichen Aspekten vorangetrieben: einerseits die Steigerung der Geschwindigkeit der Computer und andererseits eine zunehmende Miniaturisierung und Vernetzung.

Schon Licklider (1960, S. 6) wies darauf hin, dass die Benutzer über achtzig Prozent ihrer Zeit auf Ergebnisse der Stapelverarbeitung des Computers warteten. Somit war es die Antwortzeit des Computers, die die Benutzer in ihrem Arbeitsprozess signifikant beeinflusste. Um die Arbeitsressource des Computers – zu der Zeit waren es noch raumfüllende Großrechneranlagen – besser für mehr Benutzer auszunutzen, wurde das *Timesharing-Konzept* entwickelt.

Beim Timesharing handelt es sich um eine Betriebssystemeigenschaft, bei der sich mehrere Prozesse den gleichen Prozessor und Hauptspeicher teilen. Jeder Prozess

bekommt eine bestimmte Zeit zugewiesen, in der er die Systemressourcen verwenden darf. Diese Zeit muss dahingehend optimiert werden, dass das Antwortzeitverhalten des Systems und die Administrationszeit zur tatsächlichen Rechenzeit im sinnvollen Verhältnis zueinander stehen (vgl. Unland, 2001, S. 475).

Der andere wesentliche Faktor, der die Entwicklung der Computer- und Informationstechnologie signifikant beeinflusst hat, sind die hohen Budgets, die durch das Militär zur Forschung bereitgestellt wurden. Hierdurch wurde es ermöglicht, die Miniaturisierung der Computer stark zu forcieren (vgl. Weizenbaum, 1993, S. 31 f.), wodurch die wesentlichen Grundlagen für die heute selbstverständliche weltweite Vernetzung gelegt wurden (vgl. Abschnitt 2.2).

Mit der Entwicklung der Computer von elektromechanischen Rechengeräten zu Großrechneranlagen fand auch eine stetige Erweiterung der Benutzergruppe statt. Waren es in den Anfängen hauptsächlich Programmierer und Entwickler, die für sich selbst als Benutzer Programme und Systeme entwickelten, erweiterte sich der Benutzerkreis auf Personen ohne DV-Ausbildung. Dies wurde ferner durch die Entwicklung der Mikroprozessortechnik in den 1970er Jahren begünstigt. Die Folge war eine Diskrepanz zwischen dem technischen Verständnis der Entwickler einerseits und dem nur marginal technischen aber hohen fachlichen Wissen der neuen Benutzer andererseits. Somit kamen Fragen der Gestaltung von Bildschirmarbeitsplatzsystemen, insbesondere der Software-Ergonomie, auf, die die Konstruktion gebrauchstauglicher Systeme für alle Benutzer zum Ziel hatte (vgl. Abschnitt 3.1).

Die technische Entwicklung schritt immer schneller voran (vgl. Moores Gesetz, Moore, 1965, S. 114 ff.) und damit auch eine stetige Durchsetzung der Computer- und Informationstechnologie im betrieblichen Umfeld. Durch die Entwicklung von zentralen Großrechneranlagen zu dezentralen Arbeitsplatzrechnern – so genannten Personal Computern (PC) – kamen Fragen einer sinnvollen Art von Kommunikation, Kooperation und eines effektiven Datenaustausches zwischen den beteiligten Benutzern auf. Eine Vernetzung der Computer untereinander ermöglichte dies. Hierauf wird im nachfolgenden Abschnitt eingegangen.

2.2 Netzwerke

Als ein wichtiger Meilenstein der Computer- und Informationstechnologie ist der Aufsatz „As we may think“ von Vannevar Bush (1945, S. 101) zu betrachten, der das Problem der Erfassung und Verarbeitung von Informationen thematisierte und ein fiktives Gerät Namens Memex skizzierte, welches allerdings nie gebaut wurde.

Ähnlich wie Vannaver Bush in den 1940er Jahren hatte Licklider (1960) seine Vision einer Vernetzung und Kommunikation aller Computer untereinander gehabt. Hier wird der anfänglich erwähnte starke Bezug zum Militär deutlich, weil Licklider als Leiter in der vom US-Militär finanzierten Advanced Research Projects Agency (ARPA) beschäftigt war. Als Ergebnis wurde 1969 das ARPANET geschaffen, mit dem über entfernte Rechner Nachrichten ausgetauscht werden konnten. Neben dem ARPANET entwickelten sich noch eine Reihe weiterer heterogener regionaler Netze, wie z.B. das auf Funksignalen basierende Alohanet auf Hawaii von Norman Abramson. Ferner wurde für die lokale Vernetzung in Büros und Gebäuden von Robert Metcalf und David Boggs 1973 das *Ethernet* entwickelt - auch Local Area Network (LAN) genannt. Als Kommunikationsprotokoll wurde zur gleichen Zeit von Robert Kahn und Vinton Cerf das Transmission Control Protocol/Internet Protocol (*TCP/IP)* entwickelt, welches eine Kommunikation zwischen unterschiedlichen Computerplattformen und allen existierenden Netzwerken ermöglichte.

Im Laufe der Zeit entwickelte sich aus dem ARPANET das heutige *Internet* (Interconnected Networks), das seine wahre Popularität erst mit dem 1990 von Timothy Berners-Lee entwickelten World Wide Web (WWW, kurz: Web) erlangte. Das Web, basierend auf der Grundidee von Vannevar Bush, war ursprünglich für den Informationsaustausch für Wissenschaftler gedacht. Es wurde dann 1991 für die kommerzielle Nutzung geöffnet (vgl. Ceruzzi, 2003, S. 340-355; Matis, 2002, S. 303-319).

Durch die starke Verbreitung der Computer- und Informationstechnologie wurde die vorher auf den betrieblichen Kontext begrenzte Gruppe von Computerbenutzern erweitert, so dass nunmehr jeder Einzelne als Benutzer von Informations- und Kommunikationssystemen zu betrachten ist.

3 Wissenschaftsstandpunkte

3.1 Software-Ergonomie

Die Thematik der Benutzbarkeit von Software kam in den 1970er Jahren auf, als durch die Mikrocomputer eine immer größer werdende Gruppe von Benutzern ohne detaillierte IT-Kenntnisse Zugang zum Rechner fand. In der Anfangszeit waren es die Entwickler, die für sich selbst Softwareprogramme schrieben. Dies wandelte sich bis zur heutigen Zeit, in der keine fundierten IT-Kenntnisse mehr erforderlich sind, um Computer zu benutzen.

Einen ersten englischsprachlichen Übersichtsband zu der Thematik der Mensch-Computer-Interaktion veröffentlichte Martin (1973). Er stellte die Forderung auf, dass der Mensch im Fokus der Systementwicklung stehen muss und der Computer den Menschen bei seiner Arbeit unterstützen soll. Somit muss eine vernünftige Basis der Kommunikation zwischen Mensch und Computer geschaffen werden (Martin, 1973, S. 3 ff.). Im deutschsprachigen Raum wurde durch Dehning, Essig & Maass (1978) eine erste Bestandsaufnahme erhoben. Der Begriff der Software-Ergonomie wurde durch Griese (1982, S. 124) als „Anpassung der Software an den Menschen“ vorgeschlagen. Das Kunstwort Ergonomie wurde 1949 von Wissenschaftlern um Murell aus den griechischen Wörtern *ergon* für Arbeit und *nomos* für Gesetzmäßigkeit gebildet (Bubb, 1993, S. 194).

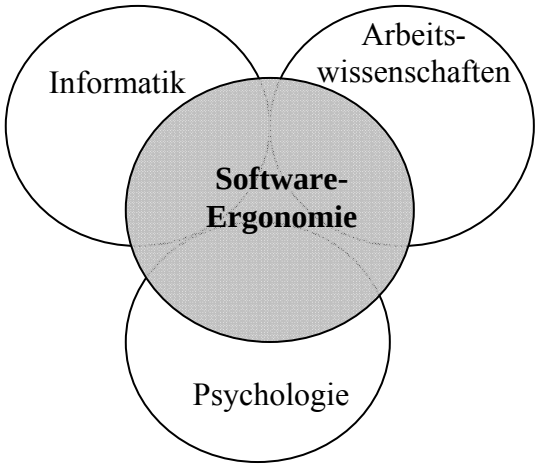


Abbildung 1 Aspekte der Software-Ergonomie

Bei der Software-Ergonomie handelt es sich im klassischen Sinne um eine Arbeitswissenschaft, die eine Anpassung der Software an den Menschen mit dem Ziel einer menschengerechten Arbeitsgestaltung hat. Es ist allerdings nicht nur die Arbeitswissenschaft per se, sondern auch die Informatik und die Psychologie, die, wie in Abb. 1 dargestellt, als disziplinäre Wurzeln der Software-Ergonomie gelten (vgl. Maass, 1993, S. 192 ff.).

Bei der Benutzung von Software kommt es immer zu dem Zusammenwirken von Aufgabe, Benutzer und Computer. Dies wird von Frese & Brodbeck (1989, S. 101) als Triade in einem ABC-Modell[4] dargestellt. Sie betonen die Schnittstellen zwischen Aufgabe, Benutzer und Computer als wesentliche Faktoren der Software-Ergonomie. Oberquelle (1991, S. 9 ff.) erweiterte dies mittels einer modifizierten Leavitt-Raute um die Komponente Organisation (vgl. Abbildung 2). Damit wird auf den besonderen Bezug zu verteilten Systemen und zusätzlich einer gegenseitigen Wechselbeziehung der Komponenten untereinander hingewiesen. Dies charakterisieren auch die vier Aspekte der Software-Ergonomie: menschengerecht, aufgabenangemessen, organisationsorientiert und technikbewusst.

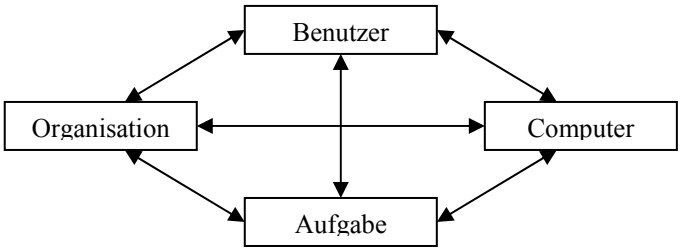


Abbildung 2 Modifizierte Leavitt-Raute (Oberquelle, 1991, S. 11)

Herczeg (2005, S. 5 ff.) weist darauf hin, dass die Software-Ergonomie nicht autark betrachtet werden kann, sondern auch im Kontext der Hardware-Ergonomie, die die technischen Rahmenbedingungen liefert. Dabei gilt es prospektiv die Beeinträchti-

[4] In der ursprünglichen Form wird nicht der Computer, sondern stattdessen allgemeingültig ein System spezifiziert, so dass man eigentlich von einem ABS-Modell sprechen müsste. Da sich die Autoren auf den konkreten Fall der Mensch-Computer-Interaktion beziehen, wird es im allgemeinen ABC-Modell genannt.

gungen der Benutzer zu reduzieren, so dass eine möglichst hohe Arbeitsproduktivität und -qualität im Sinne der Effektivität und Effizienz ermöglicht wird.

Ziel der Software-Ergonomie ist es somit, Software dahingehend zu gestalten, dass sie den Benutzer bei der Durchführung von Aufgaben am Computer im Kontext der Organisation unterstützt. Diese soll unter den Gesichtspunkten der Effektivität, Effizienz und Zufriedenheit der Benutzer, die in der internationalen Norm DIN EN ISO 9241-11 als Kernmerkmale spezifiziert sind, gestaltet werden (vgl. Abschnitt 4.4.2).

3.2 Wirtschaftsinformatik

Der im Kapitel 2 gezeigte betriebliche Einsatz von Informations- und Kommunikationssystemen führt zur Wirtschaftsinformatik. Sie hat ihren Ursprung in den 60er Jahren des 20. Jahrhunderts in der betrieblichen Datenverarbeitung. Heute gilt sie als „eine anwendungsorientierte und interdisziplinäre Wissenschaft“ (Abts & Mülder, 2004, S. 2), deren Anwendbarkeit in der Praxis als ein wesentlicher Vorteil gegenüber den anderen Disziplinen gesehen werden kann. Die Wirtschaftsinformatik wird somit als eine Disziplin der Realwissenschaft charakterisiert, die auch formal- und ingenieurwissenschaftliche Methoden anwendet (vgl. Heinrich, 2001, S. 73 f.), um sich mit Informations- und Kommunikationssystemen in der Wirtschaft und der Verwaltung zu beschäftigen. Diese Systeme sind gekennzeichnet als „soziotechnische Systeme, die menschliche und maschinelle Komponenten (Teilsysteme) als Aufgabenträger umfassen, die voneinander abhängig sind, ineinandergreifen und / oder zusammenwirken“ (vgl. WKWI, 1994, S. 80 f.). Eine „sinnvolle Integration von Betriebswirtschaftslehre und Informatik“ (Abts & Mülder, 2004, S. 2) ist somit im Kern Gegenstand der Wirtschaftsinformatik. Sie hat allerdings auch Schnittpunkte mit anderen Wissenschaften, u.a. der Psychologie, den Arbeitswissenschaften, den Rechtswissenschaften, als auch der Mathematik und Technik (vgl. Abb. 3).

Es sei angemerkt, dass die Formulierung „interdisziplinäre Wissenschaft“ in der Fachwelt kritisch hinterfragt wird. Fehling & Jahnke (1999, S. 199) charakterisieren den wissenschaftlichen Ansatz lediglich als *trans*disziplinär. Müller-Merbach (2002, S. 300 f.) sieht die Wirtschaftsinformatik als „eine Frage der persönlichen Einstel-

lung“, bei der eine *bi*disziplinäre Sichtweise als Brückenfunktion zwischen der Betriebswirtschaftslehre und der Informatik zu einer umfassenderen *inter*disziplinären Sichtweise erweitert wird. Somit sollte laut Rolf die Frage der „Interdisziplinarität und Methodenvielfalt ein Kernthema der Wirtschaftsinformatik“ (Rolf, 1998, S. 263) sein.

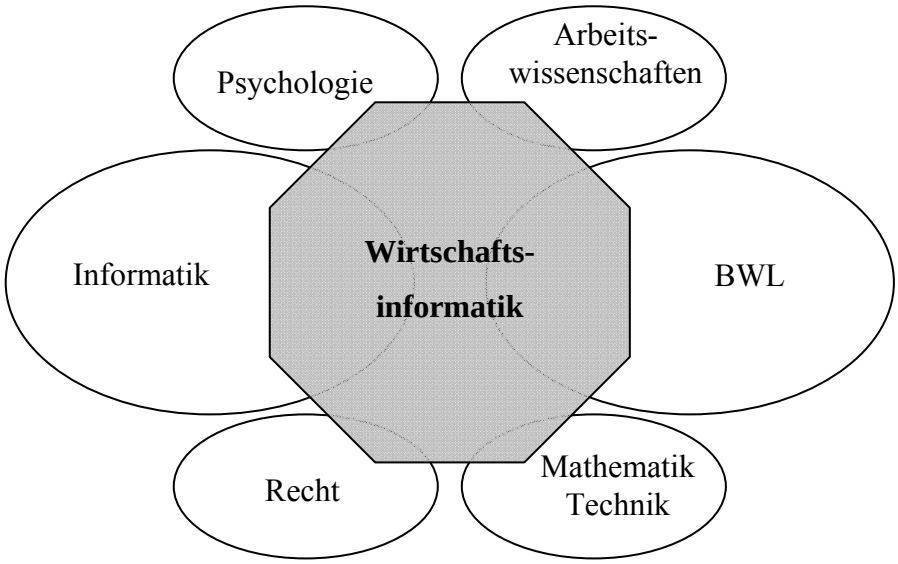


Abbildung 3 Stellung der Wirtschaftsinformatik (in Anlehnung an Abts & Müller (2004, S. 2) und Riemann (2001, S. 3))

Ein klares Ziel der Wirtschaftsinformatik lässt sich daher auch schlecht formulieren. Heinrich (2001, S. 287) kritisiert, dass der Gegenstandsbereich nicht deutlich wird. Mertens (1995, S. 48) forderte als Langfristziel die „sinnhafte Vollautomatisierung“, die durch die Wirtschaftsinformatik vorangetrieben werden sollte. Hoch (1995, S. 328 f.) stimmt Mertens Forderung zu und sieht das Ziel der Wirtschaftsinformatik sogar als „erweiterungsfähig, wenn nicht sogar -bedürftig“ an. Neben der Vollautomatisierung, nur Maschinen als Aufgabenträger zu sehen, ist noch zwischen der Teilautomatisierung, dem gemeinsamen Zusammenwirken von Mensch und Maschine, und der Nicht-Automatisierung, der alleinigen Ausführung durch den Menschen, zu unterscheiden (vgl. Ferstl & Sinz, 2001, S. 47 f.).

Informativer ist in diesem Zusammenhang eine Experten-Befragung unter Persönlichkeiten des Faches Wirtschaftsinformatik, die von Heinzl, König & Hack (2001)

durchgeführt wurde. Die Auswertung ergab, dass die Interdisziplinarität, die es zu vertiefen gilt, als Kernkompetenz angesehen wird. Die drei wichtigsten Erkenntnisziele der Wirtschaftsinformatik für die nächsten zehn Jahre sind nach dieser Befragung:

1. Komplexitätsbeherrschung in Informations- und Kommunikationssystemen
2. Netzmärkte und virtuelle Märkte
3. Anwender-/ Mensch-Maschine-Schnittstellen

Es sei darauf hingewiesen, dass dieses Befragungsergebnis sehr kritisch diskutiert wurde. Eversmann sieht die Expertenbefragung als zweifelhaft an. Er kritisiert, dass die Ausrichtung der Wirtschaftsinformatik nicht langfristig erfolgt, sondern nach der Maxime: „Mit welchen Themen erhöhen wir unsere Chancen mehr Fördergelder zu erhalten[…]?“ (Eversmann, 2002, S. 92). Eversmann erinnert an Mertens Forderung der „sinnhaften Vollautomatisierung“, die nicht ausreichend diskutiert wurde. Ferner sollte sich die Wirtschaftsinformatik nicht nur mikroökonomisch orientieren, sondern dies um den Kontext der makroökonomischen Sicht ergänzen.

3.3 Vergleich beider Disziplinen

Die im vorigen Abschnitt zitierte Befragung von Heinzl et al. (2001) ist in der Hinsicht interessant, dass als Erkenntnisziele der Wirtschaftsinformatik in den nächsten zehn Jahren schon an der dritten Stelle die Anwender-/ Mensch-Maschine-Schnittstellen genannt werden; gleich nach den IuK-Systemen und Netzmärkten. Die Benutzungsschnittstellen zwischen Benutzer und Computer und damit auch die Disziplin der Software-Ergonomie sind demnach scheinbar ein wichtiger Aspekt der Wirtschaftsinformatik. Oberquelle (2000, S. 4) stellte fest, dass sich die Wirtschaftsinformatik scheinbar noch nicht ausreichend mit der Benutzbarkeit von Softwareprodukten und überzeugenden Wirtschaftlichkeitsrechnungen beschäftigt hat, obwohl dies von eminenter Wichtigkeit wäre. Auch die WKWI (1994, S. 81) nennt ausdrücklich die Mensch-Maschine-Schnittstelle als eine Komponente von Informations- und Kommunikationssystemen, die Gegenstand der Wirtschaftsinformatik ist. Es stellt

sich somit die berechtigte Frage, warum sich die Wirtschaftsinformatik anscheinend immer noch nicht ausreichend mit der Mensch-Maschine-Schnittstelle beschäftigt.

Vergleichen wir die beiden Disziplinen, Software-Ergonomie auf der einen und Wirtschaftinformatik auf der anderen Seite (vgl. Tab. 1), so zeigen sich Überschneidungen. Beide Disziplinen werden als interdisziplinär charakterisiert. Sie haben beide die Informatik als Grundlagenwissenschaft und bedienen sich dem jeweiligen Schwerpunkt entsprechend zusätzlicher Disziplinen, die teils sowohl von der Wirtschaftsinformatik als auch die Software-Ergonomie genutzt werden.

Software-Ergonomie	**Wirtschaftsinformatik**
Arbeitswissenschaften	Arbeitswissenschaften
	Betriebswirtschaftslehre
Gestaltungswissenschaften (Design & Kunst)	
Human- und Geisteswissenschaften (Psychologie, Physiologie, Medizin, Soziologie, Linguistik)	Psychologie, Soziologie
Informatik, Technik	Informatik, Technik
	Recht

Tabelle 1 Gegenüberstellung von Software-Ergonomie und Wirtschaftsinformatik

Betrachten wir die inhaltlichen Schwerpunkte der Wirtschaftsinformatik, so zeigt sich, dass dies vielmehr eine starke Konkretisierung des Konstruktes der Leavitt-Raute (vgl. Abschnitt 3.1) im betriebswirtschaftlichen Kontext ist (vgl. Abb. 4). Dazu passt die Aussage von König & Heinzl (2002, S. 510), die als Langfristziel eine „Theorie des Kollaborationsindividualisten oder Individualkollaborateurs, in welcher zugleich die Rolle des Einzelnen und des Netzes in der Informationsgesellschaft bestimmt wird […]“ sehen.

Daraus ergibt sich, dass die Wirtschaftsinformatik „mit Hilfe von Informatik-Methoden, Modellen und Werkzeugen, Konzepte beim Aufbau von Informationssys-

temen für Organisationen bereitstellt“ (Rolf, 2004, S. 44). Die Ergonomie, in diesem Fall insbesondere die Software-Ergonomie, ist eine solche Methode, die eine konkrete Ausgestaltung der Benutzungsschnittstellen unter den genannten (wirtschaftlichen) Aspekten der Effektivität, Effizienz und Zufriedenheit der Benutzer realisiert.

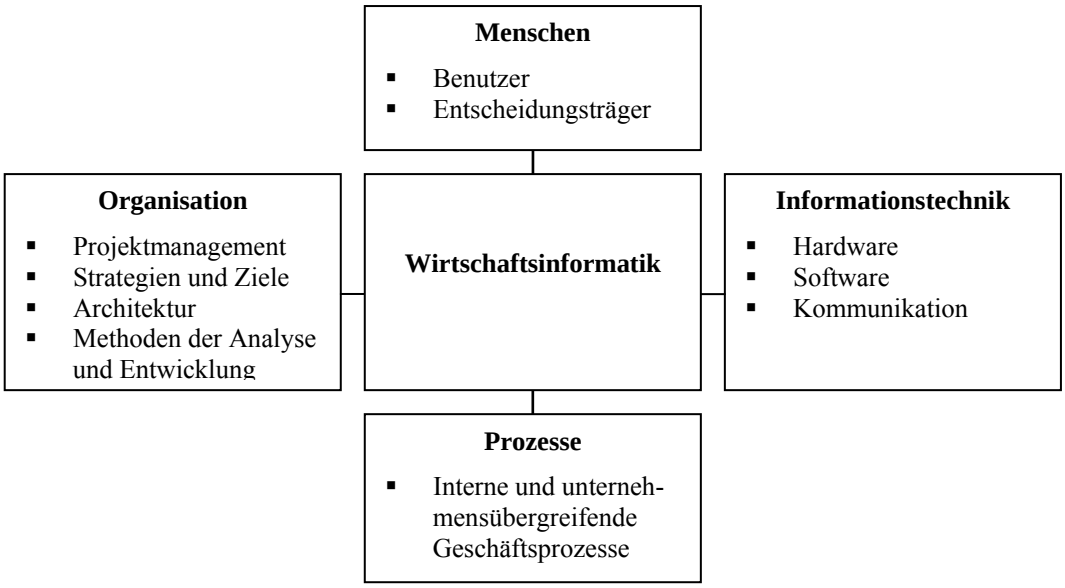


Abbildung 4 Schwerpunkte der Wirtschaftsinformatik (vgl. Abts & Mülder, 2004, S. 4)

Mit der vorangegangenen Diskussion konnte gezeigt werden, dass es sich sowohl bei der Wirtschaftsinformatik als auch die Software-Ergonomie „um eine stark anwendungsorientierte Wissenschaft handelt, bei der es um Aussagen und Erkenntnisse über die Wirklichkeit und die Lösung konkreter Probleme aus der Wirklichkeit geht“ (Schwarze, 2000, S. 25). In den nachfolgenden Kapiteln wird sich zeigen, dass das Zeitverhalten interaktiver Systeme als ein solches Wirklichkeitsproblem anzusehen ist. Die Lösungsansätze der Software-Ergonomie einerseits, als auch der Wirtschaftsinformatik andererseits bieten in Kombination einen Methodenmix, der sich für die Thematik der Antwortzeiten von Systemen sinnvoll ergänzt.

4 Zeitverhalten interaktiver Systeme

4.1 Definition der Systemantwortzeiten

Die Definition des Begriffes von Antwortzeiten eines Systems an sich wirft schon die Frage einer einheitlichen Benennung auf. In der englischen Literatur wird von „response time“ (Miller, 1968; Martin, 1973) gesprochen, während Shneiderman (1984, S. 266) dies näher definiert als „computer system’s response time“. Hierzu hat sich in der deutschen Literatur das eigenwillige Wort der „Systemresponsezeiten“ (vgl. Boucsein, Greif, Wittekamp, 1984; Alexander, 1986; Holling, 1989; Meyer, Hänze, Hildebrandt, 1999) verfestigt. Herczeg (2005, S. 107) verkürzt dies sogar nur noch auf den Begriff der Antwortzeit. Als der deutschen Sprache am nächsten zum englischen Ausdruck *system response time* ist der Begriff der „Systemantwortzeit“ (SAZ) (Hüttner, Wandke, Rätz, 1995, Kapitel 5, S. 20) zu sehen, der im Rahmen dieser Arbeit verwendet wird.

Die erste Veröffentlichung zur Problematik der Systemantwortzeiten gab es von Miller (1968), der darauf hinwies, dass Menschen mit unterschiedlichen Handlungen und Absichten verschiedene Antwortzeiten akzeptieren oder für nützlich halten. Miller geht von der Annahme aus, dass die menschliche Verhaltenweise zeitabhängig ist und signifikante Auswirkungen auf das Verhalten mit der Umwelt hat. Bei einer zwischenmenschlichen Kommunikation legt er einen Erwartungswert von zwei bis vier Sekunden für das Vorliegen einer Antwort zugrunde. Wird dieser Erwartungswert ohne eine Antwort überschritten, führt dies zu einer Beeinträchtigung in der Kommunikation. Dieses Erwartungsverhalten lässt sich auf die Mensch-Computer-Interaktion (MCI) übertragen, wobei es sich nicht auf eine Zwei-Sekunden-Regel generalisieren lässt, sondern in dem jeweiligen Kontext betrachtet werden muss. Miller führt 17 Situationen der MCI auf, in denen jeweils ein unterschiedliches Erwartungsverhalten der Antwortzeiten durch den Kontext begründet wird. Dies reicht von nicht mehr als 0,1 Sekunden für eine Aktivierung des Systems bis hin zu einer Minute zum Starten eines Programms. Er weist insbesondere darauf hin, dass das

System dem Benutzer Rückmeldung über die zu erwartende Zeitdauer der Antwort geben soll (Miller, 1968, S. 267 ff.).

Martin (1973, S. 321-332) greift Millers Artikel auf und definiert in der MCI die Antwortzeit als ein Intervall zwischen dem Drücken der letzten Taste durch den Benutzer und dem Anzeigen der Antwort auf dem Display des Computers. Während Martin darauf hinweist, dass die Standardabweichung der Antwortzeiten möglichst gering zu halten ist, belegt Miller (1968, S. 270) dies mit einem Versuch, bei dem 75% der Teilnehmer eine beidseitige Abweichung der Antwortzeit um 8% im Zeitintervall von 2 bis 4 Sekunden tolerieren. In Abbildung 5 wird mit System A eine gute Antwortzeitverteilung skizziert, in der die Standardabweichung gering ist. Dagegen zeigt System B eine schlechte Antwortzeitverteilung mit einer hohen Standardabweichung und der damit einhergehenden Frustration der Benutzer.

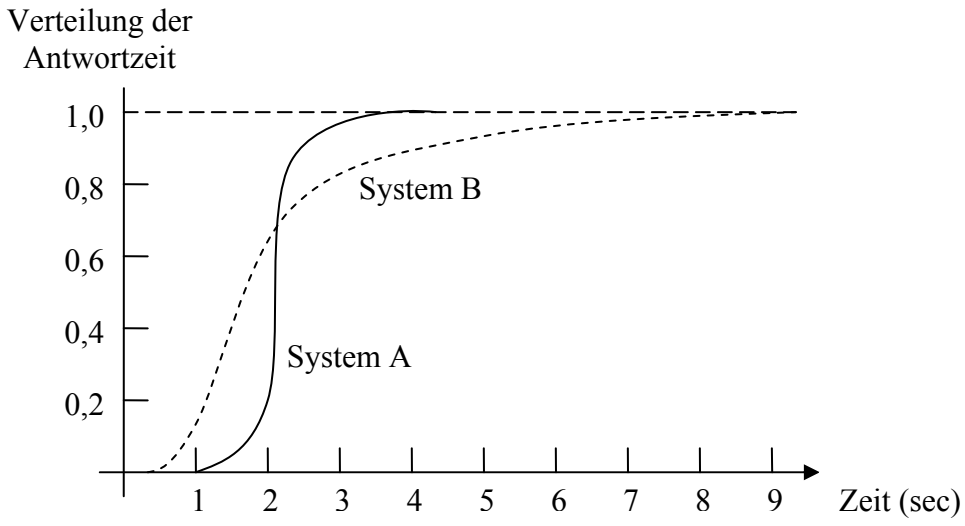


Abbildung 5 Standardabweichung der Antwortzeiten (in Anlehnung an Martin, 1973, S. 322)

Wird von Systemantwortzeiten gesprochen, so gilt es diesen Zeitraum als einen Teil der Mensch-Computer-Interaktion zu sehen. Shneiderman (1984) veranschaulicht dies mit einem einfachen Modell (vgl. Abb. 6), in dem angenommen wird, dass die Systemantwortzeit das Zeitintervall zwischen auslösender Benutzeraktivität und Antwort des Computers sei. Der Benutzer hat nach der Ausgabe Zeit zum Denken, um dann erneut einen selbigen Zyklus zu starten. Shneiderman selbst erweitert dieses Model in ein realistischeres (vgl. Abb. 7), weil davon auszugehen ist, dass der Be-

nutzer schon während der Eingabe plant und Veränderungen vornimmt. Die Planungszeit geht während der Ausgabe des Computers in die Denkzeit über. Hiernach wäre die SAZ das gesamte Zeitintervall von der auslösenden Benutzeraktivität, über die Berechnung bis zur vollständigen Antwort des Systems. Boucsein (1987, S. 165) weist darauf hin, dass die von Shneiderman (1984) eingeführten Planungs- und Denkzeiten nicht direkt trennbar sind. Herczeg (2005, S. 107) merkt an, dass sich die Zeiten wegen der Unschärfe und vielen Abhängigkeiten nur schwer messen lassen.

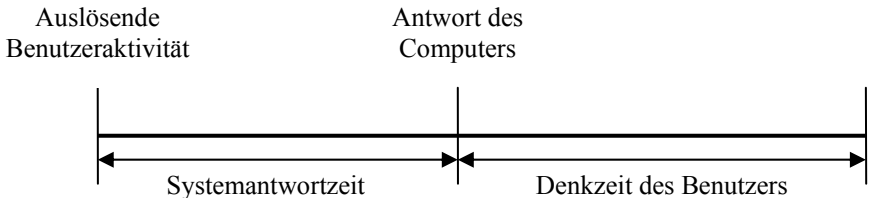


Abbildung 6 Einfaches Antwortzeitmodel (vgl. Shneiderman, 1984, S. 267)

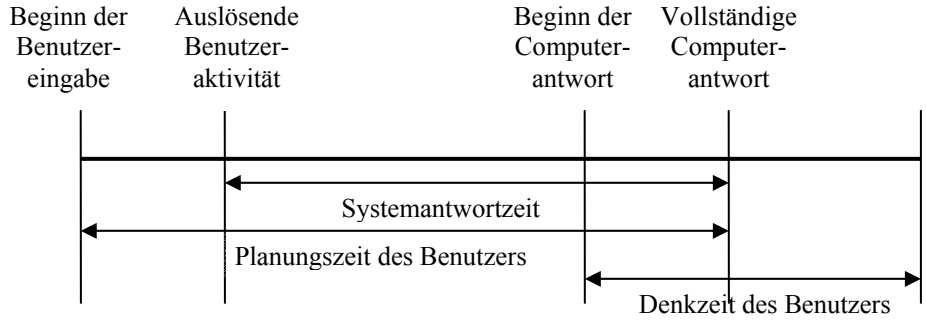


Abbildung 7 Erweitertes Antwortzeitmodel (vgl. Shneiderman, 1984, S. 267)

Boucsein (1987) erweitert und detailliert das Antwortzeitmodell von Shneiderman (1984). Er unterscheidet zwischen dem wirksamen Verhalten des Benutzers vor der Systemantwort und dem darauf antwortenden Verhalten. Das wirksame Verhalten wird detaillierter betrachtet, als die Eingabe, deren Visualisierung auf dem Display und der auslösenden Benutzeraktivität, durch die das System aktiviert wird. Das antwortende Verhalten ist das Zeitintervall der Antwortausgabe auf dem Display und der daran anschließenden Reaktion des Benutzers. Boucseins Ausführungen liegen die Annahmen zugrunde, dass der Benutzer mittels Kommandosprache über Tastatureingabe mit dem Computer interagiert und die Ausgabe der Antwort des Computers nur über den Bildschirm erfolgt.

Im Rahmen dieser Arbeit soll das detaillierte Modell Boucseins (1987, S. 164) als Basis dienen, um ein allgemeingültiges Modell für die zeitlichen Komponenten der Mensch-Computer-Interaktion zu entwickeln (vgl. Abb. 8).

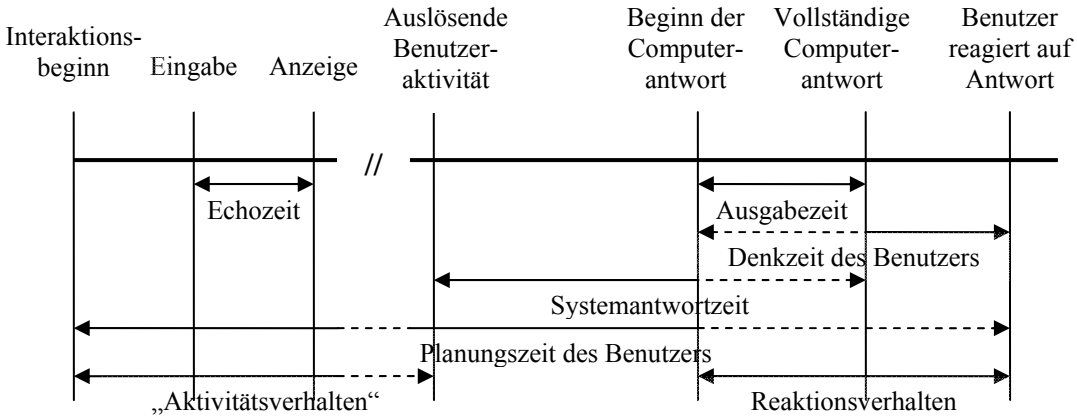


Abbildung 8 Zeitmodell der Mensch-Computer-Interaktion (in Anlehnung an Boucsein, 1987, S. 164)

Die Initialisierung der Interaktion liegt mit dem *Aktivitätsverhalten* beim Benutzer. Es gilt zwischen der Art der Eingabe zu differenzieren: vom einfachen Einschalten eines Systems über die Eingabe per Tastatur oder Maus bis hin zu umfangreicheren Interaktionstechniken. Die Zeitdifferenz zwischen Eingabe und Anzeige der selbigen im System – sei es visuell oder auditiv – wird als *Echozeit* spezifiziert. Diese Echozeit wird durch die Merkmale des Interaktionsmediums und der Art der Anzeige beeinflusst. Dem Starten einer Interaktion mit dem Computer, der *auslösenden Benutzeraktivität*, können z.B. bei der Tastatureingabe mehrere Durchläufe der Echozeit voraus gehen. In dem Moment, in dem die auslösende Benutzeraktivität, z.B. durch Drücken der Return-Taste einen Prozess im Computer startet, beginnt die *Systemantwortzeit*. Diese setzt sich aus einer Transferzeit zwischen den beteiligten Komponenten und der direkten Computer-Antwortzeit (Rechenzeit) zusammen. Das Ende der Systemantwortzeit ist nicht klar zu bestimmen. Zum einen kann der *Beginn der Computerantwort* das Ende der Systemantwortzeit spezifizieren, als aber auch erst die beendete *vollständige Computerantwort*. Dieses Zeitintervall zwischen dem Beginn und der vollständigen Antwort des Computers ist die *Ausgabezeit*. Ebenfalls in dieses Zeitintervall der Ausgabezeit könnte bei einer wahrnehmbaren verzögerten Darstellung die *Denkzeit des Benutzers* fallen, die dann mit dem *Reaktionsverhalten*

gleichzusetzen wäre. Das Reaktionsverhalten des Benutzers resultiert aus der Antwort des Computers, das in einer Reaktion endet. Der gesamte Zeitraum vom Beginn der geplanten Interaktion über die Eingabe bis zur Ausgabe durch den Computer lässt sich als Planungszeit spezifizieren. Die *Planungszeit* könnte, in Voraussicht auf die daraus resultierenden weiteren Aktionen des Benutzers, auch um die Denkzeit des Benutzers erweitert werden.

Es zeigt sich, dass sehr viele, zum Teil nicht genau zu spezifizierende, zeitliche Faktoren in der Mensch-Computer-Interaktion zu beachten sind, die maßgeblich Einfluss auf das zu erreichende Ergebnis haben. Im weiteren Verlauf dieser Arbeit soll der Fokus auf den Systemantwortzeiten liegen. Sie wird hier definiert als die Reaktionszeit des Systems mit dem Zeitintervall zwischen der auslösenden Benutzeraktivität und der dadurch resultierenden vollständigen Systemausgabe.

In den nachfolgenden Abschnitten sollen die Auswirkungen der Systemantwortzeiten auf die Mensch-Computer-Interaktion näher betrachtet werden. Als erstes werden die psychologischen und physiologischen Aspekte besprochen (Abschnitt 4.2). Hierzu werden die kognitiven Leistungen, das Wirken von Belastung und Beanspruchung beschrieben. Dies erst allgemein und dann differenzierter auf Arbeitsplatztypen. Es folgt eine mögliche Betrachtungsdifferenzierung anhand von soziographischen Aspekten der Benutzer (Abschnitt 4.3). Daran schließt sich eine detaillierte Untersuchung der für die Mensch-Computer-Interaktion relevanten Normen (Abschnitt 4.4) unter der besonderen Betrachtung der Systemantwortzeiten an.

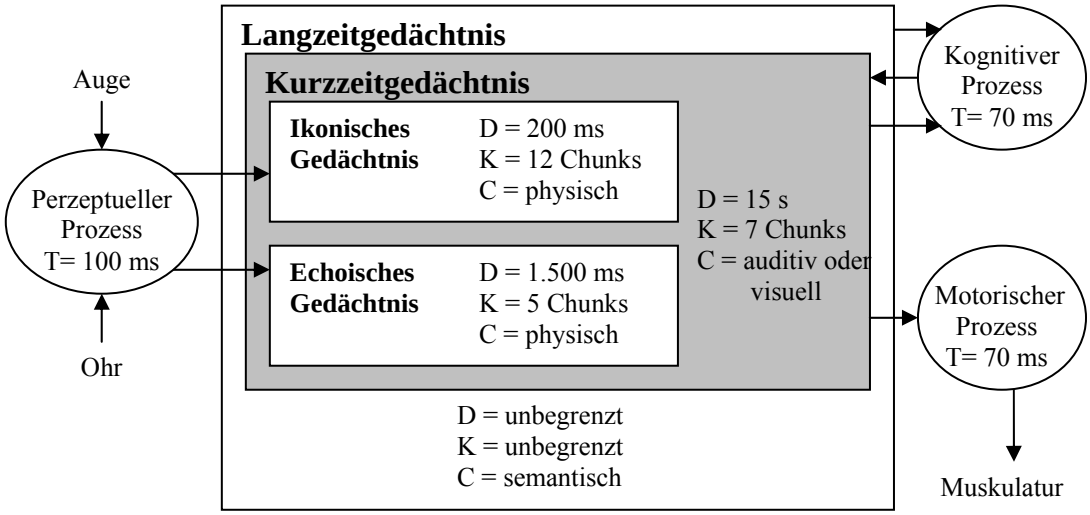
4.2 Psychologische und physiologische Aspekte

4.2.1 Kognitive Leistung

Die Art der kognitiven Informationsaufnahme und -verarbeitung ist in der Mensch-Computer-Interaktion von Interesse, weil die Verarbeitungszeiten und die Informationsspeicherungskapazität des Menschen berücksichtigt werden müssen. Das Gedächtnissystem lässt sich hierbei in drei Typen unterscheiden:

- Sensorisches Gedächtnis
- Kurzzeitgedächtnis
- Langzeitgedächtnis

Das *sensorische Gedächtnis* nimmt die Reizeindrücke auf. In dem Modell von Card, Moran, Newell (1983) wird dies als perzeptueller Prozessor beschrieben, der die Sinnesreize zu Einheiten verschmilzt (vgl. Abb. 9).



Legende: D: Dauer; K: Anzahl; C: Code; T: Zykluszeit

Abbildung 9 Gedächtnisse und Prozessoren (vgl. Heinecke, 2004, S. 54 nach Card et al., 1983, S. 26)

Der Reiz wird dann in das *Kurzzeitgedächtnis* übertragen. Die besondere Charakteristik dieses Gedächtnisses – auch Arbeitsgedächtnis genannt – liegt in der beschränkten Informationskapazität und der kurzen Behaltensdauer. Die Informationseinheiten werden nach Miller (1956) als *Chunks* bezeichnet. Ihre Größe variiert situations- und personspezifisch und kann sowohl einzelne Buchstaben und Zahlen, als

auch Begrifflichkeiten umfassen. Miller (1956) kam durch Untersuchungen zu dem Schluss, dass wir eine Kurzzeitgedächtniskapazität von sieben Chunks haben. Nach Card et al. (1983, S. 25 ff.) können wir zusätzlich noch zwischen dem visuellen ikonischen Gedächtnis, das mit bis zu 12 Chunks und nur 200 ms eine relativ kurze Speicherzeit hat, und dem auditiven, echoischen Gedächtnis mit nur 5 Chunks und 1.500 ms Speicherzeit unterscheiden. Das Behaltensintervall des Kurzzeitgedächtnisses lässt sich mit ca. 15 Sekunden angeben. Interessant ist in dieser Hinsicht, dass die Gedächtnisleistung mit steigendem Zeitintervall des Informationsabrufes monoton abnimmt (vgl. Abb. 10).

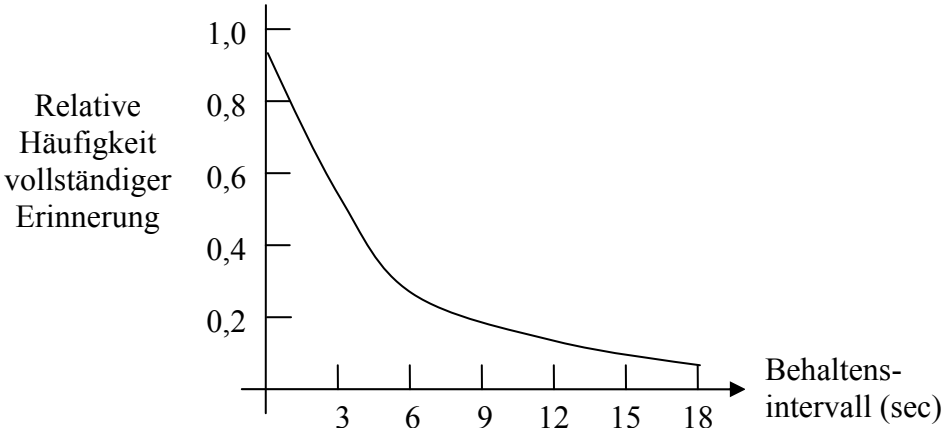


Abbildung 10 Erinnerungskurve des Kurzzeitgedächtnisses (vgl. Peterson & Peterson, 1959, S. 195)

Das *Langzeitgedächtnis* dagegen zeichnet sich durch eine scheinbare unbegrenzte Speicherungsdauer und -kapazität aus. Damit die Informationen lang anhaltend gespeichert werden können, sowohl prozedural als auch semantisch, sind Chunking und elaborierendes Wiederholen wichtige Hilfsmittel. Den Informationen werden zum langfristigen Speichern Bedeutungen zugewiesen (vgl. Zimbardo, 1995, S. 324 ff.).

Für die direkte Mensch-Computer-Interaktion und damit die Systemantwortzeiten im Speziellen, sind die Informationsverarbeitung und deren Geschwindigkeit im Kurzzeitgedächtnis von besonderem Interesse. Systembedingte Antwortzeiten sollten möglichst gering gehalten werden. Die maximal kognitiv mögliche Speicherungsdauer von 15 Sekunden gilt es nicht zu überschreiten. Ansonsten wird der Benutzer in seiner Interaktion gestört, weil er sich nicht mehr ausreichend an seine geplante Handlung erinnern kann.

4.2.2 Beanspruchung und Belastung

Die psychologischen und physiologischen Aspekte im Sinne der Arbeitspsychologie bedingen sich gegenseitig. Eine Belastung ergibt sich aus einem Arbeitsprozess und dem Arbeitsumfeld. Dies führt beim Menschen, entsprechend seines persönlichen Leistungsangebotes, zu einer Beanspruchung (vgl. Hardenacke, Peetz, Wichardt, 1985, S. 68 ff.). Gemäß der DIN EN ISO 10075 Teil 1 (ursprünglich DIN 33405) wird die *psychische Belastung* definiert als die Gesamtheit der externen Einflüsse, die auf den Menschen psychisch einwirken. Diese Belastungen wirken sich sowohl durch die objektive Belastung als auch durch subjektive Einschätzung des persönlichen Leistungsangebots als Beanspruchung auf den Menschen aus (Gros, 1994, S. 96) – vergleiche Abbildung 11. Die *psychische Beanspruchung* ist eine „unmittelbare […] Auswirkung der psychischen Belastung im Individuum in Anhängigkeit von seinen jeweiligen überdauernden und augenblicklichen Voraussetzungen, einschließlich der individuellen Bewältigungsstrategien“ (DIN EN ISO 10075-1, 2000, S. 3).

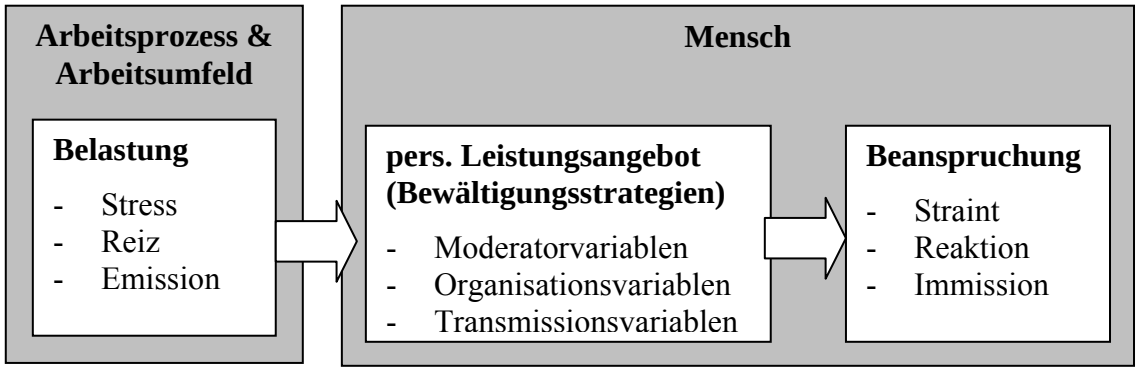


Abbildung 11 Belastungs-Beanspruchungsmodell (in Anlehnung an Gros, 1994, S. 96; Hardenacke, Peetz, Wichardt, 1985, S. 69)

Richter (2000, S. 10 f.) stellt dar, dass zwischen Personen und der Umwelt in Belastungssituationen „komplexe und dynamische Interaktions- und Transaktionsprozesse“ ausgelöst werden, die keine direkte, also auch keine lineare Beziehung zwischen dem Reiz der Belastung und der Beanspruchung als Reaktion geben. Abbildung 12 gibt einen Überblick über die physischen und psychischen Auswirkungen, die durch Belastungen des Arbeitsprozesses in Verbindung mit dem Arbeitsumfeld auftreten

können. Es zeigen sich starke Überschneidungen und nicht eindeutig definierbare Grenzen zwischen den physischen und psychischen Beanspruchungen.

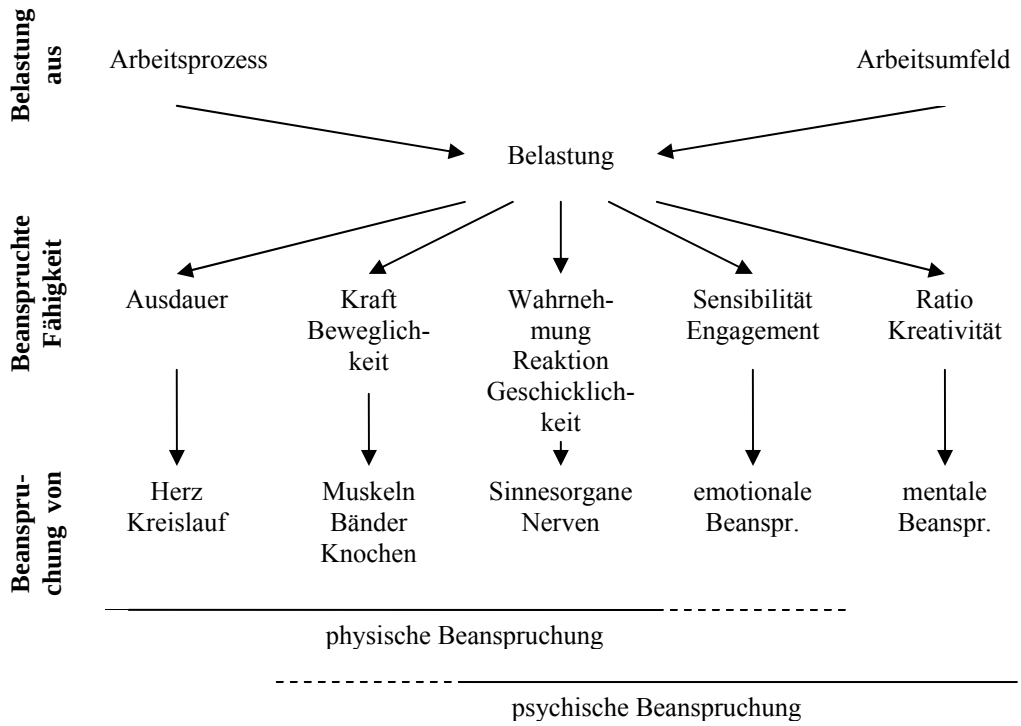


Abbildung 12 Beanspruchte Fähigkeiten (Hardenacke, Peetz, Wichardt, 1985, S. 72)

Es können bei physischer und psychischer Beanspruchung sowohl kurzfristige als auch langfristig beeinträchtigende Folgen eintreten. In einer Betrachtung internationaler Studien resümiert Bödeker (2003, S. 133), dass eine Evidenz zwischen psychischen Faktoren wie Job-Zufriedenheit, soziales Klima, psychische Anforderung und dem Handlungsspielraum zu gesundheitlichen Auswirkungen vielfach vorhanden ist, auch wenn einige Begriffe und Konzepte unscharf beschrieben werden. Insbesondere die Differenzierung von qualitativen und quantitativen Anforderungen führt zu Über- bzw. Unterforderungen, die sich direkt auf die psychische Belastung und Beanspruchung auswirken (Hacker, 1998, S. 30 ff.).

Bevor nachfolgend auf den Faktor Stress als psychische Belastung im Kontext des Bildschirmarbeitsplatzes eingegangen und ein Bezug zu den Systemantwortzeiten hergestellt wird, muss vorher der Arbeitsplatz an sich und dessen Umfeld näher betrachtet werden.

4.2.3 Arbeitsplatztypen

Bei der Bewertung von Arbeitsplätzen mit computergestützten Tätigkeiten muss ein Leitbild zugrunde gelegt werden, welches eine explizite Sichtweise ermöglicht. Es bietet sich der Werkzeug & Material-Ansatz (WAM) an, weil er „die Gegenstände und Konzepte des Anwendungsbereichs als Grundlage des softwaretechnischen Modells“ nimmt (Züllighoven, 1998, S. 4 f.). Der WAM-Ansatz ermöglicht eine genaue Analyse der menschlichen Arbeit und überführt dies mittels Entwurfsmetaphern zur Konstruktion von Anwendungssystemen. Mit Werkzeuge werden wiederholende Arbeitsabläufe und -handlungen beschrieben, die den jeweiligen Aufgaben angepasst werden. Ein Automat ermöglicht eine automatisierte Aufgabenerledigung, die im Voraus genau spezifiziert wird. Als Materialien werden Arbeitsgegenstände angesehen, die mittels Werkzeug und / oder Automat zum Arbeitsergebnis transformiert werden. Von besonderer Bedeutung in unserem Fall ist die Entwurfsmetapher der Arbeitsumgebung, da sie den Raum beschreibt, in dem Arbeitsaufgaben erledigt werden. Da die Gestaltung der Arbeitsumgebung bzw. des Arbeitsumfeldes als eine Quelle für Belastungen gilt (vgl. Abschnitt 4.2.2), ist deren Gestaltung von besonderer Wichtigkeit und soll weitergehend betrachtet werden. Für eine detaillierte Beschäftigung mit dem WAM-Ansatz und der Konstruktion interaktiver Systeme sei auf Züllighoven (1998 und 2005) verwiesen.

Die verschiedenartigen Tätigkeiten ermöglichen die Konstruktion universeller Leitbilder von Arbeitsplatztypen, die es nach Aufgaben- und Unterstützungsumfang zu differenzieren gilt (vgl. Züllighoven, 1998, S. 93 ff.):

- Funktionsarbeitsplatz für eigenverantwortliche Expertentätigkeit
- Gruppenarbeitsplatz für eigenverantwortliche, kooperative Aufgabenerledigung
- Selbstbedienungsautomat

Allgemein lassen sich die Arbeitsplatztypen nach ihren Anteilen an situativ flexiblen und repetitiven Tätigkeiten, den vorhandenen Fach- und IT-Kenntnissen sowie den Ausstattungsmerkmalen des Arbeitsplatzes unterscheiden. *Funktionsarbeitsplätze* zeichnen sich dadurch aus, dass sie auf einen bestimmten Aufgabenbereich ausge-

richtet sind, der sich durch Wiederholungen und die Forderung nach einer hohen Geschwindigkeit charakterisieren lässt (vgl. Züllighoven, 1998, S. 94). Als Beispiel sei der Back-Office-Bereich genannt. Durch dessen Anforderungen ergäbe sich eine hohe Erwartungshaltung bei den Benutzern, dass das System möglichst schnell und beständig in seinem Antwortzeitverhalten ist.

Expertentätigkeiten dagegen zeichnen sich dadurch aus, dass sie eigenverantwortlich, kreativ in einem ständig wechselnden Aufgabenfeld mit hoher Komplexität agieren und in ihrer Tätigkeit durch Anwendungssysteme unterstützt werden (vgl. Züllighoven, 1998, S. 80 f.). Die Komplexität der zu bewältigen Aufgaben legt einen entsprechend hohen Verarbeitungsaufwand im Anwendungssystem nahe, so dass bei Expertentätigkeiten von höheren tolerierbaren Systemantwortzeiten im Gegensatz zum reinen Funktionsarbeitsplatz ausgegangen werden kann.

Während Funktionsarbeitsplätze für nur einen Benutzer ausgerichtet sind, werden *Gruppenarbeitsplätze* von mehreren Benutzern genutzt. Dies erfordert einen gesicherten Informationsaustausch und Abstimmung zwischen den Gruppenmitgliedern (vgl. Züllighoven, 1998, S. 456). Eine Aufgabenteilung zwischen den Gruppenmitgliedern mit unterschiedlichen Qualifikationen (vgl. Züllighoven, 1998, S. 94) kann zu zusätzlichen Verzögerungen im Arbeitsablauf führen. Antwortzeiten sind somit nicht nur durch das System an sich gegeben, sondern werden auch durch den zeitlichen Verlauf des Gruppenprozesses bedingt.

Ein Selbstbedienungsautomat wird meist für eine bestimmte Dienstleistung konzipiert. Die Benutzergruppe und deren Fachwissen sind meist als sehr heterogen anzusehen. Dadurch ergeben sich besondere Anforderungen an umfangreichen Informations- und Hilfemöglichkeiten (vgl. Züllighoven, 1998, S. 97 f.). Als Beispiel seien Bank- und Fahrkartenautomaten genannt. Hier können keine Detailkenntnisse der Benutzer vorausgesetzt werden und es obliegt einer guten Benutzerführung, dass der Benutzer – insbesondere beim Fahrkartenautomat – möglichst schnell seine Aufgabe erledigen kann. Somit gilt es die Systemantwortzeiten möglichst gering zu halten und den Benutzer über den aktuellen Systemzustand zu informieren.

4.2.4 Stress am Bildschirmarbeitsplatz

Der Einsatz von Computern als Bildschirmarbeitsplätze im betrieblichen Umfeld dient primär der Erweiterung von Arbeitstätigkeiten der Benutzer unter dem Gesichtspunkt der Effizienz. Durch diesen Einsatz wird der Arbeitsplatz des Menschen an sich tangiert, so dass hier gewisse (gesetzliche) Anforderungen eingehalten werden müssen, auf die verstärkt in Abschnitt 4.4 eingegangen wird. Ausgangspunkt hierfür ist die eingehende Untersuchung von Çakir, Reuter, von Schmude & Armbruster (1978), die den ergonomischen Aspekt der Gestaltung von Bildschirmarbeitsplätzen untersuchten. Es zeigte sich, dass bei der Nutzung des Bildschirmarbeitsplatzes durch die eingesetzte Computerhardware und -software eine Belastung als Stress für den Benutzer auftrat.

Stress kann nach Greif (1991, S. 13) definiert werden als „ein *subjektiv* intensiv unangenehmer Spannungszustand, der aus der Befürchtung entsteht, daß eine stark aversive, subjektiv zeitlich nahe (oder bereits eingetretene) und subjektiv lang andauernde Situation sehr wahrscheinlich nicht vollständig kontrollierbar ist, deren Vermeidung aber subjektiv wichtig erscheint“. Die Reaktionen auf die Stresssituation können sowohl kurz- als auch mittel- bis langfristig sein. Dies zeigt sich in den in Abbildung 12 genannten physiologischen und psychischen Beanspruchungen, die sich auch auf das grundsätzliche Verhalten des Benutzers auswirken können. Stressoren sind hypothetische Konstrukte, die mit hoher Wahrscheinlichkeit Stress auslösen werden (vgl. Zapf & Frese, 1993, S. 658 f.). Frese & Brodbeck (1989, S. 170) weisen darauf hin, dass alltägliche, kleinere Unannehmlichkeiten – die so genannten Mikrostressoren – für den Stressverlauf am Arbeitsplatz besonders wichtig sind.

Bei der Computerarbeit lassen sich die Stressbedingungen in physische und psychische unterscheiden. Als physische Stressbedingung kann es durch die Nutzung von Bildschirmarbeitsplätzen zu Augenschmerzen und –beschwerden bei den Benutzern kommen. Hierbei hängt es von der Art der Arbeitstätigkeit ab. Ferner gelten sitzende Körperhaltung, Bewegungsarmut bzw. einseitige körperliche Belastung als physische Stressfaktoren. Als psychische Stressbedingungen werden unter anderem das Gefühl des Zeitdrucks, der Überwachung des Arbeiters, die Abstraktheit der Arbeit, die

Angst vor Arbeitslosigkeit und auch die Systemantwortzeiten genannt (Frese & Brodbeck, 1989, S. 177 ff.).

Boucsein et al. (1984) wiesen darauf hin, dass Systemantwortzeiten als ein Belastungsfaktor bei Bildschirm-Dialogtätigkeiten angesehen werden können und diese möglichst gering zu halten sind. Sie beziehen sich auf Timesharing-Systeme, bei denen die Anzahl der Benutzer die Dauer und Streuung der Systemantwortzeit direkt beeinflusst. Boucsein et al. stellten damals die These auf, dass durch eine Vernetzung dezentraler Systeme die Systemantwortzeiten verkürzt werden könnten. Demnach wären sowohl die zeitliche Intensität der Arbeit, als auch die benötigte Zeitdauer des Systems mögliche Stressoren, die auf die Benutzer einwirken.

Als Intensität kann nicht nur einseitig die Wartezeit des Benutzers auf das System, sondern es muss die gesamt Interaktion zwischen dem Benutzer und dem Computer betrachtet werden. So muss auch der Computer auf den Benutzer warten. Es treten somit mehreren Wartezeiten auf, die einer variierenden Verteilung unterliegen und nicht konstant sind. Es gilt sowohl die mittlere Antwortzeit als auch die Streuung der Antwortzeit zu betrachten. Die Streuung der Antwortzeit führt durch ihre Ungewissheit – sowohl die Ereignisungewissheit, als auch die zeitliche Ungewissheit – zu negativen physiologischen und emotional-kognitiven Reaktionen beim Benutzer (vgl. Boucsein et al, 1984, S. 118 ff.).

Die Wahrnehmung der Zeit erfolgt im Gehirn durch die Wahrnehmung der Geschwindigkeit von Bewegungen und sensorischen Prozessen (Granit, 1985, S. 61 f.). Städtler (2003, S. 1248) führt aus, dass bei Ereignissen und interessanten Arbeitsdurchführungen die Zeit als schneller vergehend empfunden wird als bei monotoner Arbeit oder ungefüllten Intervallen. Daraus lassen sich für die Systemantwortzeiten nach Carbonell, Elkin & Nicherson (1968, S. 135-142) drei wesentliche Einflussfaktoren ableiten:

1. Die Antwortzeit selbst
2. Die Art der Tätigkeit
3. Merkmale der Benutzer

Während die Systemantwortzeit bereits in Abschnitt 4.1 beschrieben wurde, wird auf die Merkmale der Benutzer in Abschnitt 4.3 näher eingegangen. Die Art der Tätigkeiten ergibt sich aus der computergestützten Arbeit an sich. Hierzu untersuchte Kühlmann (1993) die Stressbedingungen und Ansätze zur Stressbewältigung. Belastungen, die durch Arbeit am Computer entstehen, sind nicht rein objektiv messbar, sondern sind subjektiv und werden stark durch die Persönlichkeitsmerkmale der Benutzer geprägt. Kühlmann unterteilte die Art der Belastungen in vier Klassen, vgl. Tabelle 2. Hierbei sind Systemfehler und Systemmängel als Stressbedingungen hervorzuheben, weil diese eine geringe Kontrollierbarkeit durch den Benutzer ermöglichen. Lange Antwortzeiten gelten demnach als ein Systemmangel. Dagegen hat der Benutzer eine hohe Kontrollierbarkeit auf Stressbedingungen, die auf Bedienerfehler und Bedienerunsicherheit zurückzuführen sind.

Stressbedingung	**Beispiel**	**Situationskontrolle**
Systemfehler	Ausfälle; Softwarefehler	gering
Systemmängel	Lange Antwortzeiten; Umständliche Benutzerführung	gering
Bedienerfehler	Unbeabsichtigtes Datenlöschen; Erfolgloser externer Speicherzugriff	hoch
Bedienerunsicherheit	Lückenhafte Erinnerung; Unklarheit über Systemstatus	hoch

Tabelle 2 Beschreibung von Stressbedingungen (vgl. Kühlmann, 1993, S. 234 f.)

Bei technisch bedingten Stressbedingungen wie Systemfehlern und -mängeln bemühen sich die Benutzer, diese durch ursachenorientierte Bewältigungsversuche zu beseitigen. Dem gegenüber werden bei stark kontrollierbaren Situationen symptomorientierte Bewältigungsversuche unternommen. Als Ergebnis seiner Untersuchung kommt Kühlmann zu dem Schluss, dass Benutzer bei Stressbedingungen, die sie selbst beeinflussen können, mehr Bewältigungsversuche unternehmen als in Situationen mit gering kontrollierbaren Stressoren. Das psychische und physische Wohlbefinden wird durch den Erfolg bzw. Misserfolg dieser Bewältigungsversuche beeinflusst. Ferner weist Kühlmann darauf hin, dass nicht nur die Kontrollierbarkeit die Stresssituation beeinflusst, sondern auch die Wichtigkeit der zu verrichtenden Aufgabe.

4.3 Soziographische Aspekte

4.3.1 Benutzergruppen

Die Merkmale der Benutzer von Informationssystemen ermöglichen eine Klassifizierung in Gruppen nach unterschiedlichen Persönlichkeitsmerkmalen. Anhand dieser sind differenzierte Aussagen in Bezug auf das Zeitempfinden der Systemantwortzeiten möglich. Zwischen folgenden Merkmalen soll unterschieden werden:

- Alter
- besondere Anforderungen
- Erfahrung der Benutzer

Bei den aufgeführten Persönlichkeitsmerkmalen handelt es sich um eine Auswahl, die in Bezug auf Systemantwortzeiten besonders betrachtet werden sollte. Während beim Alter (Abschnitt 4.3.2) eine dynamische Veränderung über den Lebenszeitraum stattfindet, sind besondere Anforderungen (Abschnitt 4.3.3) bei Personen mit Erkrankungen und Behinderungen evident. Insbesondere die Erfahrungen der Benutzer (Abschnitt 4.3.4) verändern sich über die Zeit oder bei der Anpassung an Situationen, z.B. den Arbeitsbedingungen (vgl. Frese, 1983, S. 11), so dass bezüglich der Systemantwortzeiten eine andere Erwartungshaltung vorhanden ist.

4.3.2 Alter

Beim Alter können wir zwischen den drei Personkreisen der Kinder, der Erwachsen und dem der älteren Menschen unterscheiden. Es ist bekannt, dass bei Kindern bis zum 14. Lebensjahr die Zeiteinschätzung äußerst ungenau ist (vgl. Städtler, 2003, S. 1248). Dies sorgt für eine zusätzliche Verzerrung in der subjektiven Wahrnehmung der Zeitdauer gegenüber der von Erwachsenen. Bei Erwachsenen kann man im Allgemeinen – lässt man subjektive person- und arbeitsspezifische Besonderheiten außer Acht – bei normalem Wachbewusstsein von einer Zeitwahrnehmung sprechen, die mit Zeitmessungen übereinstimmt (vgl. Münzel, 1993, S. 3). Allerdings verändert sich die kognitive Leistungsfähigkeit des Menschen im Laufe des Lebens. Die visuellen und auditiven Wahrnehmungen verschlechtern sich (vgl. Burmester, 2001, S. 12 f.). Bei älteren Menschen – gemeint sind Menschen, die das 65. Lebensjahr über-

schritten haben – „verändert sich zwar die Art der kognitiven Auseinandersetzung mit der Umwelt, aber nicht notwendigerweise die generelle Leistungsfähigkeit“ (Frese & Brodbeck, 1989, S. 94 f.). Czaja & Sharit (1993) wiesen in ihrer Untersuchung signifikante Alterseffekte bei der Arbeit am Computer nach. Es zeigte sich eine stärkere Ermüdung, höhere Fehleranfälligkeit und längere Antwortzeit durch die Benutzerinnen. Es stellt sich die Frage, inwiefern sich eine systembedingte Antwortzeitverzögerung auf ältere Benutzer auswirkt bzw. als Belastung empfunden wird.

4.3.3 Besondere Anforderungen

Mehr als zehn Prozent der deutschen Bevölkerung – ca. 8,4 Million Menschen – sind als Behinderte anerkannt. Zu beachten ist, dass zum einen 77 % der Behinderten älter als 55 Jahre sind und zum anderen die Erwerbsquote bei den 25- bis 45-jährigen Behinderten 72 % beträgt (Pfaff et al., 2004, S. 1181 ff.). Damit fällt auf diese Benutzergruppe ein besonderes Augenmerk, da sie einen bedeutenden Anteil der Bevölkerung stellt und die Nutzung von Computern ihnen eine bedingte Teilnahme am gesellschaftlichen Leben ermöglicht. Es muss daher näher betrachtet werden, bei welchen Arten von Erkrankungen und Behinderungen sich die Systemantwortzeiten besonders auf den Benutzer auswirken. Als Leitfaden der Zugänglichkeit (engl. accessibility) in der Mensch-Computer-Interaktion gilt die ISO/TS 16071 (vgl. auch Abschnitt 4.4.5). Sie nennt spezielle Benutzergruppen, die besondere Anforderungen an die Interaktion und die Benutzung haben.

Exemplarisch genannt seien Menschen mit körperlichen und kognitiven Erkrankungen, mit dadurch bedingter vorübergehender Arbeitsunfähigkeit und vielfachen Körperbehinderungen (vgl. ISO/TS 16071, 2003, S. 7 f). Bei *Körperbehinderungen* sind ggf. speziell auf den Benutzer angepasste Interaktionsgeräte erforderlich (vgl. Weist, 2004, S. 67). Die Individualisierung der zeitlichen Interaktion ist von Bedeutung, damit die Benutzer – bedingt durch ihre Behinderungen – ausreichend Zeit haben, um mit dem System zu interagieren. Bei *kognitiven Behinderungen* gibt es kaum Erkenntnisse in Bezug auf die Mensch-Computer-Interaktion und damit auch keine Erkenntnisse über die Auswirkungen der Systemantwortzeiten. Hier kommt es darauf an, dies an den Bedürfnissen der Menschen anzupassen (vgl. Weist, 2004, S. 69).

4.3.4 Erfahrungen der Benutzer

Nach Zeidler & Zellner (1994, S. 124 ff.) und Herczeg (2005, S. 67 ff.) erfolgt eine Klassifizierung nicht nach den fachlichen Erfahrungen, sondern nur nach dem Vorwissen der Benutzer im Umgang mit den Geräten der Informationstechnik. Es wird zwischen folgenden Gruppen differenziert:

- *Unerfahrene Benutzer*, so genannte Anfänger, die entweder kaum Erfahrung haben oder mit neuen Anwendungssystemen konfrontiert werden;
- *Gelegentliche Nutzer*, die grobe Kenntnisse haben, aber das System zu selten nutzen, um eine Routine zu entwickeln, bzw. die Nutzung nur über einen kurzen Zeitraum erfolgt;
- *Routinebenutzer*, die sich in ein System eingearbeitet haben und es regelmäßig und intensiv nutzen;
- *Experten* haben langjährige Erfahrungen als Routinebenutzer, ein hohes Wissen und erwarten eine komprimierte Darstellung;

Anhand dieser Gruppeneinteilung lässt sich in Bezug auf die Systemantwortzeiten ableiten, dass Experten aufgrund ihres fundierten Wissens eine andere Erwartungshaltung in Bezug auf die Antwortzeit des Systems haben als unerfahrene Benutzer. Der Grund hierfür liegt in der Möglichkeit der Vergleichbarkeit mit anderen Kontexten und der langjährigen Nutzung des Systems. Experten dürften, mit Blick auf den Wunsch einer komprimierten Darstellung, ein stärker ausgeprägtes Zeitgefühl bezüglich des Systems haben und schnelle Antwortzeiten präferieren. In Bezug auf die Erfahrung, Schulbildung und Ausbildung weist die DIN EN ISO 10075 Teil 2 (2000, S. 8) insbesondere auf die Entwicklung einer psychischen Sättigung hin (vgl. Abschnitt 4.4.5).

4.4 Normen

4.4.1 Grundlage

Für den effizienten und effektiven Einsatz von Bildschirmarbeitsplätzen ist insbesondere eine gebrauchstaugliche Gestaltung der Software-Systeme wichtig. Grundlage hierfür ist die Bildschirmarbeitsverordnung (BildschArbV, Anhang 20-22), deren Anforderung seit dem 01.01.2000 für alle Software-Programme bei Bildschirmarbeitsplätzen bindende Gesetzesvorschrift ist. Sie ergibt sich aus dem Arbeitsschutzgesetz, das in §4 Satz 1 Ziffer 3 (ArbSchG) fordert, „bei den Maßnahmen sind der Stand von Technik, […] sowie sonstige gesicherte arbeitswissenschaftliche Erkenntnisse zu berücksichtigen“. DIN-Normen gelten als entsprechende Erkenntnisse, die eine Konkretisierung der Anforderung an Produkte spezifizieren.[5]

Abbildung 13 Aufbau der Vorschriften zur Software-Ergonomie (Bräutigam, Schneider, 2003, S. 17)

[5] Es sei darauf hingewiesen, dass es nicht nur ein Deutsches Institut für Normung (DIN) gibt, sondern auch Europäische Normen (EN) der CEN (Comité Européen de Normalisation) und Internationale Normen des International Standardisation Organisation (ISO). Somit kann eine Norm im Sinne der Standardisierung sowohl den Anforderungen der ISO, EN als auch DIN entsprechen.

Eine Übersicht bietet Abbildung 13, die den Aufbau der Vorschriften zur Software-Ergonomie als konkretisierendes Modell darstellt. Die rechtlichen Aspekte sollen in dieser Arbeit nicht näher beschreiben werden, sondern vielmehr das Zusammenwirken zwischen Mensch und Arbeitsmittel in Bezug auf die Systemantwortzeiten. Die für Softwareprodukte relevante DIN-Norm, ist die Normreihe 9241 mit dem Haupttitel „Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmarbeitsgeräten“. Die nachfolgende Betrachtung bezieht sich auf Teil 11, der grundsätzliche Leitsätze zur Gebrauchstauglichkeit beinhaltet, sowie Teil 10 mit den Grundsätzen der Dialoggestaltung, beschäftigt. Anschließend werden die DIN-Norm 9126, die sich mit der Software-Qualität an sich beschäftigt, sowie weitere relevante Normen, die einen Bezug zu den Systemantwortzeiten aufweisen, betrachtet.

4.4.2 DIN EN ISO 9241-11

Die DIN EN ISO 9241-11 stellt heraus, „dass die Gebrauchstauglichkeit vom Nutzungskontext abhängt und dass die besonderen Umstände, unter denen das Produkt benutzt wird, den Grad der Gebrauchstauglichkeit beeinflussen“ (DIN EN ISO 9241-11, 1998, S. 3). Der Grad der Gebrauchstauglichkeit wird als Erreichung von Zielen mit Effektivität, Effizienz und der Zufriedenstellung spezifiziert:

- *Effektivität* spezifiziert die Genauigkeit und Vollständigkeit, mit der ein Benutzer ein bestimmtes Ziel erreicht.
- *Effizienz* spezifiziert den eingesetzten Aufwand, der zur effektiven Erreichung des Zieles erforderlich ist.
- *Zufriedenstellung* charakterisiert die positive Einstellung und Freiheit von Beeinträchtigungen bei der Nutzung des Produktes durch den Nutzer.

Im Sinne einer allgemeinen Gebrauchstauglichkeit ist die Zeit ein signifikantes Maß der Effizienz. Diese lässt sich definieren als die erforderliche bzw. produktive Zeit im Verhältnis zu einer vorher definierten Effektivität, die vom jeweiligen Ziel abhängt (vgl. a.a.O. S.11 f.). Somit ist das Zeitniveau ein signifikanter Faktor, der nicht nur durch die Charakteristika des Benutzers, sondern auch die Beschaffenheit des verwendeten Produktes bzw. Systems als Arbeitsmittel kennzeichnet.

Bezüglich der Effektivität und Effizienz der Zielerreichung ist ein direkter Bezug zum Grundsatz der Aufgabenangemessenheit gegeben, der in Teil 10 der DIN EN ISO 9241 beschrieben wird. Der Teil 11 stellt somit die Grundlage für gebrauchstaugliche Systeme dar. Ferner weist dieser Normteil Beziehungen zu anderen Normen wie der ISO 9126 (vgl. Abschnitt 4.4.4) und anderen Teilen der DIN EN ISO 9241 auf (vgl. a.a.O. S. 16 f.). Daher wird von Geis, Dzida & Redtenbacher (2004, S. 44 f.) angeregt, die ISO 9241 dahingehend zu überarbeiten und den aktuellen Teil 11 als neuen ISO 9241 – Teil 1 zu spezifizieren; trotz der zu erwartenden Referenzierungsprobleme.

4.4.3 DIN EN ISO 9241-10

Die DIN EN ISO 9241 Teil 10 hat ihren Ursprung in der DIN 66234 und enthält sieben Grundsätze der Dialoggestaltung:

- Aufgabenangemessenheit
- Selbstbeschreibungsfähigkeit
- Steuerbarkeit
- Erwartungskonformität
- Fehlertoleranz
- Individualisierbarkeit
- Lernförderlichkeit

Für eine genaue Spezifikation der jeweiligen Grundsätze sei der interessierte Leser auf die Norm verwiesen. An dieser Stelle sollen nur die Grundsätze detailliert betrachtet werden, die sich auf das Zeitverhalten beziehen. Dies sind im Kern die Grundsätze Erwartungskonformität, Individualisierbarkeit und Steuerbarkeit

Der Grundsatz der *Erwartungskonformität* wird wie folgt definiert: „Ein Dialog ist erwartungskonform, wenn er konsistent ist und den Merkmalen des Benutzers entspricht, z.B. seinen Kenntnissen aus dem Aufgabengebiet, seiner Ausbildung und seiner Erfahrung sowie den allgemein anerkannten Konventionen.“ (DIN EN ISO 9241, 1996, S. 6). In den Empfehlungen der DIN-Norm heißt es hierzu, dass der Benutzter darüber zu informieren ist, wenn „voraussichtlich erhebliche Abweichungen

von der erwarteten Antwortzeit“ (a.a.O. S. 7) zu erwarten sind. Dies kann als einfache Mitteilung oder grafische Warte-Anzeige erfolgen.

Interessant hieran ist, dass von einer *erheblichen Abweichung* der zu *erwartenden Antwortzeit* ausgegangen wird. Dies lässt Raum zur Interpretation, ob bei signifikant geringen bzw. nicht vorhandenen Abweichungen der zu erwartenden Antwortzeit keine Rückmeldungen erforderlich sind. Ferner stellt sich die Frage bezüglich der Erwartung, die an die Antwortzeit des Dialogsystems gestellt wird. Wodurch wird diese Wartezeit bedingt? Schneider (1998a, S. 93 f.) interpretiert es dahingehend, dass das System möglichst unmittelbar reagieren bzw. dass eine Verzögerung nicht bemerkbar sein sollte. Er kritisiert eine einfache Warte-Anzeige dahingehend, dass diese dem Benutzer keine spezifischen Informationen über den aktuellen Bearbeitungszustand gibt. Ferner weist Schneider darauf hin, dass die Benutzererwartung hinsichtlich der Systemantwort zu ermitteln sei. Die Bearbeitungszeit von Prozessen ist für jeden Computer individuell und führt somit zu der Problematik, dass Verzögerungszeiten nicht pauschal ermittelt werden können.

Es lässt sich somit als Anforderung an die DIN-Norm die Forderung erheben, dass gemäß dem Grundsatz der *Selbstbeschreibungsfähigkeit* unabhängig von einer Abweichung der zu erwartenden Antwortzeit, immer eine unmittelbare Rückmeldung über den Systemzustand bzw. den Arbeitsfortschritt und ggf. der zu erwartenden Antwortzeit durch das System an den Benutzer erfolgen muss. Als Begründung hierfür sind benutzerspezifische Erwartungshaltungen und Tolleranzschwellen anzuführen, die zusätzlich zum jeweiligen Arbeitskontext variieren und somit nicht pauschalisiert angegeben werden können.

Auf die Interaktionsmöglichkeit des Benutzers geht der Grundsatz der *Individualisierbarkeit* ein: „Ein Dialog ist individualisierbar, wenn das Dialogsystem Anpassungen an die Erfordernisse der Arbeitsaufgabe sowie die individuellen Fähigkeiten und Vorlieben des Benutzers zulässt.“ (a.a.O. S. 8). In der fünften Empfehlung hierzu heißt es, dass der Benutzer die Möglichkeit haben sollte, die Zeitparameter der

Dialogfunktionen individuell anzupassen. Als Beispiel wird die Geschwindigkeit des Scrollens angeführt.

Generell sollte den Benutzern in einer sinnvollen Weise die Möglichkeit offeriert werden, Geschwindigkeit der Interaktion und ggf. der Systemantwortzeit ihren jeweiligen Erfordernissen, den individuellen Fähigkeiten und Vorlieben anzupassen. Dies deckt sich mit einer Empfehlung des Grundsatzes der *Steuerbarkeit*, die fordert, dass die Geschwindigkeit eines Dialoges nicht vom Dialogsystem vorgeschrieben werden sollte (vgl. a.a.O. S. 6). Es muss hierbei allerdings die technisch bedingte, minimal mögliche Systemantwortzeitdauer bedacht werden, die nicht unterschritten werden kann.

4.4.4 ISO/IEC 9126

Während die vorigen Abschnitte die software-ergonomischen Aspekte der DIN EN ISO 9241-Serie behandelten, gilt es nunmehr auch die software-technischen Aspekte zu betrachten. Hierzu liefert die ISO/IEC 9126 Norm – ehemals DIN 66272 – ein Qualitätsmodell für Softwareprodukte.

Kernpunkte für Softwarequalität sind Funktionalität, Zuverlässigkeit, Gebrauchstauglichkeit, Effizienz, Wartbarkeit und Übertragbarkeit. Als Effizienz wird das Leistungsniveau zwischen der Software an sich und anderen verwendeten Ressourcen, wie z.B. anderen Software-, Hardwarekomponenten, Materialen und den Benutzern verstanden (vgl. ISO/IEC 9126, 1991, S. 2 f.). Eine Konkretisierung des Effizienz-Merkmales bezieht sich auf das Zeitverhalten. Darunter werden Antwortzeit, Verarbeitungszeit und Durchsatzraten in der Funktionsdurchführung verstanden (vgl. a.a.O. S. 10). Somit ist das Zeitverhalten eines Softwaresystems als Effizienzkriterium wichtiger Qualitätsfaktor.

Meyer, Vogt, Glier (2005a) kritisieren, dass es kaum kontext-unabhängige, quantitative Werte gibt, aus denen sich konkrete Zeitangaben für die jeweiligen Anforderungen des Softwareeinsatzkontextes ableiten lassen. Sie weisen darauf hin, dass sich

oftmals die Benutzer dem teils willkürlichen Zeitverhalten der Systeme anpassen müssen und dies meistens auch tun. Dies widerspricht allerdings dem Leitbild der Software-Ergonomie. Lediglich bei der Direktmanipulation lässt sich nach Shneiderman, Plaisant (2005, S. 473) feststellen, dass für eine erfolgreiche Interaktion eine Antwortzeit im Bereich von 50 bis 150 Millisekunden erforderlich ist.

4.4.5 Weitere Normen

Neben den genannten Normen gibt es noch eine Reihe weiterer Normen, die sich mit der software-ergonomischen Gestaltung interaktiver Systeme beschäftigen. Auf diese sollen im Rahmen dieser Arbeit allerdings nicht näher eingegangen werden, da sie im Kern auf die besprochene Normreihe 9241 Bezug nehmen und bezüglich der Systemantwortzeiten keine neuen Erkenntnisse liefern.[6]

Es soll an dieser Stelle auf die im vorigen Text schon zitierte *DIN EN ISO 10075* „Ergonomische Grundlagen bezüglich psychischer Arbeitsbelastung“ eingegangen werden. Der erste Teil befasst sich allgemein mit den Begriffen der psychischen Belastung und Beanspruchung (vgl. Abschnitt 4.2.2), während der dritte Teil Verfahren zur Messung und Erfassung psychischer Arbeitsbelastungen vertieft, die hier nicht weiter behandelt werden sollen.

Im zweiten Teil werden generalisierte und schon diskutierte Gestaltungsgrundsätze beschrieben, die fordern, dass „das Arbeitssystem an den Nutzer anzupassen“ (DIN EN ISO 10075, 2000, S. 3) ist und personelle Faktoren mit den Wechselwirkungen der technischen und organisatorischen Faktoren zu berücksichtigen sind (vgl. a.a.O. S. 2). Es wird gefordert „die Intensität der Arbeitsbelastung zu reduzieren oder zu optimieren, [und] die Zeit der Exposition zu begrenzen“ (a.a.O. S. 4). In Bezug auf die Systemantwortzeit heißt es im Unterpunkt Zeitverzögerungen hierzu:

---

[6] Hierzu wurden die DIN EN ISO 13407 „Benutzer-orientierte Gestaltung interaktiver Systeme“ und DIN EN ISO 14915 „Software-Ergonomie für Multimedia-Benutzungsschnittstellen“ Teile 1-3 durchgearbeitet.

„Eine zeitverzögerte Antwort des Systems erfordert vom Operator, die Antwort des Systems geistig vorwegzunehmen […]. Zeitverzögerungen sollten daher vermieden werden. Wenn dies nicht möglich ist, sollten Beschleunigungen (quickening) oder Vorwert-Anzeigen benutzt werden“ (a.a.O. S. 5). Zeitverzögerungen in der Systemantwort sind ein Charakteristikum der Steuerbarkeit und erhöhen die Arbeitsbelastung, die verringert werden sollte (vgl. a.a.O. S. 6).

Eine Überprüfung der Umsetzbarkeit der Empfehlungen der DIN EN ISO 10075 Teil 2 findet sich in Nachreiner, Meyer, Schomann & Hildebrand (1998), die zu dem Ergebnis kommen, dass nur Personen mit fundiertem arbeitspsychologischen / ergonomischen Fachwissen das Verfahren sinnvoll einsetzen können.

Eine weitere interessante Norm ist die ISO/TS 16071 „Ergonomics of human-system interaction – Guidance on accessibility for human-computer interfaces“, die bereits in Abschnitt 4.3.3 genannt wurde. Sie enthält einen Leitfaden bezüglich der Zugänglichkeit in der Mensch-Computer-Interaktion für einen weiten Personenkreis von Benutzern, die in ihrem Handlungsraum physisch und/oder kognitiv eingeschränkt sind. Die ISO/TS 16071 nimmt Bezug auf die schon behandelten Normenreihe 9241 Teile 10 bis 17 und DIN EN ISO 13407.

Die Leistung des Systems, die sich in Effektivität und Effizienz zeigt, und die Zufriedenheit der Benutzer sind wichtige Kriterien, anhand derer die Zugänglichkeit von Systemen und Umgebung in speziellen Kontexten bestimmt werden kann (vgl. ISO 16071, 2003, S. 6). In Bezug auf Antwortzeiten wird im Kern gefordert, dass das Zeitintervall der Benutzereingaben durch die Benutzer angepasst, ggf. auch deaktiviert werden sollte (vgl. a.a.O. S. 12). Des Weiteren sollen den Benutzern bei zeitsensitiven Informationspräsentationen Möglichkeiten der Pause oder des Stoppens gegeben werden (vgl. a.a.O. S. 13).

5 Systemantwortzeiten von Anwendungssystemen

5.1 Grundlagen und Modellierungsaspekte

5.1.1 Definition

Informationssysteme bestehen aus zwei wesentlichen Komponenten: Zum einen aus dem automatisierten Teil eines Informationssystems, das nach Ferstl & Sinz (2001, S. 4) als Anwendungssystem bezeichnet wird. Dieser besteht aus Hardware-, Software- und Netzwerkkomponenten sowie den zu verarbeiteten Daten. Die zweite Komponente der Informationssysteme ist im Sinne der Mensch-Computer-Interaktion der Benutzer. Während in den vorangegangenen Kapiteln ausführlich auf den Benutzer und seine Wahrnehmung der Systemantwortzeiten eingegangen wurde, wird in diesem Kapitel der Fokus auf die technischen Komponenten gelegt.

Werden die einzelnen Komponenten eines Anwendungssystems genauer betrachtet, so wird die Hardware immer leistungsfähiger und die Software immer umfangreicher. Gleichzeitig fand in den letzten Jahrzehnten eine zunehmende lokale als auch globale Vernetzung von Anwendungssystemen – den verteilten Systemen – statt, wodurch zusätzliche Leistungskapazitäten verfügbar wurden. „Ein verteiltes System ist eine Menge voneinander unabhängiger Computer, die dem Benutzer wie ein einzelnes, kohärentes System erscheinen“ (Tanenbaum & van Stehen, 2003, S. 18). Es kann somit nicht unterstellt werden, dass der Benutzter zwangsläufig beurteilen kann, ob er ein autonomes oder ein verteiltes System nutzt. Dies ist in Bezug auf die Erwartungshaltung von Systemantwortzeiten wichtig. Laut Hüttner et al. (1995, S. 20) stellt die Systemantwortzeit bei autonomen Personal-Computern kein Problem mehr da. Allerdings besteht eine besondere Relevanz der Systemantwortzeiten innerhalb von Netzwerken. Lamport (1978, S. 558) stellte sogar die These auf, dass insbesondere die zeitliche Verzögerung in verteilten Systemen das entscheidende Charakteristikum gegenüber autonomen Einzelsystemen sei.

Daher gilt es die unterschiedlichen vernetzten Systeme, deren Einsatzkontexte und deren Charakteristika in Bezug auf das Systemantwortzeitverhalten und die damit

verbundene Erwartungshaltung und Auswirkungen auf die Benutzer näher zu betrachten. Dix (2003, S. 332 f.) unterscheidet dies in einer Matrix in zweierlei Hinsicht räumlich. Zum einen in lokale und globale Netze und zum anderen nach der Einsatzart, ob diese ortsgebunden oder veränderbar sind (vgl. Abbildung 14). Anhand dieser Klassen lassen sich die verschiedenen existierenden Netze eingruppieren.

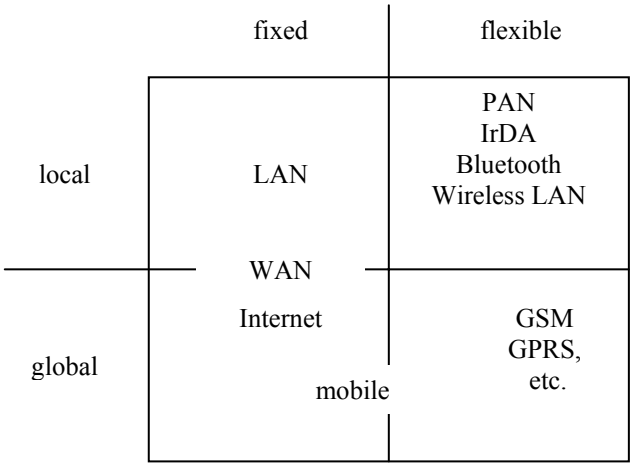


Abbildung 14 Netzklassen (Dix, 2003, S. 332)

Bevor die einzelnen Rechnernetze näher betrachtet werden, müssen weitergehende Grundlagen und analytische Modellierungsaspekte verteilter Systeme gelegt werden. Hierzu werden als nächstes die Qualitätskriterien (Abschnitt 5.1.2) skizziert um dann die Leistungskenngrößen und physikalischen Eigenschaften, die unabdingbare Grenzen darstellen, zu betrachten (Abschnitt 5.1.3). Daran schließen mathematisch-analytische Aspekte der Leistungen der Netzknoten (Abschnitt 5.1.4) sowie der Netzauslastung und Wartezeiten (Abschnitt 5.1.5) an.

Nachdem die Grundlagen gelegt wurden, können die Systemantwortzeiten in den Anwendungssystemen detailliert betrachtet werden. Im Abschnitt 5.2 werden Einzelsysteme anhand von Hard- und Software Komponenten beschrieben. Dies wird im darauf folgenden Abschnitt 5.3 auf verteilte Systeme ausgeweitet. Beginnend mit der Client-Server-Architektur über lokale Netze, Weitverkehrsnetze, bis hin zu mobilen Systemen. Auf das Internet wird dabei im Abschnitt 5.4 näher eingegangen.

5.1.2 Qualitätskriterien

Anwendungssysteme sind in unserem täglichen Leben allgegenwärtig. Die Anforderungen, die an sie gestellt werden, lassen sich in quantitative und qualitative Eigenschaften differenzieren. Während die quantitativen Eigenschaften zähl- und messbare Größen sind, sind die qualitativen Eigenschaften – so genannte nicht-funktionale Eigenschaften – nicht exakt messbar, haben aber eine ebensolche Wichtigkeit. In diesem Abschnitt sollen die qualitativen Eigenschaften als *Qualitäts- bzw. Dienstgütekriterien (engl. Quality of Services, QoS)* beschrieben werden, während die quantitativen Eigenschaften im nächsten Abschnitt vertieft werden (vgl. Abschnitt 5.1.3). Es sei darauf hingewiesen, dass in der Literatur der Begriff der Dienstgüte nicht immer exakt verwendet wird und teilweise, je nach Betrachtungsweise, quantitative und qualitative Eigenschaften vermischt werden.

Ein Anwendungssystem muss, damit es überhaupt genutzt werden kann, verfügbar sein. Die **Verfügbarkeit** (engl. availability) ist somit ein Maß der Systemverfügbarkeit für die Benutzer. Gute Verfügbarkeit resultiert aus einer erhöhten *Zuverlässigkeit* und Robustheit von Hardwarekomponenten und Software. Allerdings ist es wichtiger, dass ein System bei einer fehlerhaften Komponente weiterhin funktioniert, als dass eine Komponente nie ausfällt (vgl. Dyson & Longshaw, 2004, S. 15 f.). Exemplarisch sind in Tabelle 3 prozentuale Verfügbarkeiten mit den daraus resultierenden jährlichen Ausfallzeiten dargestellt.

Verfügbarkeit	**Jährliche Ausfallzeiten eines Anwendungssystems**		
	Tage	**Stunden**	**Minuten**
99 %	3	15	36
99,9 %		8	45,6
99,99 %			52,56
99,999 %			5,26

Tabelle 3 Verfügbarkeit und resultierende Ausfallzeiten

Es ist evident, dass eine sinnvolle Abwägung zwischen den zusätzlichen Kosten – vermutlich exponentiell steigend – und einer möglichst geringen Ausfallfallzeit und deren Nutzen erforderlich ist. Nach Menascé, Almeida & Dowdy (2004, S. 16) kann die Nichtverfügbarkeit von Anwendungssystemen im E-Commerce-Bereich einen Kundenverlust implizieren, während Schneider (1998b, S. 144) die Konsequenzen der Nichtverfügbarkeit sogar zum Schaden für Leben und Besitz ausweitet. Demnach kann auch eine extrem lange Systemantwortzeit, bei der der Benutzer nicht mehr gewillt ist auf das Anwendungssystem zu warten, mit der Nichtverfügbarkeit eines Anwendungssystems gleichgesetzt werden.

Die **Leistung** (engl. performance) eines Systems ist nach der Verfügbarkeit ein entscheidendes Kriterium und eng an die Systemantwortzeiten, die durch quantitative und physische Eigenschaften (vgl. Abschnitt 5.1.3) bedingt sind, gekoppelt. Hierzu gehören unter anderem Auslastung, Datendurchsatz und Datenbankanbindung. Dyson & Longshaw (2004, S. 16) führen aus, dass nicht individuelle gute Leistung, sondern eine konstant gute Leistung ein gutes Anwendungssystem charakterisiert. Es lässt sich somit festhalten, dass für eine gute Systemleistung die Streuung der Antwortzeiten möglichst gering zu halten ist.

Unter **Skalierbarkeit** (engl. scalability) wird die Effizienz eines gut funktionierenden Anwendungssystems bei einer größer werdenden Auslastung verstanden (vgl. Stein, 2004, S. 174). So muss sichergestellt werden, dass die Systemantwortzeit gleich bleibend gut bleibt, auch wenn die Benutzeranfragen im zeitlichen Verlauf variieren oder ansteigen (vgl. Dyson & Longshaw, 2004, S. 17). Abbildung 15 zeigt exemplarisch System A als ein mit der Systemlast gut skaliertes System, während dies bei System B nicht zutrifft und die Systemantwortzeit mit steigender Systemlast ab einem Punkt exponentiell steigt.

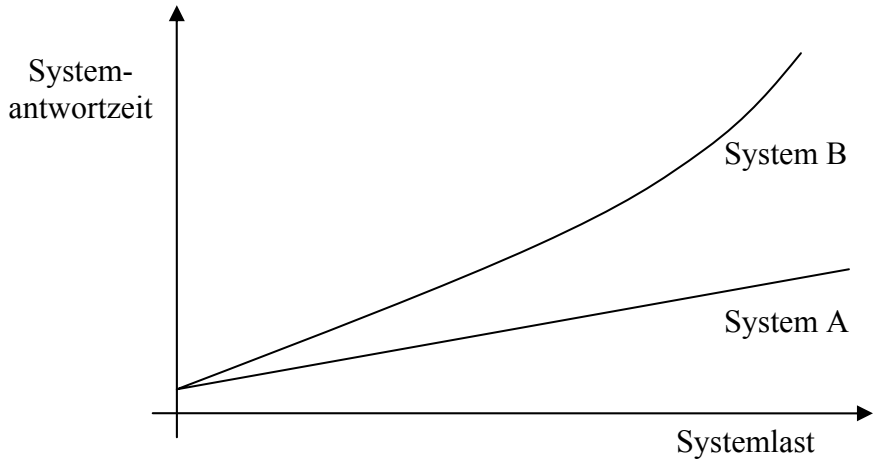


Abbildung 15 Skalierbarkeit von Anwendungssystemen (vgl. Menascé, Almeida, Dowdy, 2004, S. 20)

Die **Sicherheit** (engl. security) eines Anwendungssystems besteht nach Pfleeger & Pfleeger (2003, S. 10 f.) aus den drei Komponenten: Vertraulichkeit, Integrität und Verfügbarkeit. Ein System kann allerdings nie zu 100 % sicher sein. Somit gilt es, das System möglichst gut abzusichern, so dass die Funktionalität und die Einhaltung der Sicherheitsrichtlinien gewährleistet ist (vgl. Dyson &Longshaw, 2004, S. 17). Es muss somit sichergestellt werden, dass trotz der durchzuführenden Sicherheitsmechanismen keine signifikante Verschlechterung der Systemantwortzeit eintritt.

Die **Kompatibilität** (engl. compatibility) und **Portabilität** (engl. portability) beziehen sich sowohl auf Software als auch auf Hardware. In Bezug auf Software sind damit die Kompatibilität von verschiedenen Versionen untereinander sowie die Übertragbarkeit auf unterschiedliche Systemumgebungen gemeint (vgl. Abts & Mülder, 2004, S. 85 f). Portabilität bezeichnet die Migrierung eines bestehenden Anwendungssystems auf neue Hardware-Komponenten oder Betriebssysteme (vgl. Dyson & Longshaw, 2004, S. 18). Die Erweiterung eines Systems um einzelne Komponenten wirkt sich damit auf das Leistungsverhalten des Systems aus.

Im Zuge einer Migrierung muss bedacht werden, dass Systemantwortzeiten von den ursprünglichen Werten signifikant abweichen können und nicht mit den gewohnten Erwartungswerten der Benutzer übereinstimmen. Es sollte daher sichergestellt werden, dass die durchschnittlichen Antwortzeiten auf bestehendem Niveau beibehalten

bzw. verbessert werden, Schwankungen möglichst gering gehalten und die Benutzer im Vorwege über die Systemveränderungen informiert werden.

Die **Handhabbarkeit** (engl. manageability) eines Anwendungssystems erfordert eine stetige Überwachung der Funktionsfähigkeit und unmittelbare Anpassung des Systems an neue Gegebenheiten, um die Zuverlässigkeit des Gesamtsystems sicherzustellen (vgl. Meinel & Sack, 2004, S. 232). Dyson & Longshaw (2004, S. 17) weisen darauf hin, dass die Handhabbarkeit nur schwer zu definieren ist, weil die Theorie von einer umfassenden Überwachung aller Parameter ausgeht, bei der sich die unmittelbare Optimierung des Systems am Benutzerverhalten orientiert. Dies lässt sich in der Praxis aber nur schwer realisieren. Zum einen müssen aus der Informationsflut der Überwachungsparameter die wichtigsten herausgefiltert werden. Zum anderen sind die Systeme derart komplex, dass aus der Vielzahl von möglichen Handlungsalternativen die richtigen gewählt werden müssen.

Da Systemantwortzeiten eine direkte Auswirkung auf die Handhabbarkeit von Anwendungssystemen haben, gilt es, diese zu überwachen und system-, bzw. komponentenbasierte Grenzwerte zu definieren, die bei einer Über- oder Unterschreitung Hinweismeldungen erzeugen oder Anpassungsmechanismen starten.

Die genannten quantitativen Eigenschaften sind die Kernpunkte der zu beachtenden Leistungsaspekte von Anwendungssystemen. In der Literatur finden sich teilweise auch noch eine Reihe von Erweiterungen der Leistungsaspekte mit Begriffen wie Integrität, Wartbarkeit, Flexibilität usw., die im Rahmen dieser Arbeit aber nicht näher betrachtet werden sollen.

5.1.3 Leistungskenngrößen und physikalische Eigenschaften

Leistungskenngrößen, als quantitative Eigenschaften, sind messbare Größen, die sowohl benutzerbezogen, als auch technikbezogen sein können. Benutzerbezogene Leistungskenngrößen wurden vom American National Standards Institute (ANSI) in der Referenz ANSI X3.102 (1992) als Leistungs-Rahmenwerk für Datenkommunikationssysteme spezifiziert (vergleiche Tabelle 4). Das Modell impliziert einen verbindungsorientierten Dienst zwischen Benutzern, in dem die Funktionen des Zugriffs, der Übertragung von Benutzerdaten und der Abkoppelung des Dienstes anhand der Kriterien Geschwindigkeit, Korrektheit und Zuverlässigkeit bewertet werden.

<table>
<tr><th>Kriterium / Funktion</th><th>Geschwindigkeit</th><th>Genauigkeit</th><th>Zuverlässigkeit</th></tr>
<tr><td>Zugriff</td><td>Zugriffsdauer</td><td>Wahrscheinlichkeit des falschen Zugriffs</td><td>Wahrscheinlichkeit des verweigerten Zugriffs oder einer Zugriffsunterbrechung</td></tr>
<tr><td rowspan="2">Übertragung von Benutzerdaten</td><td rowspan="2">Datenübertragungsrate</td><td>Wahrscheinlichkeit eines Block-, Bitfehlers und der falschen Zustellung</td><td>Verlustwahrscheinlichkeit eines Blocks oder Bits</td></tr>
<tr><td colspan="2">Wahrscheinlichkeit für verweigerte Übertragung</td></tr>
<tr><td>Abkoppelung</td><td>Abkoppelungsdauer</td><td colspan="2">Wahrscheinlichkeit für verweigerte Abkopplung</td></tr>
</table>

Tabelle 4 Benutzerbezogene Leistungskenngrößen (vgl. ANSI X3.102, S. 4)

Dieses Modell verdeutlicht anhand der drei Zeitkomponenten der Geschwindigkeit und den Wahrscheinlichkeiten des Fehlerauftrittes, die direkte Auswirkung auf die Systemantwortzeit und die Verfügbarkeit des Anwendungssystems.

Werden die Leistungskenngrößen detailliert betrachtet, so gilt es zwischen Leistungsparametern und Leistungsschwankungen zu differenzieren. Als Leistung werden in der Literatur Bandbreite, Datenrate, Durchsatz, Antwortzeit und die Laufzeit zusammengefasst. Als Leistungsschwankungen werden Jitter, Fehlerrate, Latenzzeit und eine zu definierende Garantie bezeichnet.

Die *Bandbreite* wird gelegentlich irrtümlich mit dem Durchsatz gleichgestellt. Als *Durchsatz* wird die tatsächliche Menge an fehlerfreier Übertragung von Nutzeinheiten – *Datenrate* in Bits pro Sekunde (bps) – definiert, während die *Bandbreite* die technisch maximal mögliche Datenrate angibt (vgl. Meinel & Sack, 2004, S. 223 f.). Die Datenrate ist abhängig von der physikalischen Eigenschaft des jeweiligen Mediums. Tabelle 5 zeigt eine Zusammenstellung der aktuellen Netzwerkarten mit den jeweils maximal möglichen Distanzen, Bandbreiten und der daraus resultierenden Latenzzeit (Verzögerungszeit). Es wird zwischen kabelgebundenen und kabellosen Verbindungen differenziert. Die Vermittlungszeit berechnet sich nach der Formel in Gray & Reuter (1993, S. 58), unter den Annahmen der Lichtgeschwindigkeit im Glasfaserkabel von $C_{m,g} = 194.865\ km/s$, in der Luft von $C_{m,l} = 299.792\ km/s$ und einer nur durch das Übertragungsmedium beeinflussten Latenz, wie folgt:

$$\ddot{U}bermittlungszeit(NachrichtenBits) = \frac{Entfernung}{C_m} + \frac{NachrichtenBits}{Bandbreite}\text{sec} \qquad (5.1)$$

Netzwerkart	Beispiel	Einheit	Distanz	Latenzzeit	Bandbreite (max)
Kabelgebunden:					
Microprozessorsystem	CPU	Platine	0,1 m	0,513 ns	
Microprozessor-Cluster	CPUs	System	1 m	5,13 ns	1 Gbps
Cluster	Vernetzte Computer		100 m	0,51 µs	1 Gbps
LAN	Ethernet	Gebäude	1 km	5,1 µs	1 Gbps
MAN	ATM	Stadt	100 km	0,51 ms	2,5 Gbps
WAN	IP Routing	Kontinent	10.000 km	51,3 ms	10 Gbps
Kabellos:					
WPAN	Bluetooth	Raum	30 m	0,1 µs	2,2 Mbps
WLAN	WiFi	Gebäude	1,5 km	5,0 µs	54 Mbps
WMAN	WiMAX	Stadt	50 km	0,16 ms	109 Mbps
WWAN	GSM	Kontinent	Zellengröße		2 Mbps

Tabelle 5 Eigenschaften von Rechnernetzen (in Anlehnung an: Gray & Reuter, 1993, S. 59; Meinel & Sack, 2004, S. 197; Coulouris, Dollimore & Kindberg, 2005, S. 70)

Die zur Verfügung stehende Bandbreite des Übertragungsmediums impliziert damit eine distanzabhängige Latenzzeit, die die Systemantwortzeit beeinflusst. Bei der zur Verfügung stehenden Bandbreite gilt es zwischen fester Mindestbandbreite sowie variabler und verfügbarer Bandbreite zu differenzieren. Diese hängt von den von den Benutzern zu tragenden Kosten, den jeweiligen Anwendungsbedarfen und den freien Netzkapazitäten ab. Bei verbindungslosen Diensten wie dem Internet können Bandbreite, Verzögerungen und Jitter nicht garantiert werden, weil die zu übermittelnden Daten von der zur Verfügung stehenden Bandbreite abhängig sind. Bei einer hohen Aus- bzw. Überlastung der Bandbreite nehmen die Verzögerungen stetig zu, so dass eine Übermittlung nur nach und nach so gut wie möglich erfolgen kann (vgl. Winzerling, 2001, S. 25 ff).

Die Leistungsschwankungen können unterschiedliche Ursachen haben (vgl. Abb. 16). Sie werden durch die Benutzer aber nur als Gesamtverzögerung bemerkt, so dass sie die genaue Ursache nicht bestimmen können. Zum einen gibt es die schon genannte Laufzeitverzögerung als unabdingbare physikalische Eigenschaft des jeweiligen Mediums. Zum anderen können Wartezeiten durch die Verarbeitungsleistung und -dauer in den Netzknoten (vgl. Abschnitt 5.1.4) oder durch die Auslastung der Netze (vgl. Abschnitt 5.1.5) entstehen.

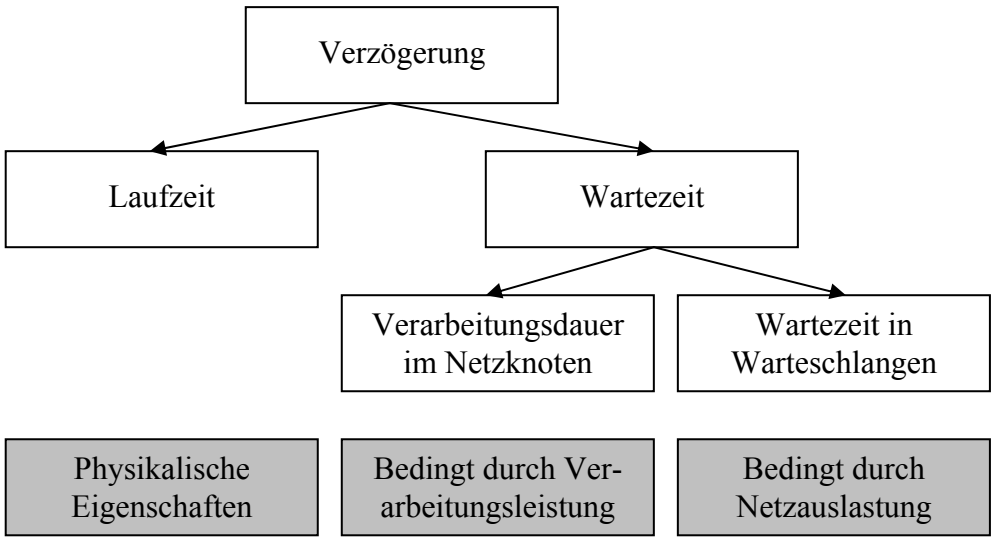


Abbildung 16 Verzögerungszusammensetzung (vgl. Stein, 2004, S. 171)

Gemäß der angewendeten Formel (5.1) lässt sich die verbindungsorientierte Übermittlungszeit allgemein spezifizieren als:

$$\ddot{U}bermittlungszeit(NBits) = Verzögerung + \frac{NBits}{Datenrate} \tag{5.2}$$

Die Verzögerung setzt sich wie folgt zusammen:

$$Verzögerung = (Laufzeit + Wartezeit) = (\frac{Entfernung}{C_m} + t_{Verarbeitung} + t_{Warteschlange}) \tag{5.3}$$

Damit lässt sich die verbindungsorientierte Übermittlungszeit in Abhängigkeit der Nachrichten-Bits-Größe (NBits) definieren:

$$\ddot{U}bermittlungszeit(NBits) = (\frac{Entfernung}{C_m} + t_{Verarbeitung} + t_{Warteschlange}) + \frac{NBits}{Datenrate} \tag{5.4}$$

Neben der *Latenzzeit* sind *Jitter* eine weitere Kenngröße der Leistungsschwankung. Sie geben die maximalen Schwankungen der Verzögerung an und sind daher für den Benutzer – insbesondere bei Multimedia-Daten – von besonderer Wichtigkeit. Coulouris, Dollimore & Kindberg (2005, S. 49) nennen als Beispiel Audiodaten, die in unterschiedlichen Zeitintervallen gespielt und somit verzerrt werden.

Jitter lassen sich nach ihrem Verbindungsverhalten unterscheiden in asynchron, synchron und isochron. Beim asynchronen Verhalten ist die Übermittlungszeit der Datenpakete zwischen dem Sender und Empfänger nicht spezifiziert und variiert stark. Diese im Extremfall sehr hohen Zeiten sind aber für bestimmte Daten akzeptabel und nicht problematisch, z.B. beim E-Mail-Versand. Beim synchronen Verhalten sind im Gegensatz dazu die oberen Grenzwerte der Verweildauer festgelegt. Diese Mindestanforderung – wenn auch nicht hinreichend – gilt bei der Sprach- und Bilderübertragung. Dagegen ist beim isochronen Verhalten die Verweildauer aller Pakete gleich (vgl. Meinel & Sack, 2004, S. 224 f.; Stein, 2004, S. 173 f.).

In Bezug auf Mediendaten gibt es noch eine weitere Verzögerungsgröße – den *Skew*. Es handelt sich hierbei um die synchrone Übertragung von Daten, bei der ein Versatz auftreten kann. Bei den parallelen Datenströmen kann es somit zu unterschiedlichem Jitter kommen. Daher ist der maximale Skew die Summe aus allen maximalen Jitter (Mühlhäuser, 2002, S. 860).

Neben dem Jitter sind noch *Fehlerraten* als Leistungsschwankungen näher zu betrachten. Fehlerraten bestimmen eine maximal zugesicherte Wahrscheinlichkeit für den Datenverlust oder die Datenverfälschung bei der Übertragung. Hierbei gilt es für die jeweilige Anwendung eine Restfehlerwahrscheinlichkeit zu spezifizieren, bei der ein Fehlererkennungs- und Korrektur-Verfahren im sinnvollen Kosten/Nutzen-Verhältnis steht (vgl. Meinel & Sack, 2004, S. 225; Stein, 2004, S. 169).

Um einen ordnungsgemäßen Betrieb von Systemen und Anwendungen sicherzustellen, werden Verträge zwischen den Dienstanbietern und den Dienstnutzern geschlossen – die Service Level Agreements (SLA). Die gegebenen Garantien umfassen Leistungsparameter, die zu festgelegten Werten, immer in Verbindung mit einer zu bestimmenden Wahrscheinlichkeit, dem Benutzer vom Dienstanbieter garantiert werden. Als Grundlage hierzu dienen die oben spezifizierten Größen wie Bandbreite, Durchsatz, Verfügbarkeit und die bedingten Leistungsschwankungen als wichtige Leistungskenngrößen (vgl. Menascé et al., 2004, S. 44).

5.1.4 Leistungen der Netzknoten

Netzknoten sind Rechnersysteme, die aus Hardware- und Software-Komponenten bestehen, die in Abschnitt 5.2 näher betrachtet werden. In diesem Abschnitt sollen mathematisch-analytische Ansatzpunkte aufgezeigt werden, die in einem solchem System als Modellierungsgrundlage dienen. Zur Vertiefung der mathematischen Modellierung wird für stochastische Grundlagen auf Hübner (2002) und in Bezug auf Rechnersysteme auf Menascé et al. (2004, S. 251 ff.) und Bolch, Greiner, de Meer & Trivedi (1998) verwiesen.

Ganz allgemein betrachtet gibt es eingehende Aufträge, die bearbeitet werden und abgeschlossen das System verlassen. Dies lässt sich an einem Bediensystem mit einem eingehenden Datenkanal und einem verarbeitenden Bediener (Server) abbilden (vgl. Abb. 17). Die einzeln ankommenden Aufträge werden mittels der Ankunftsrate λ modelliert, die als Mittel die Zahl von Auftragsankünften je Zeiteinheit angibt. Danach sammeln sich die Aufträge in einer Warteschlange mit der Länge c. Nach dem

F*irst Come First Served* Prinzip (FCFS) und werden dann von einem Bediener *s* (Server) mit der Bedienrate μ bearbeitet. Die Bedienrate gibt die Anzahl der Bedienungen je Zeiteinheit an.

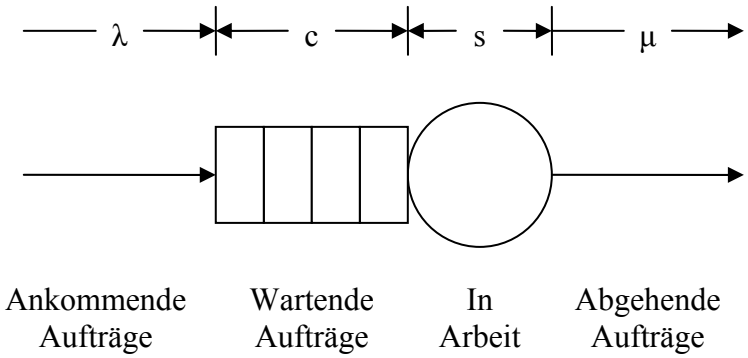


Abbildung 17 Grundlegendes Bedienmodell

Anhand dieses Schemas lässt sich eine allgemeingültige Klassifikation von Bediensystemen entwickeln, die sich aus fünf Komponenten zusammensetzt:

$$A|B|s|c|R$$

Dabei gilt:
- *A*: bestimmt die Art des Ankunftsprozesses
- *B*: bestimmt die Art des Bedienvorganges
- *s*: bestimmt die Anzahl der Bedieneranzahl (server)
- *c*: bestimmt die Warteraumgröße (capability)
- *R*: bestimmt die Bedienreihenfolge

Gemäß den oben genannten Annahmen wird nur ein Bediener zugrunde gelegt und zur Vereinfachung wird der Warteraum als unendlich modelliert, auch wenn in der Realität ein endlicher Warteraum angenommen werden muss. Der Ankunfts- und Bedienvorgange basiert auf einer Markov-Eigenschaft *M*. Dies bedeutet eine weitere rechnerische Vereinfachung, da bei der Markov-Eigenschaft angenommen wird, dass nur eine geringe Anzahl von Einflussgrößen das Zeitverhalten eines Systems beeinflusst. Dies kann bei Netzknoten unterstellt werden. Damit erhalten wir folgendes Bediensystem:

$$M|M|1|\infty$$

Mit Hilfe dieses Modells ist es nun möglich, Kenngrößen zu berechnen. Damit ein System noch arbeiten kann, muss die Auslastung *p* des Systems bestimmt werden. Die Auslastung wird definiert als:

$$p = \frac{Ankunftsrate}{Bedienrate} = \frac{\lambda}{\mu} \tag{5.5}$$

Der nahe liegenden Forderung nach p < 1 muss mit Vorsicht entgegen getreten werden. Die Annahme geht von nur einem Bediener aus und beschreibt seine Beschäftigung zu dem betrachteten Zeitpunkt. Dies führt dazu, dass lediglich bei $\lambda < \mu$ von einer Gleichgewichtsbedingung gesprochen werden kann, während bei $\lambda \geq \mu$ kein Gleichgewicht entsteht, da entweder mehr Aufträge reinkommen, als abgearbeitet werden können, bzw. keine Möglichkeit besteht die Warteschlange zu reduzieren.

Daher ist es interessant, den Ankunftsprozess der Aufträge näher zu betrachten. Es kann davon ausgegangen werden, dass der Ankunftsprozess exponentiell über die Zeit verteilt ist. Unter der Annahme, dass in dem Zeitintervall t genau k Aufträge einkommen gilt:

$$p(k) = \frac{(\lambda t)^k}{k} e^{-\lambda t} \tag{5.6}$$

Diese Verteilung wird auch *Poissonverteilung* genannt. Sie setzt voraus, dass die eingehenden Aufträge aus einer unabhängigen Menge kommen und die Zeit zwischen den Auftragseingängen exponentiell verteilt ist.

In Bezug auf die Systemantwortzeiten ist es daher wichtig, die Leistungskapazitäten der Systeme zu bestimmen, weil es durch ein nicht vorhandenes Gleichgewicht zu zeitlichen Verzögerungen kommen kann, die es im Vorwege zu vermeiden gilt. Anhand des Bedienmodells in Abbildung 17 lässt sich ableiten, dass die Länge der Warteschlange *c* die durchschnittliche Wartezeit *w* bestimmt, während die durchschnittliche Bedienzeit als *E[s]* angegeben werden kann. Daraus ergibt sich eine Verweilzeit (Gesamtantwortzeit) von:

$$EW = w + E[s] \tag{5.7}$$

Diese mittlere Verweilzeit als Erwartungswert *EW*, bzw. die daraus abgeleiteten Quantile sind somit aussagekräftige Leistungsmaße eines Systems. Ferner lässt sich

auch die Streuung angeben. Hierzu benötigen wir die mittlere Kundenanzahl EX_k und die Streuung $StrX_n$ der Kundenanzahl:

$$EX_k = \frac{\lambda / \mu}{1 - \lambda / \mu} = \frac{p}{1 - p} \tag{5.8}$$

$$StrX_n = \frac{\sqrt{\lambda / \mu}}{1 - \lambda / \mu} = \frac{\sqrt{p}}{1 - p} \tag{5.9}$$

Mittels der Formel von Little lässt sich die mittlere Verweilzeit durch die mittlere Kundenanzahl EX_k und die mittleren Ankunftsrate $\overline{\lambda}$ leicht berechnen:

$$EW = \frac{EX_k}{\overline{\lambda}} \tag{5.10}$$

Es gibt natürlich noch eine Reihe weiterer Bediensysteme und Modelle, die im Rahmen dieser Arbeit nicht weiter betrachtet werden können, da dies zu umfangreich werden würde.

5.1.5 Netzauslastung

Bei der Übertragung von Datenpaketen in paketvermittelten Netzwerken können Verzögerungen an unterschiedlichen Stellen auftreten (vgl. Meinel & Sack, 2004, S. 227 ff.). Es gilt hierbei zwischen den verschiedenen Verzögerungen zu differenzieren. Bei der Verarbeitung in den Vermittlungsrechner kann es zu *Verarbeitungsverzögerungen* (d_{proc}) kommen, die heute meist im Mikrosekundenbereich liegen. Proportional zur Anzahl der Datenpakete und abhängig von der jeweiligen Netzauslastung entwickelt sich die *Warteschlangenverzögerung* (d_{queue}), die stark variieren und bei Routern meist im Mikro- bis Millisekundenbereich liegen. Die Schnelligkeit der Verbindungsrechner und die Bandbreite der Verbindung bestimmen die *Versendeverzögerungen* (d_{trans}), die beim Absenden eines kompletten Datenpaketes entstehen. *Laufzeitverzögerungen* (d_{prop}) geben die Übertragungszeit an und sind abhängig von den Eigenschaften des Übertragungsmediums (vgl. Abschnitt 5.1.3). Aus den vier Einzelverzögerungen lässt sich die Gesamtverzögerung (*d*) ermitteln, die insbesondere durch die Warteschlangen- und Laufzeitverzögerungen geprägt ist. Eine geringe Verbindungsbandbreite führt zusätzlich zu einer hohen Gesamtverzögerung.

$$d = d_{proc} + d_{queue} + d_{trans} + d_{prop} \tag{5.11}$$

5.2 Einzelsystem

5.2.1 Hardware

Nachdem nun die Grundlagen und Modellierungsaspekte der Systemantwortzeiten von Anwendungssystemen dargelegt wurden, werden nachfolgend die Systeme detailliert betrachtet. Die Betrachtung wird in diesem Kapitel mit dem Einzelsystem und den Systemkomponenten begonnen und dann sukzessiv auf vernetzte Systeme erweitert.

Die Hauptkomponenten, aus denen sich ein Rechner zusammensetzt, sind in Abbildung 18 skizziert. Das Kernstück besteht aus der Zentraleinheit, die sich wiederum aus Speicher, Cache und Prozessor mit Steuer- und Rechenwerk zusammensetzt. Daran angeschlossen sind Peripheriegeräte für die Ein- und Ausgabe sowie für die Datenspeicherung. Bevor wir das Antwortzeitverhalten der Einzelkomponenten betrachten, sei der Einwand von Raskin (2000) eingeworfen, der beklagt, dass allein das Starten von Anwendungssystemen zu lange dauert und den Benutzer in seiner Interaktion behindert.

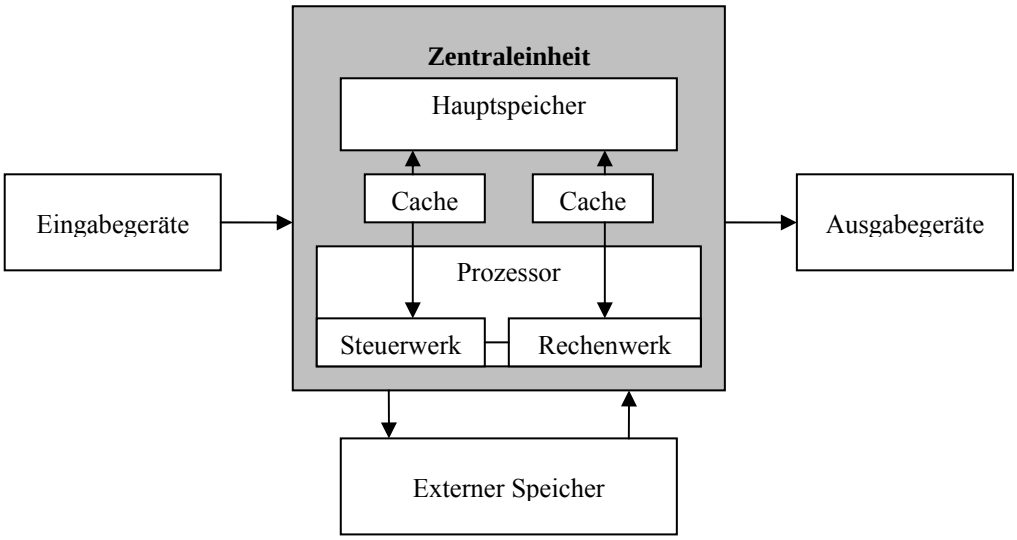


Abbildung 18 Rechnergrundstruktur (in Anlehnung an Abts, Mülder, 2004, S. 38; Mertens, Bodendorf, König, Picot, Schumann, Hess, 2005, S. 14)

Betrachten wir nun die einzelnen Komponenten genauer. Gemäß der Hypothese von Moore (1965) verdoppelt sich die Leistung des Prozessors (Central Processing Unit,

CPU) alle 18 Monate, so dass hierbei davon auszugehen ist, dass seitens der CPU eine weiter stetige Entwicklung für die Zukunft zu erwarten ist. Dies wirft die Frage der Wichtigkeit der CPU-Komponente auf, da durch deren kontinuierliche Leistungssteigerung eine Beeinträchtigung im Sinne der Systemantwortzeiten nicht auftreten dürfte. Mit Bezug auf die Warteschlangentheorie (vgl. Abschnitt 5.1.4) in Bolch et al. (1998, S. 209 ff.) weisen Mißbach et al. (2005, S. 137 f.) darauf hin, dass die Antwortzeit des Systems von der CPU-Last abhängig ist. Die CPU-Last ergibt sich durch die Anzahl der gleichzeitigen Benutzerprozesse. Durch steigende Benutzerprozesse erhöhen sich die Wartezeit und die Auslastung der CPU, die sich dann durch erhöhte Antwortzeiten – auch bei den Benutzern – bemerkbar machen. Abbildung 19 skizziert das Verhältnis von CPU Auslastung zur Antwortzeit. Es zeigt sich, dass lediglich bei einer niedrigen CPU Auslastung ein linearer Zusammenhang besteht, dieser aber mit steigender Last mit einer Auslastung von über 70 % exponentiell steigt. Zur weitergehenden Veranschaulichung wurden Leistungskurven von Systemen mit ein, zwei und vier Prozessoren abgetragen. Es zeigt sich, dass bei mehreren Prozessoren mit steigender Last eine bessere Antwortzeit gegenüber Systemen mit nur einem Prozessor geliefert werden kann.

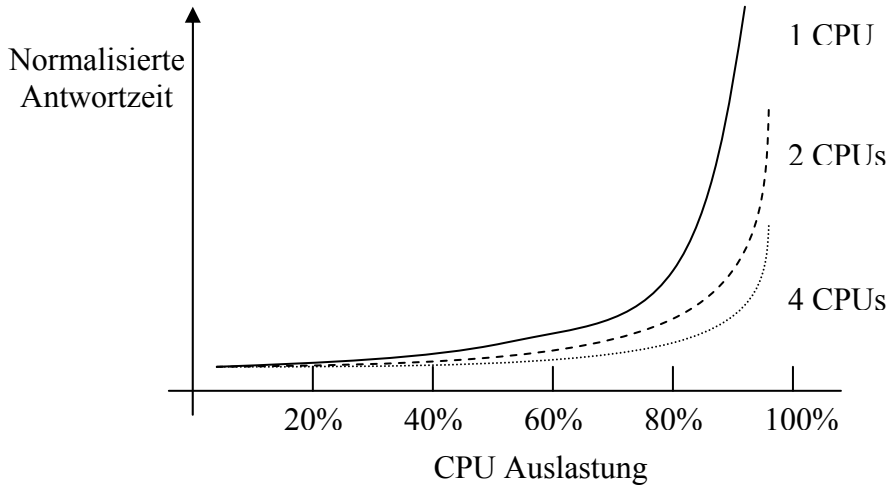


Abbildung 19 CPU Auslastung und Antwortzeit (in Anlehnung an Mißbach, 2005, S. 140)

Nelson & Tantawi (1989) führten Untersuchungen der Systemantwortzeit in parallelen Systemen durch. Sie unterschieden hierbei zwischen Bedienplätzen mit eigener Warteschlange (distributed (D)) oder einer gemeinsamen Warteschlange (centralized (C)). Ferner wird die Auftragsbearbeitung unterschieden zwischen einer an einen

Bedienplatz gebundenen Auftragsbearbeitung (no splitted (NS)) und einer unabhängigen auf mehrere Bedienplätze verteilten Bearbeitung (splitted (S)). Daher kann zwischen vier Modellen der parallelen Verarbeitung unterschieden werden (vgl. Tab. 6). In ihrem Vergleich der Systemantwortzeiten dieser vier Modelle kamen Nelson und Tantawi zu dem Ergebnis, dass die schlechteste Leistung in Systemen mit verteilter Warteschlange und an Server gebundene Auftragsverarbeitung (D/NS) erzielt wurde. Die beste Leistung wurde mit C/NS-Systemen erreicht in der eine gemeinsame Warteschlange für alle Bedienplätze vorgehalten wird und die jeweilige Auftragsbearbeitung bei der Planung an nur einen Server vergeben wird und nicht parallel auf mehrere gleichzeitig. Als Grund hierfür geben sie die geringere Anzahl von leistungsstärkeren Prozessoren an und dem Ausschluss von Leistungseinbußen durch einzelne Warteschlangen.

	Auftragsbearbeitung	
Warteschlange	An einen Server gebunden	Parallel auf verschiedenen Servern
Eigene für jeden Server	D/NS	D/S
Eine gemeinsame für alle	*C/NS*	C/S

Tabelle 6 Modelle der parallelen Verarbeitung

Bei der Gestaltung von System ist immer zu bedenken, dass durch die einzelnen Komponenten systembedingt immer eine minimale Antwortzeit gibt, die nicht unterschritten werden kann. So werden Prozessoren immer schneller, aber diese Leistungssteigerung lässt sich nicht auf Laufwerke übertragen (vgl. Williams & Smith, o.J.). Während der Hauptspeicher mit dem Cache in der Zentraleinheit mit dem Prozessor zusammengefasst ist, sind Laufwerke als externe Speicher zwar in ihrer Kapazität vergrößert worden, aber die Umdrehgeschwindigkeit liegt z. Zt. physisch bedingt bei 7.200 Umdrehungen in der Minute. Bei Ein- und Ausgabegeräten ist die Systemantwortzeit ebenfalls durch das jeweilige Medium bedingt. Die Eingabe erfolgt mittels Tastatur und / oder Maus. Hier gibt es die Empfehlung, dass Handlungen wie Typing, Cursor-Bewegung und Maus-Auswahl nur 50-150 Millisekunden betragen sollten (vgl. Shneiderman & Plaisant, 2005, S. 473). Die Ausgabe mittels Bildschirm erfolgt heutzutage im Millisekundenbereich.

5.2.2 Software

Seit der Entwicklung der Computer besteht die allgemeine Haltung, dass die nächste Hardware-Generation bedeutende Leistungsverbesserungen mit sich bringt, so dass sich um die Leistung der Software keine Gedanken gemacht werden muss. Allerdings ermöglicht die leistungsstärkere Hardware auch komplexere Software, so dass auch deren Effizienz bedacht werden muss (vgl. Smith, 1993, S. 527). Dix (1987) gibt zu bedenken, dass man nicht dem Mythos unterliegen darf, dass es die unendlich schnelle Maschine gibt, so dass sich Softwareentwickler nicht um die Leistung kümmern müssten.

Der komplexe Aufbau der Software zeigt sich in dem schematischen Schichtenmodell (vgl. Abb. 20). Auf die Hardware setzt das Betriebssystem auf, das die Prozessorverwaltung steuert und damit direkten Einfluss auf die systembedingten Parameter wie Durchsatz, Antwortzeit und Prozessorauslastung hat. Die Middleware fungiert als Dienstleistungssoftware, die einen Datenaustausch zwischen verschiedenen, sonst entkoppelten heterogenen Softwarekomponenten ermöglicht. Das Anwendungsprogramm stellt die Schnittstelle zum Benutzer her. Durch das Zusammenwirken aller Hard- und Softwarekomponenten ist es erforderlich, den gesamten Software-Lebenszyklus zu betrachten.

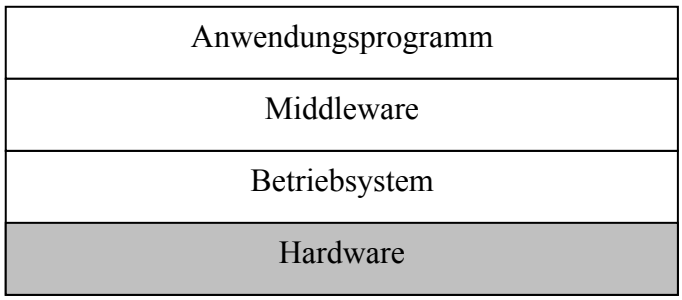


Abbildung 20 Systemaufbau (in Anlehnung an Abts & Mülder, 2004, S. 35)

Die Leistung von Systemen besteht aus zwei Dimensionen: Ansprechempfindlichkeit (Systemantwortzeit oder Durchsatz) und Skalierbarkeit. Smith & Williams (2002, S. 10 f.) weisen in diesem Zusammenhang auf verfestigte Mythen hin, die dieses behindern. So wird irrig angenommen, dass für die Leistung erst dann etwas getan werden kann, wenn auch etwas gemessen werden kann. Außerdem würden angeblich hand-

habbare Leistungen zu viel Zeit benötigen und die Modelle wären zu komplex und zu teuer. Dieses stimmt nicht, denn seitens der Software, wie auch der Hardware, ist es schon frühzeitig möglich die erforderlichen Leistungsparameter im Entwicklungsprozess zu spezifizieren. Dadurch lässt sich die Gesamtprojektzeit reduzieren und nachträgliche Verbesserungen vermeiden. Hierzu schlägt Smith (1993, S. 509 ff.) die Methodik des Software Performance Engineerings (SPE) vor. Der Fokus liegt hierbei in der Einhaltung von Leistungsparametern eines Systems – wie Antwortzeit und Durchsatz – schon in der Entwicklung und über den ganzen Lebenszyklus. Es ist ein iterativer Prozess, der darauf basiert, dass Modelle für aussagekräftige Leistungsvorhersagen geschaffen werden und Werkzeuge vorhanden sind, die Studien und Methodenanwendung in der Systementwicklung ermöglichen. Hier liegt wiederum auch die Schwierigkeit, verlässliche Aussagen über die Anforderungen an zukünftige Systeme und Umgebungen zu spezifizieren. Die Abwesenheit von Problemen bedeutet nicht, dass diese nicht vorhanden sind. Software-Modelle leisten somit ihren Beitrag, schon im Entwicklungsstadium architektonische und designspezifische Leistungsprobleme aufzudecken (vgl. Smith & Williams, 2002, S. 72).

Die software-ergonomische Gestaltung der Systeme im Bezug auf die Systemantwortzeiten hat somit ebenfalls eine hohe Bedeutung. Das Interaktiondesign stellt die einzige Möglichkeit dar, die Benutzer über die Systemantwortzeiten und deren Verzögerungen zu informieren. Trotz dieser wesentlichen Bedeutung gibt es kaum Untersuchungsergebnisse und Gestaltungsempfehlungen. Meyer, Shinar, Bitan & Leiser (1996) kommen in ihren Untersuchungen zu dem Ergebnis, dass die Benutzer dynamische und grafische Fortschrittsanzeigen gegenüber einfachen statischen Hinweisen vorziehen. Durch dynamische Restwartezeitanzeigen wird die Verarbeitungszeit kürzer wahrgenommen und eine höhere Zufriedenheit der Benutzer erreicht. Laut Nielsen (1993, S. 135 ff.) sind solch detaillierte Rückmeldungen erst ab 10 sec Wartezeit nötig. Bei Zeiten zwischen 2 und 4 Sekunden reicht ein einfacher Beschäftigungshinweis aus, wie z.B. eine Veränderung des Maus-Zeigers vom Pfeil zur Sanduhr. Als Grund gibt Nielsen an, dass die Benutzer sonst angesichts der schnell wechselnden Informationen überanstrengt und gestresst werden würden.

5.3 Verteilte Systeme

5.3.1 Client-Server-Architektur

Verteilte Systeme haben die Charakteristik, dass mehrere Systeme zusammenwirken. Daher soll an dieser Stelle erst einmal das Paradigma der Client-Server-Architektur eingeführt werden, um im Anschluss daran die einzelnen Netzarten differenziert zu betrachten.

Client-Server-Architekturen setzten sich in den 1990er Jahren durch und ersetzten die Mainframesysteme. Die Architektur basiert auf dem einfachen Prinzip, dass der Benutzer mittels eines Clients bei den Servern angebotene Dienste abruft (vgl. Abb. 21). Hierzu sendet der Client eine Anfrage (*request*) an den Server, der eine Antwort (*reply*) zurücksendet. Die Clients dienen somit lediglich der Benutzerinteraktion und Datenpräsentation. Der Vorteil dieser Client-Server-Systeme liegt gegenüber den alten Mainframesystemen in der Herstellerunabhängigkeit der eingesetzten Hardware und Software sowie der dadurch ermöglichten größeren Informationsbasis und schnelleren Systemantwortzeiten. Die meisten Dienste des Internets (vgl. Abschnitt 5.4) basieren auf der Client-Server-Architektur (vgl. Schwartz, 2001, S. 96 f.; Mertens et al., 2005, S. 40 f.)

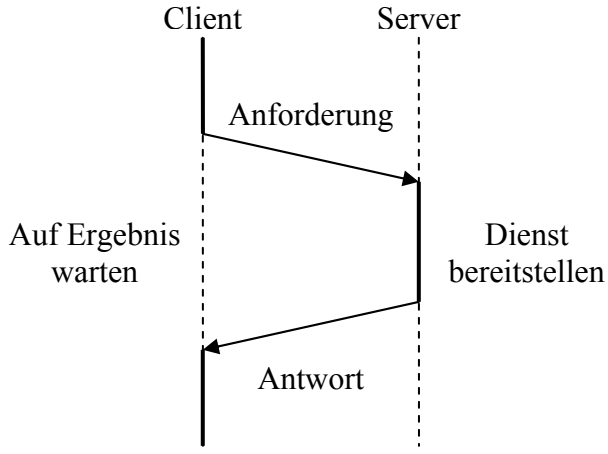


Abbildung 21 Allgemeine Client-Server-Kommunikation

Die Client-Server-Struktur kann durch eine Vielzahl von Clients und Servern gekennzeichnet sein. So können Server, die spezielle Dienste haben, von anderen Ser-

vern aufgerufen werden. Hierbei kann es zu Wartezeiten und verlängerten Systemantwortzeiten kommen. Woodside (1993; S. 394 ff.) betrachtet dieses Problem und benennt die Software als den Flaschenhals, da die Bedienzeit stark lastabhängig ist. Für Field, Harrison & Parry (1998, S. 75 ff.) ist die Systemantwortzeit in Client-Server-Systemen ein Schlüsselmerkmal. Client, Netzwerk und Server gelten als Verzögerungsquellen. Sie weisen darauf hin, dass die Art der Nachrichtenlängenverteilungen bedeutend für den Durchsatz im Ethernet ist und gehen davon aus, dass sich die Antwortzeitdichte für willkürliche Nachrichtenlängenverteilung approximierbar ist. Maccabee (1996) führt die *Ende-zu-Ende* Methode bei der Bestimmung von Systemantwortzeit ein, in dem er argumentiert, dass nicht nur die technische Sichtweise betrachtet werden darf, sondern vielmehr die Benutzersicht. Bei Mainframe-Systemen war das Ziel 95% aller Transaktionen des *Customer Information Control System* (CICS) innerhalb von 3 Sekunden Systemantwortzeit abzuarbeiten. Dieses Ziel galt es auch in der Client-Server-Umgebung einzuhalten. Maccabee stellte fest, dass die geographische Lage von Servern wichtig ist. Server, die über ein lokales Netz (LAN) angeschlossen waren, waren schneller als geografisch verteilte Server. Ursache war die langsame Verbindung zu dem entfernten Standort.

5.3.2 Lokales Netz

In der lokalen Rechnervernetzung hat sich die LAN-Technologie durch die große Bandbreite von mittlerweile bis zu 1 Gbps (vgl. Abschnitt 5.1.3), durchgesetzt. Ein früher Übersichtsartikel über die Leistungsprobleme der LANs lässt sich in Bux (1984) finden. Bux betrachtet Aspekte des Durchsatzes und der dadurch bedingten Verzögerungen bei folgenden Standardverfahren: Carrier Sense Multiple Access/Collision Detect (CSMA/CD), Token Ring und Token Bus. Beim *CSMA/CD-Verfahren* zeigte sich, dass die Verzögerung exponentiell zum Durchsatz steigt und somit bei steigender Geschwindigkeit die Effizienz des CSMA/CD-Verfahrens signifikant fällt. Dagegen führte die Erhöhung des asymmetrischen Verkehrs beim Token Bus zu einer leichten Reduzierung der Verzögerung. Die Ursache hierfür liegt in dem mitgeführten Overhead des Tokens, der kleiner wurde.

Durch den Einsatz von Client-Server-Systemen über LAN-Verbindungen in Unternehmen sollten die Systemantwortzeiten möglichst gering gehalten werden, damit die Benutzer bei ihrer Arbeit nicht gestört werden. Doherty & Thadani (1982) weisen darauf hin, dass sich die Dauer der Systemantwortzeit auf die Dauer der Benutzerantwortzeit auswirkt. Je schneller die Systemantwortzeit, desto schneller war die Reaktionszeit der Benutzer (vgl. Abb. 22). Von daher wäre es interessant zu sehen, wie es um die Systemantwortzeiten lokaler, verteilter Anwendungssysteme bestimmt ist, doch leider ließen sich hierzu keine Ergebnisse finden.

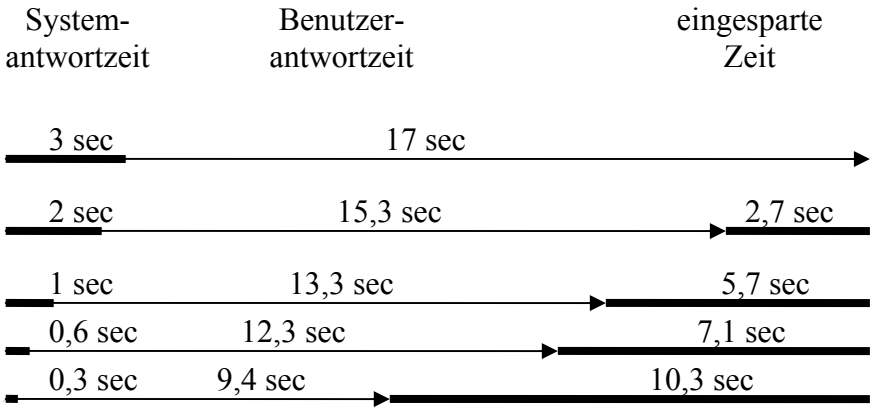


Abbildung 22 Systemantwortzeit im Verhältnis zur Benutzerantwortzeit (vgl. Doherty & Thadani, 1982)

5.3.3 Lokale Funknetzwerke

Die Nutzung lokaler Funknetze und -verbindungen erfreut sich seit den vergangenen Jahren sowohl im Privat- als auch im Geschäftsbereich einer steigenden Beliebtheit. Der Einsatz der sichtbedingten Infrarot-Verbindung (IrDA) und des weiterentwickelten Bluetooth Funknetzes (IEEE[7] 802.15) ist nur auf kurzer Distanz als Ersatz herkömmlicher Kabelverbindungen zwischen Geräten möglich (vgl. Stein, 2004, S. 240). Dagegen ermöglichen drahtlose lokale Netzwerke (Wireless LAN, WLAN) nach dem IEEE 802.11 Standard einen größeren Mobilitätsraum. Ein entscheidendes Charakteristikum des WLANs ist die Bandbreite und der Durchsatz. So liegt die maximale Bandbreite zurzeit bei 54 Mbit/s. Der tatsächlich erreichbare Durchsatz hängt allerdings von der Entfernung zum Funksender und der Datenrate ab. Lenk (2005, S.

[7] Institute of Electrical and Electronics Engineers, Berufsverband mit Standardisierungsgremien

314) zeigt, dass bei größeren Entfernung niedrigere Datenraten zu bevorzugen sind, da diese einen konstanteren Datendurchsatz ermöglichen.

Aufgrund des durch die Entfernung bedingten Datendurchsatzes kann es zu variierenden netzbedingten Laufzeiten und damit zu variierenden Systemantwortzeiten kommen. Ferner gilt es zu bedenken, dass es – gewollt oder ungewollt – bei einem Verlassen des Funknetzes zu einem Verbindungsabbruch kommt. Dies führt bei einer vom Funknetz abhängigen Tätigkeit zu deren Abbruch, so dass der Vorgang erneut durchgeführt werden muss. Durch den Wechsel in den asynchronen Modus kann es bei einer erneuten Verbindung mittels Funknetz zu erforderlichen Synchronisationszeiten und Datenabgleichen kommen, die die erneute Arbeitsaufnahme wesentlich verzögern können. Hier gilt es die Benutzer über die Möglichkeit der Steuerbarkeit und die Transparenz des Systemzustands zu informieren.

5.3.4 Weitverkehrsnetze

Unter Weitverkehrsnetzen werden Metropolitan Area Networks (MAN) und Wide Area Networks (WAN) verstanden. Sie sind großflächigere Netze als LANs und erfordern einen eigenen Netzbetreiber (vgl. Stein, 2004, S. 161). Beispielsweise hat ein MAN als regionales Netz eine Ausdehnung von ca. 100km (vgl. Tab. 5, S. 46). Aufgrund der größeren Netzdistanzen kommt es zu höheren Latenzzeiten, die beachtet werden müssen. Als Verzögerungsbeispiel nennt Tanenbaum (2003, S. 28) die Uhrensynchronisation, die im LAN im Millisekundenbereich möglich ist, aber im WAN mehrere hundert Millisekunden dauern kann.

Ein globales Weitverkehrsnetz ist das *Internet* (Interconnected Networks). Es basiert auf dem Client-Server Prinzip, da Clients große Datenmengen von Servern abrufen. Protokolle, die auf das Internet aufsetzen sind unter anderem *File Transfer Protocol (FTP)* zur Dateiübertragung, *Simple Mail Transfer Protocol (SMTP)* zum Versand elektronischer Post sowie *Hypertext Transfer Protocol (HTTP)* zur Übertragung von Webseiten. Eine differenzierte Betrachtung der Systemantwortzeiten im World Wide Web (WWW) erfolgt aufgrund der wichtigen gesellschaftlichen und geschäftlichen Bedeutung in Abschnitt 5.4.

5.3.5 Mobile Systeme

In den frühen 1990er Jahren wurde die zweite Generation (2G) der digitalen Mobilkommunikation mit dem *Global System for Mobile Communications* (GSM) eingeführt (vgl. Halonen, Romero, Melero, 2003, S. xxv). Die anfängliche Fokussierung der Infrastruktur auf die Sprachübermittlung wurde durch die veränderten Anforderungen zu einem globalen Daten- und Medienservice erweitert (vgl. Dix, 2003, S. 355). Die Leistungen des GSM wurden mit dem datenpaketorientiertem *General Packet Radio Service* (GPRS) ergänzt. Anzuführen ist die Einführung des W*ireless Application Protocol* (WAP) zur Transformation des Internets auf Mobiltelefone. Die höheren Anforderungen, die Daten- und Mediendienste an die mobilen Systeme stellten, machten es erforderlich, dass eine schnellere dritte Generation (3G) entwickelt werden musste. Das Universal Mobile Telecommunikations System (UMTS) ist eine erste Entwicklung in diese Richtung (vgl. Halonen et al. 2003, S. xxv).

Die Mobilfunksysteme unterscheiden sich von den stationären Systemen in der möglichen Datenrate und der daraus resultierenden Übertragungszeit. Während bei kabelgebundenen Systemen wie WAN ein maximal möglicher Datendurchsatz von bis zu 10 Gbit erreicht wird (vgl. Abschnitt 5.1.3), ist dies beim Mobilfunk noch nicht erreichbar (vgl. Abb. 23).

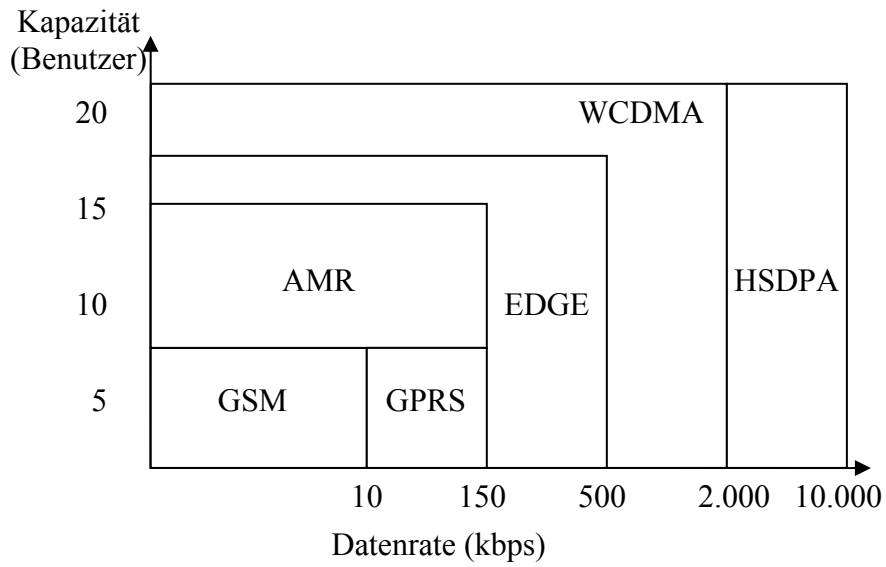


Abbildung 23 Entwicklung der Mobilfunkstandards (vgl. Halonen et al., 2003, S. xxvi)

Die Datenrate ist beim GSM auf lediglich 9,6 kbit/s beschränkt. Durch Weiterentwicklungen konnten mit GPRS 160 kbit/s – wobei allerdings ein Wert von ca. 115 kbit/s realistisch ist – und *Enhanced Date Rates for GSM Evolution* (EDGE) von 220 kbit/s erzielt werden (vgl. Stein, 2004, S. 333). Die *Adaptive Multi-Rate* (AMR) ermöglicht lediglich eine bessere Telefonqualität. Viel interessanter ist die Erhöhung der Datenrate durch die Entwicklung des UMTS. Hier bieten *Wideband CDMA* (WCDMA) mit bis zu 2 Mbps (vgl. Melero, Toskala, Hakalin, Tolli, 2003, S. 532) und *High Speed Downlinks Packet Access* (HSDPA) bis zu 10 Mbps bessere Rahmenbedingungen für die hohen Anforderungen der multimedialen Datendienste.

Mit der Verbreitung der Mobilfontechnik stieg das Bedürfnis der Benutzer mobil zu arbeiten und mit dem Internet vernetzt zu sein. Wir unterscheiden hierbei zwischen Notebooks, Personal Digital Assistants (PDAs) und Mobiltelefonen. *Notebooks* sind ähnlich wie Computer. Sie sind nur kleiner gebaut und dadurch flexibler in der Handhabung und in der Wahl des Benutzungsortes. Ferner können sie sich kabelgebunden oder kabellos in Netzwerke einwählen und sind vernetzt. *PDA*s sind wesentlich kompakter und haben einen kleineren Bildschirm. Sie sind meistens nicht mit dem Netz verbunden, so dass sie auf Informationen nicht unmittelbar zugreifen können. Dagegen sind *Mobiltelefone* meistens immer mit dem Netz verbunden, haben dafür aber auch nur einen sehr kleinen Bildschirm und einen geringeren Funktionsumfang (vgl. Dix, Finlay, Abowd, Beale, 2004, S. 756 f.). Dies führt zu der Forderung, dass ein System klar und deutlich angeben muss, wenn es keine Verbindung zum Netz hat. Ferner soll das System die Arbeit auch im Offline-Modus mittels zwischengespeicherten Daten ermöglichen und sich automatisch synchronisieren, wenn die Verbindung wiederhergestellt ist (vgl. Pearrow, 2002, S. 123). Durch diese Rückmeldungen wird es dem Benutzer ermöglicht, aktuelle Handlungsspielräume und Systemreaktionen richtig einzuschätzen.

Als eine solche Anforderung an mobile Systeme gilt es auch die Systemantwortzeiten zu spezifizieren. Die durch mobile Zugriffe geringen Datenraten lassen eine andere Erwartungshaltung der Benutzer gegenüber denen stationärer Systeme implizieren. Insbesondere der Arbeitsfluss netzwerkabhängiger Arbeit kann durch die Nut-

zung mobiler Systemen beeinträchtigt, wenn nicht sogar durch ständige Verbindungsabbrüche unmöglich gemacht werden.

Eine Vergleichsanalyse der Endbenutzerleistung von GPRS und EGPRS Technologie findet sich in (Gomez, Sanchez, Cuny, Kuure, Paavonen, 2003, S. 333 ff.). Sie vergleichen die Systemantwortzeiten von GPRS und EGPRS in den vier Kontexten: Webbrowser, WAP, MMS und Streaming. Eine Zusammenstellung der Ergebnisse liefert Tabelle 7. Es zeigt sich, dass durch den Einsatz von EGPRS die Systemantwortzeiten um über 50 % reduziert werden können. Trotz allem sind die Zeiten, insbesondere beim Web, viel zu lange und dürften für die Benutzer nicht tolerabel sein. Interessant ist in diesem Zusammenhang, dass Gomez et al. (2003, S. 235 ff.) zeigen, dass es beim Web – insbesondere im EGPRS Fall – keinen linearen Zusammenhang zwischen Webseitengröße und Systemantwortzeit gibt. Beim WAP dagegen, insbesondere bei GPRS mit der Nutzung von UDP, zeigt sich ein linearer Zusammenhang. Dieser Zusammenhang bestätigt sich auch bei der Größe von MMS.

	Größe	**Systemantwortzeit**		**Zeit-einsparung**
		GPRS	**EGPRS**	
Webbrowser	100 kB	~ 23 sec	~ 10sec	~ 56 %
WAP	1,4 kB	~ 3,2 sec	~ 2,9 sec	~ 10 %
MMS	30 kB	~ 47 sec	~ 34 sec	~ 28 %
Streaming	40 kbps	~ 17 sec	~ 12 sec	~ 30 %

Tabelle 7 Systemantwortzeiten bei GPRS und EGPRS (nach Gomez et al., 2003, S. 233 ff.)

Abschließend gehen Gomez et al. (2003, S. 341 f.) darauf ein, dass Systemantwortzeiten, insbesondere durch Netzwerkverzögerung, starken Einfluss auf Onlinespiele haben. Allerdings gilt es zwischen den verschiedenen Spielarten wie Actionspielen, echtzeitbasierten Strategiespielen und rundenbasierten Strategiespielen zu differenzieren. Es lässt sich aber die Aussage machen, dass die Systemantwortzeit bei Spielservern im Netzwerk um die Hälfte geringer ist als bei Peer-to-peer-Spielen.

Interessant ist in diesem Zusammenhang die technische Spezifikation des *3rd Generation Partnership Project* (3GPP), die die Service- und Systemaspekte und deren Leistungsvermögen festgelegt hat. Unter dem Punkt Trägerdienste werden Verzögerungen und deren Streuung explizit als Charakteristika genannt.

	Medium	**Anwendungsbeispiel**	**Datenrate**	**Verzögerung**	**Art**
Echtzeit-/ Gesprächs-Dienste	Audio	Gespräch	4-25 kb/s	< 150 msec	Ende-zu-Ende
	Video	Bildtelefon	32-384 kb/s	< 150 msec	
	Daten	Telemetrie, Telnet,	< 28,8 kb/s	< 250 msec	
		interaktive Spiele	< 1 KB	< 250 msec	
Interaktive Dienste	Audio	Gesprächsnachricht	4-13 kb/s	< 1 sec	
	Daten	Webseiten (HTML)		< 4 sec / Seite	
		Transaktionen		< 4 sec	
		E-Mail (Zugriff)		< 4 sec	
Streaming	Audio	Sprache & Musik	5-128 kb/s	< 10 sec	Starten
	Video	(Echtzeit-) Filme	2-384 kb/s	< 10 sec	
	Daten	Informationstransfer	< 384 kb/s	< 10 sec	

Tabelle 8 Leistungserwartung der Endbenutzer (vgl. 3GPP, 2005, S. 15 f.)

In der Tabelle 8 wird durch die 3GPP eine Klassifizierung der Dienste in Echtzeit, Interaktiv und Streaming und der jeweiligen Differenzierung nach Mediumart vorgenommen. Interessant ist die Auflistung dahingehend, dass Verzögerungen, also Systemantwortzeiten, spezifiziert wurden. Im Echtzeitbereich wurden sie – vergleichbar mit der Direkten Manipulation – im Millisekunden-Bereich festgelegt; wobei für Audio- und Videodienste eine kürzere Verzögerungszeit gefordert wird als bei Daten. Erstaunlich ist der Bereich der interaktiven Datendienste wie Webseiten, in dem ein Verzögerungsgrenzwert von 4 Sekunden festgelegt wurde, obwohl dies eigentlich schon ein nicht tolerabler Wert ist.

Es zeigt sich, dass es für mobile Systeme kaum Studien und Ergebnisse gibt. Aufgrund des veränderten Anwendungskontextes gegenüber stationären Systemen wird von den Benutzern scheinbar eine höhere Systemantwortzeit toleriert.

5.4 Internet

5.4.1 Bedeutung des Webs

In diesem Abschnitt soll das Web mit der technischen Infrastruktur des Internets im Bezug auf die Systemantwortzeiten betrachtet werden. Als Internet wird das globale Netzwerk bezeichnet, das sich aus vielen Einzelnetzwerken (LAN und WAN) zusammensetzt und zum Informationsaustausch die TCP/IP-Protokollfamilie nutzt (vgl. Mertens et al., 2005, S. 43). Die Besonderheit des Internets liegt in dessen rasanter Entwicklung (vgl. Abb. 22). Waren im Jahre 1981 nur 213 Computer angeschlossen, so wurden im Juli 2005 mehr als 353 Millionen Computer gezählt. Mit dem rasanten Anstieg der Hosts stieg auch die Anzahl der Benutzer, die in ihren Kenntnissen und Anforderungen an das Web sehr heterogen sind. Des Weiteren erlang das Internet durch den E-Commerce eine zunehmend ökonomische Bedeutung. Im Rahmen der neuen Institutionsökonomik lassen sich dadurch die Transaktionskosten senken.

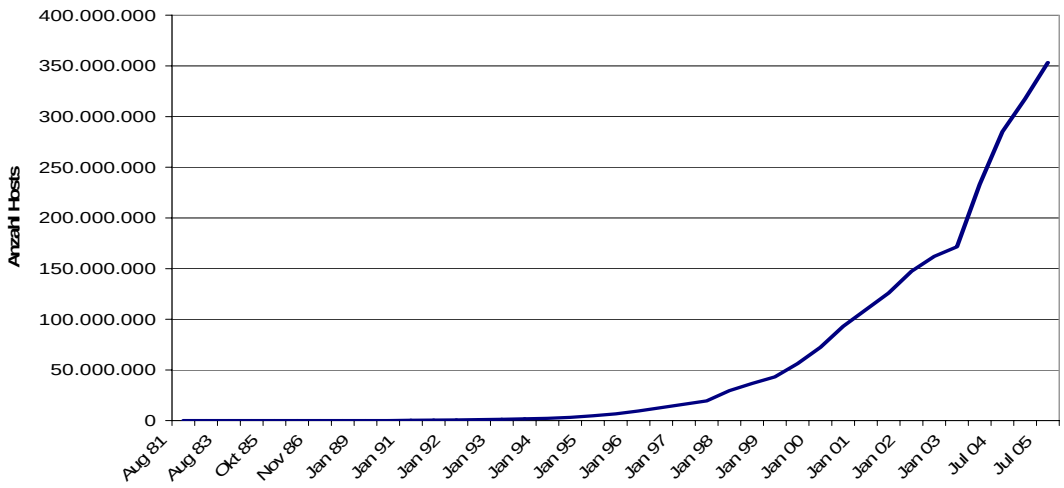


Abbildung 24 Wachstumsentwicklung des Internets (vgl. ICS, 2005)

Es ist erforderlich zwei Sichten auf das Internet einzunehmen. Zum einen die technische Sicht (Abschnitt 5.4.2), die durch die Infrastruktur bedingt ist. Die Leistungsfähigkeit von Webservices gilt es mittels Dienstgüteparametern zu überwachen (Abschnitt 5.4.3). Ferner werden Ansätze zur technischen Optimierung von Systemantwortzeiten aufgezeigt (Abschnitt 5.4.4). Auf der anderen Seite gilt es die Benutzersicht mit der Wahrnehmung von Systemantwortzeiten als Verzögerung zu betrachten (Abschnitt 5.4.5).

5.4.2 Technische Infrastruktur

Die technische Infrastruktur des Internet ist sehr komplex. In einem einfachen Modell wird in Abb. 25 der Weg einer Benutzeranfrage über einen Client an einen Webserver dargestellt. Diese Anfrage läuft über eine Vielzahl von Routern, Gateways und Hosts durch das Internet. Durch die Vielzahl der Komponenten zeigt sich, dass es viele Stellen gibt, an denen es zu Verzögerungen kommen kann.

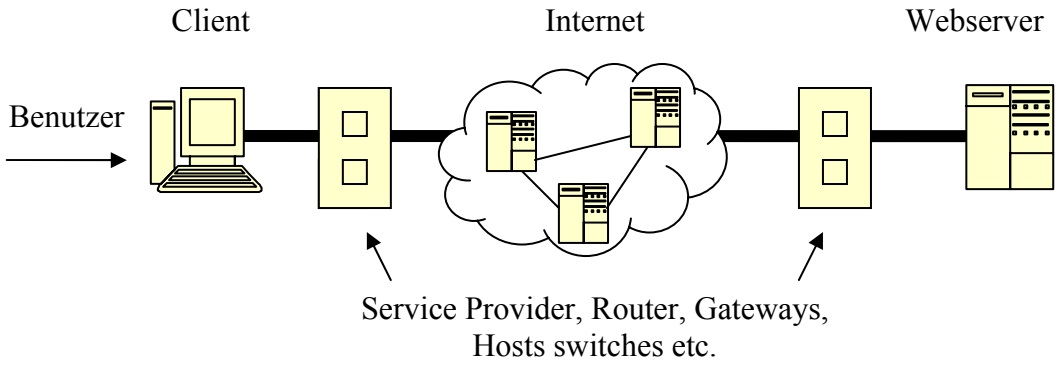


Abbildung 25 Einfaches Web-Modell (vgl. Cremonesi & Serazzi, 2002, S. 161)

Als die drei wesentlichen Hauptkomponenten, die die Systemantwortzeit im Internet beeinflussen, gilt es den Client, das Netzwerk und den Server näher zu betrachten (vgl. Abb. 26). Während beim Browser durch die Ein- und Ausgabe, sowie die Bearbeitung Verarbeitungszeiten entstehen, ist die Systemantwortzeit im Netzwerk durch die Verbindungszeiten zum Internet Service Provider (ISP) und der allgemeinen Laufzeit geprägt. Die Systemantwortzeit des Servers setzt sich im Wesentlichen aus den Komponenten der Ein- und Ausgabe, der Bearbeitung im Server sowie durch die Vernetzung mit anderen Systemen und Systemkomponenten zusammen.

Browser-Zeit		**Netzwerk-Zeit**			**E-Commerce Server-Zeit**		
Bearbeitung	I/O	Verbind-ungszeit: Browser zum ISP	Internet Zeit	Verbind-ungszeit: ISP zum Server	Bearbeitung	I/O	Ver-netzung
Engpass							

Abbildung 26 Aufschlüsselung der Systemantwortzeit (vgl. Menascé, Almeida, Dowdy, 2004, S. 13)

Die Systemantwortzeit im Web beginnt genauer betrachtet beim Klick des Benutzers im Browser (vgl. Abb. 27). Liegt der Inhalt im Cache des Clients vor, so wird die Anfrage in sehr kurzer Zeit beantwortet (R_{cache}). Ansonsten erfolgt eine HTTP-Anfrage über das Netzwerk zum Server, der dann die Daten ausliefert. Die gesamte Systemantwortzeit (R_{total})wird auch als *Ende-zu-Ende Antwortzeit* bezeichnet, weil sie die Zeitspanne vom Beginn der Benutzeranfrage bis zu deren vollständigen Antwort beim Benutzer angibt und damit der tatsächlichen Benutzerwahrnehmung entspricht; anders als die reine Auslieferungszeit des Servers (R_{Server}).

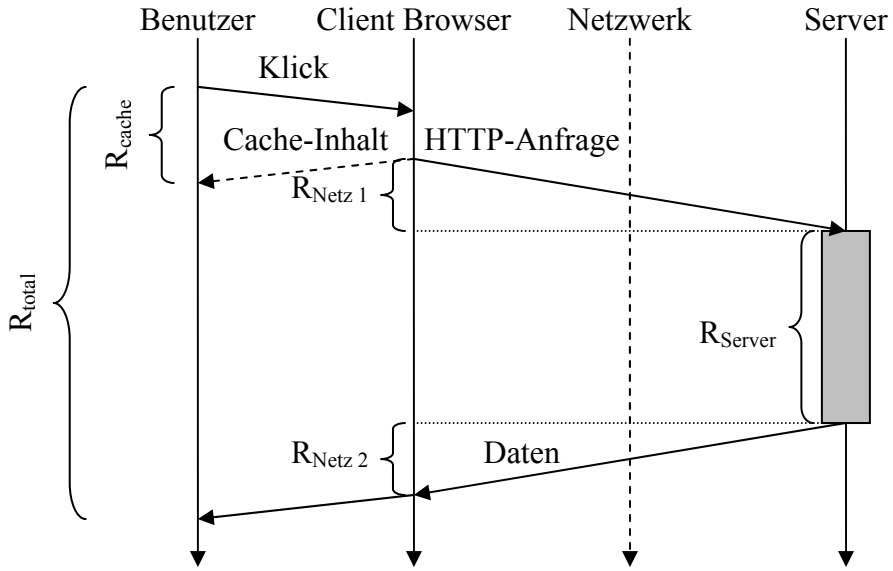


Abbildung 27 HTTP Transaktion (vgl. Menascé & Almeida, 2002, S. 135)

Der Verbindungsaufbau und Datenaustausch zwischen Client und Server ist sehr komplex und zeitintensiv (vgl. Abb. 28). Als erstes findet ein *DNS Lockup* statt, in dem die alphanumerische Adresse in eine IP Adresse umgewandelt wird. Darauf findet die erste *TCP Verbindung* zwischen dem Client und dem Server statt. Werden zusätzlicher Daten anderer Server benötigt, findet ein *Redirection* statt. Danach wird die vom Client gestellte HTTP Anfrage mit dem *Download des ersten Paketes* vom Server beantwortet. Hat dieses einwandfrei funktioniert, so erfolgt der eigentliche Inhaltsdownload mit sämtlichen eingebetteten Elementen (vgl. Zhi, 2001). Im HTTP 1.0 Protokoll muss hierfür jeweils eine neue TCP Verbindung aufgebaut werden. Beim HTTP 1.1 entfällt der Verbindungsaufbau für jedes Element und reduziert damit die Antwortzeit signifikant (RFC2616, 1999, S. 43).

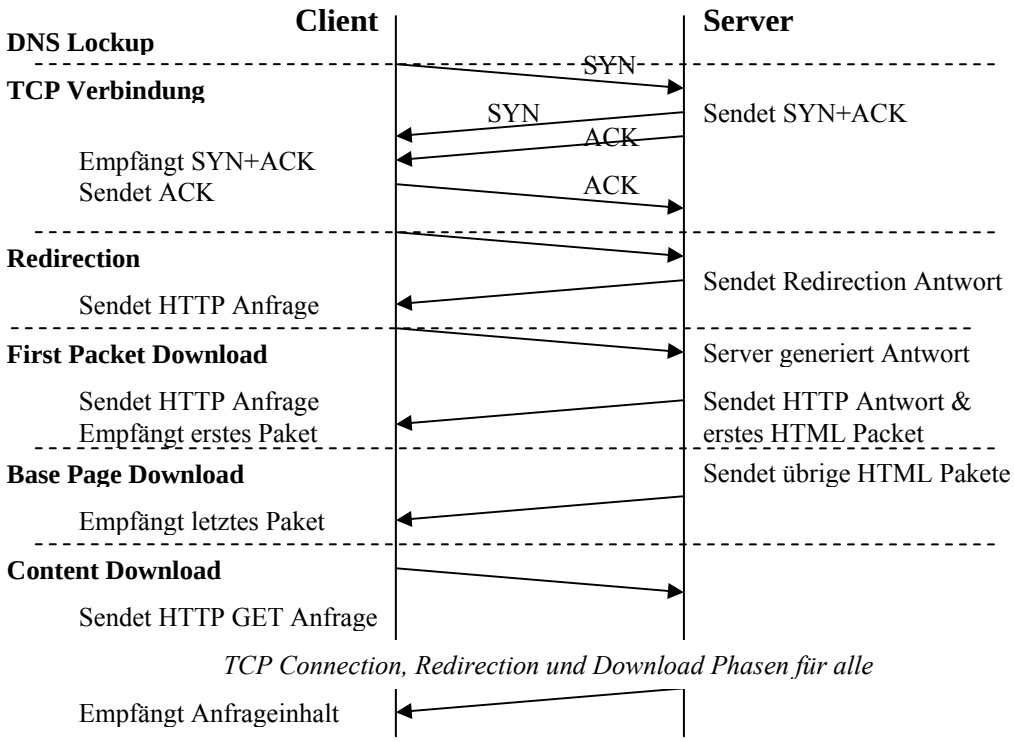


Abbildung 28 Systemantwortzeitkomponenten von Webseiten (in Anlehnung an Zhi, 2001, S.2)

Zu bedenken gilt es allerdings, dass die an den Server gerichtete Anfrage meist nicht von ihm alleine beantwortet werden kann, sondern zur Bearbeitung an andere weitergeleitet werden muss. So bestehen E-Commerce-Systeme im einfachen Fall aus einem Web-Server, dem ein Anwendungsserver und ein Datenbankserver nachgestellt sind. Meist ist die Systemarchitektur um einiges komplexer. Bei großen E-Commerce-Systemen werden mehrere Web-Server parallel betrieben, so dass ein Loadbalancer vorgelagert wird, der die Aufgabe hat, die Anfragen entsprechend festgelegter Regeln zu verteilen (vgl. Menascé & Almeida, 2002, S. 161 f.).

Systemantwortzeiten sind im Web somit eine sehr komplexe Thematik. Eine Vielzahl von Komponenten wirkt sich direkt auf sie aus, so dass bei einer Optimierung sehr differenziert geschaut werden muss, welche Komponente die Verzögerung verursacht und wie dies behoben werden kann. Nicht zu vergessen sind die Benutzer, die meist kein differenziertes Wissen über die komplexen Systemarchitekturen haben und für die nur das Gesamtergebnis zählt, dass die Ladezeit kurz ist.

5.4.3 Dienstgüte bei Webservices

Durch die Verlagerung von Geschäftsprozessen von unternehmensinternen Systemen zu externen Anbietern – den Application Service Providern (ASP) – sind Dienstgüteparameter ein entscheidendes Qualitäts- und Auswahlkriterium. Als Bewertungsgrundlage sehen Berbner, Heckmann, Mauthe, Steinmetz (2004, S.268 ff.) eine Reihe von Parametern (vgl. Tab 9). Es gilt hier zwischen quantitativ messbaren Kriterien wie Verfügbarkeit, Leistungsfähigkeit, Fehlerhäufigkeit und so genannten „weichen" Kriterien wie Sicherheit, Reputation und Kosten, die durch den Nutzer mittels Bewertungsmatrix bewertet werden, zu unterscheiden.

Parameter	**Teilkriterien**	**Bewertungsart**
Verfügbarkeit		Quantitativ messbar
Leistungsfähigkeit	Durchsatz, Antwortzeit	
Fehlerhäufigkeit	Fehleranzahl je Zeitintervall	
Sicherheit	Authentizität, Autorisierung, Vertraulichkeit und Datenverschlüsselung	Bewertung durch Nutzerseite
Reputation	Positive Erfahrungen, Referenzen	
Kosten	Abrechnungsart	

Tabelle 9 Dienstgüteparameter von Webservices (vgl. Berbner et al., 2004, S. 270 f.)

Systemantwortzeiten sind neben dem Durchsatz ein integraler Bestandteil der Leistungsfähigkeit und damit ein entscheidendes Auswahlkriterium. Als optimaler Wert gilt eine möglichst geringe Systemantwortzeit. Um in einem Bewertungsverfahren Gewichtungsfaktoren festzulegen, gilt es normalisierte Werte zu ermitteln. Hierzu wird unter allen zur Auswahl stehenden Diensten die längste Antwortzeit durch die des jeweiligen Dienstes dividiert. Der auszuwählende Dienst sollte somit einen möglichst hohen normalisierten Wert der Systemantwortzeit haben.

Damit lässt sich festhalten, dass die Systemantwortzeiten als ein wichtiges Bewertungskriterium bei der Verlagerung von geschäftskritischen Prozessen zu externen Serviceanbietern anzusehen sind.

5.4.4 Ansätze zur technische Optimierung

Um die Systemantwortzeiten zwischen Server und Client zu reduzieren, bietet sich das Cachen an. Hierbei werden die in Webseiten eingebundenen Objekte nicht direkt vom Server, sondern aus dem Cachespeicher ausgeliefert. Dies ermöglicht neben einer reduzierten Verzögerungszeit, die auch als solche wahrgenommen wird, einen reduzierten Netzwerkverkehr, als auch eine Reduzierung der Serverlast. Dabei gilt es allerdings zu bedenken, dass die im Cache hinterlegten Objekte zum Auslieferungszeitpunkt noch aktuell sein müssen. Yuan & Chi (2003, S. 23 ff.) weisen darauf hin, dass die Abfragezeit mit steigender Objekt-/ Einheiten-Anzahl zunimmt. In ihrer Studie untersuchten sie 1,36 Mio. Webanfragen. Sie fanden heraus, dass viele Webserver falsch konfiguriert sind. Durch Optimierung ließen sich 30% der nicht gecachten Objekte zusätzlich im Cache vorhalten und 30-80% der Gültigkeitsprüfungen vermeiden.

Bhalekar & Baras (2004) betrachten die Besonderheit der Satellitennetzwerkverbindung zwischen Benutzer und WWW-Server. Aufgrund der großen Distanz zwischen Erde und Satellit (ca. 36.000 km) und der dadurch bedingten langen Verbindungswege – von der Erde über den Satelliten zum Hub und vom Hub zurück über den Satelliten zur Erde – kommt es zu Verzögerungszeiten von mehr als einer halben Sekunde. Da sich diese Zeit signifikant auf die Benutzer auswirkt, gilt es diese zu minimieren. Bhalekar und Baras bedienen sich hier der Feststellung, dass 80% der Webanfragen lediglich 20% der Onlineressourcen betreffen und im Umkehrschluss dann die restlichen 20% der Anfragen sich auf die übrigen 80% der Webressourcen beziehen. Durch Caching beim Benutzer, so argumentieren sie, ließen sich mehr als 40% der Verzögerungen reduzieren, weil die abgefragten Inhalte direkt aus dem Cache und nicht vom Server mittels Satellitenverbindung ausgeliefert werden müssten. Damit die Inhalte im Cache aktuell sind, werden diese spätestens alle 24 Stunden automatisch aktualisiert bzw. bei es wird bei Anfragen der Zeitstempel verglichen.

Der Ansatz von Bhalekar und Baras ist interessant, weil sie mit der Satellitenverbindung eine Verbindungsart ansprechen, die zum einen wenig betrachtet wird und zum

anderen in einer Erweiterung bei einem mobilen Einsatz sicher auch für mobile Systeme (vgl. Abschnitt 5.5) von Interesse sein dürfte. Allerdings geben die Autoren selbst die Einschränkungen, dass sie sich auf Benutzer in kleinen Büros und von Heimarbeitsplätzen beziehen. Sicherheits- und Datenschutzaspekte müssen weitergehend betrachtet werden. Ferner konnten sie keine Aussagen über Kosten und Nutzen machen.

Einen weiteren interessanten Aspekt betrachten Garg et al. (2002, S. 329 ff.), die die Thematik der dynamischen Objekte im Intranet aufgreifen und feststellen, dass – wider deren ursprüngliche Erwartung – die Intranetleistung als Flaschenhals anzusehen ist und damit dem Internet gleichzusetzen wäre. Um Systemantwortzeiten zu reduzieren, schlagen sie ebenfalls das Prinzip des Cachen vor. Bei dynamischen Inhalten muss bedacht werden, dass die Webseiten erst generiert werden müssen. Dies entweder bei Abruf, bei Änderungen oder periodisch. Die Auslieferung von dynamischen Seiten durch den Cache führte in der Studie zu einer Reduzierung der Systemantwortzeit von 80%. Ferner empfehlen sie bei nicht cachebaren dynamischen Webseiten eine Umwandlung in statische Webseiten, sodass dadurch Netzwerklast und -verkehr reduziert werden. Garg et al. geben aber zu bedenken, dass sich die Ergebnisse nicht einfach auf Web-Applikationen übertragen lassen und näher untersucht werden müssen.

Eine Untersuchung zu den Auswirkungen dynamischer Webseiten-Generierung auf die Systemantwortzeit im Internet lässt sich in Tichkosky, Arlitt & Williamson (2003) finden. Dort werden sowohl Server-Software wie Perl, PHP und Java, als auch die statische und dynamische Auslieferung von Webseiten betrachtet. Sie kommen zu dem Ergebnis, dass die Erzeugung dynamischer Seiten eine achtfach höhere Auslastung des Servers mit sich bringt. Ferner ist die Java-Serversoftware den anderen Sprachen überlegen, auch wenn PHP bei kleinen dynamischen Inhalten noch mithalten kann.

5.4.5 Benutzersicht

Lange und variierende Systemantwortzeiten wirken sich, wie schon dargestellt, negativ auf die Benutzer aus. Im Internet sind Systemantwortzeiten von besonderem Interesse, da sie kaum vom Benutzer beeinflusst werden können, sondern infrastrukturtechnisch bedingt von sehr vielen Komponenten abhängen (vgl. Abschnitt 5.4.2). Des Weiteren ist die stetig wachsende Bedeutung der geschäftlichen Abwicklungen über das Internet und im Speziellen des Webs zu beachten. Lange Systemantwortzeiten können Geschäftsabwicklungen behindern oder sogar ganz verhindern und nachhaltig auswirken. Untersuchungen von Bouch, Kuchinsky & Bhatti (2000a, 2000b) zeigen, dass sich die Benutzerwahrnehmung von Systemantwortzeiten im Web signifikant auf das Image der Firma und deren Produkte überträgt. Die persönliche Wartebereitschaft des Benutzers wird bedingt durch seine Erfahrungen und dem zugrunde liegenden konzeptuellen Modell, das er vom Web hat.

Dyson & Longshaw (2004, S. 32) stellen dar, dass die Erwartungshaltung auch durch das technische System bedingt ist. Beim Intranet sind langsame Systemantwortzeiten durch nachvollziehbar komplexe Funktionen vertretbar. Sie werden aber bei einfachen Aktionen nicht toleriert. Man spricht davon, dass sich bei den Benutzern angemessene Erwartungen bilden. Dies lässt sich ebenfalls auf das Extranet übertragen. Anders verhält es sich beim Internet. Hier gilt die Benutzererwartung einem schnellen Internetsystem. Die Komplexität der durchzuführenden Funktionen wird allerdings nicht betrachtet. Dadurch bilden sich bei den Benutzern keine Erwartungswerte für die Systemantwortzeiten.

Nah (2004) untersucht die Toleranz von Wartezeiten bei Benutzern im Web. Generell ist zu bedenken, dass die Wartezeit-Toleranz durch die jeweilige Aufgabe bedingt ist. Bei der Informationssuche im Web liegt die maximale Wartebereitschaft bei ca. zwei Sekunden und deckt sich mit der frühen Studie von Miller (1968). Fortschrittsanzeigen können den Benutzer Informationen über die zu erwartende Wartezeit liefern. Dadurch ließe sich eigentlich die wahrgenommene Wartezeit reduzieren. Der Nutzen wird allerdings in der Studie von Hui & Zhou (1996) angezweifelt, die

zeigen, dass die Wartezeitanzeige die wahrgenommene Wartezeit nicht reduziert. Nah (2004) weißt darauf hin, dass deren Studie entgegen der allgemeinen Auffassung ist und weitere empirische Studien von Nöten sind. Nielsen (2001, S. 259) schlägt Fortschrittsanzeigen bei Applets vor, vor denen die Antwortzeiten mehr als 10 Sekunden benötigt. Ferner sollte der Vorgang vorzeitig beendet werden können.

Es stellt sich somit die Frage, welche Wartezeit als eine optimale zu fixieren ist. Wie schon ausgeführt, hängt dies von der Art der Interaktion und der jeweiligen Benutzererwartung und Toleranz zusammen, so dass eine exakte allgemeingültige Bestimmung nicht möglich ist, sondern nur für spezielle Anwendungsfälle gilt (vgl. Meyer, Vogt & Glier, 2005a,b). An dieser Stelle soll ein Überblick über vorgeschlagene Systemantwortzeiten gegeben werden, um Anhaltspunkte aufzuzeigen (vgl. Tab. 10). Während Nielsen (2001, S. 42) als Minimalziel eine Antwortzeit unter 10 Sekunden vorschlägt, hat sich dies in neueren Ergebnissen auf 2-3 Sekunden reduziert (vgl. Nah, 2004; Dyson & Longshaw, 2004, S. 62). Interessant ist hierbei, dass Dyson & Longshaw eine Differenzierung der Antwortzeit nach der Aufgabentätigkeit vornehmen und bei Extranet-Verbindungen eine größere Wartezeit-Toleranz unterstellen, da Extranets zwischen mehreren Unternehmen genutzt werden (vgl. Mertens et al., 2005, S. 48 f.). Es liegen allerdings noch immer zu wenig empirische Befunde vor, aus denen sich Empfehlungen für Systemantwortzeiten ableiten lassen.

Zeit	**Interaktionsart**	**Quelle**
1 sec	Kein Feedback erforderlich	Nielsen (1993, S. 135)
< 2 sec	Informationsabfrage	Nah (2004)
< 3 sec	Homepage	Dyson/Longshaw (2004, S. 62)
< 6 sec	Inhaltsseiten (statisch/dynamisch)	Dyson/Longshaw (2004, S. 62)
< 10 sec	Einfache Suchoperation	Dyson/Longshaw (2004, S. 62)
10 sec	Aufmerksamkeitslimit	Nielsen (1993, S. 135)
< 20 sec	Erweiterte Suchoperation	Dyson/Longshaw (2004, S. 62)
< 20 sec	Anfrage Bestellstatus Extranet	Dyson/Longshaw (2004, S. 63)
< 30 sec	Anfrage Bestellbestätigung Extranet	Dyson/Longshaw (2004, S. 63)

Tabelle 10 Systemantwortzeiten von Webseiten

6 Wirtschaftliche Aspekte der Systemantwortzeiten

6.1 Wirtschaftlichkeit von Informationssystemen

6.1.1 Bedeutung der Informationssysteme

Im Jahr 1991 beklagten Bauknecht, Tjoa und Draxler (1991, S. I), dass es „mit den derzeit vorliegenden Methoden nicht möglich ist, den Nutzen von Informationssystemen exakt zu quantifizieren“. Carr (2003, S. 41 ff.) stellte die provokante und viel diskutierte These *IT doesn't matter* auf und hinterfragte den Wettbewerbsvorteil, der durch direkte Investitionen in den Einsatz der IT entsteht. Carr zeigt hierzu eine Parallelität zur Elektrizität auf. Er argumentiert, dass die Elektrizität zum Allgemeingut geworden ist, und überträgt dies auf die Informationstechnik, die seiner Meinung nach ebenfalls ein Allgemeingut ist und keinem Unternehmen Vorteile bringt. Metcalfe (2004) gibt zu bedenken, dass die kontroverse Diskussion über Carrs Artikel sich mehr auf den Titel als auf den eigentlichen Inhalt bezieht. „IT ist für jeden wichtig“ bringt es Metcalf (2004, S. 100) auf den Punkt. Carr selbst relativiert seine Thesen später in seinem Buch *IT does matter* (Carr, 2004) und weist auf den wichtigen und nicht zu vernachlässigenden Bezug zwischen der eingesetzten Technik, den Menschen und der zu bewältigenden Aufgabe hin.

Damit zeigt diese Triade – Mensch-Technik-Aufgabe – einen direkten Bezug zur Software-Ergonomie auf und bietet einen Ansatz für die Forderung von Oberquelle (2000, S. 4ff), dass die direkten und indirekten Kosten der (Un-)Benutzbarkeit kritisch durchleuchtet werden müssen und Potential für mehr Effizienz bieten.

6.1.2 Produktivitätsparadoxon

Die Forderung nach Effizienz führt allerdings zu der grundsätzlichen Diskussion des Produktivitätsparadoxons. Dies besagt, dass trotz stetig leistungsstärkerer und günstigerer Informationstechnologien (Jovanovic, Rousseau, 2003, S. 14 ff.) die Produktivität stagniert (Gründler, 1997, S. 2). Als Gründe gibt Gründler (1997, S. 74 ff.) unter anderem Schwierigkeiten bei der Messung der In- und Output-Größen, fehlende bzw. verzögerte Realisierung von Produktionsvorteilen sowie Missmanagement

und politische Widerstände beim Einsatz der IT an. Potthof (1998, S. 54 ff.) untersuchte zu dieser Thematik 49 Studien – hauptsächlich aus den USA – und klassifiziert die Erklärungspunkte für das Produktivitätsparadoxon in methodische Defizite und reale Probleme. Als methodische Defizite werden die schon genannten Messprobleme der In- und Outputgrößen, zeitliche Differenz zwischen Investition und Nutzen sowie unzureichende Kenntnisse von Auswirkungen zusammengefasst. Reale Probleme umfassen den Einflussfaktor der IT auf das Unternehmen, keine generalisierbaren Produktivitätsgewinne durch die IT, den Faktor Mensch mit dem Hang zu Planungs- und Einführungsfehlern sowie die Akzeptanzproblematik.

Es gibt allerdings auch Studien von Brynjolfsson & Hitt (2003) und Jorgenson & Stiroh (2000), die das Produktivitätsparadoxon widerlegen. Während Jorgenson & Stiroh (2000) die Nachhaltigkeit der Produktivität an die Entwicklung der Halbleiterproduktion – und damit an Moores-Gesetz – koppeln, stützen Brynjolfsson & Hitt (2003) ihre These gegen das Produktivitätsparadoxon auf Beobachtungen, die sie im Zeitraum der späten 1980er bis in die frühen 1990er Jahre machten. Demnach resultiert die Produktivitätssteigerung nicht nur aus dem Einsatz von IT, sondern auch aus dem Vorhandensein eines organisatorischen Rahmens (Brynjolfsson & Hitt, 2003, S. 26 f.). Dies stützt Potthofs (1998, S. 63) Feststellung, dass ein großer Mangel an systematischen Wirtschaftlichkeitsbetrachtungen fehlt und der Nutzen durch IT weitergehend kritisch hinterfragt werden muss.

6.1.3 Wirtschaftlichkeitsvergleich

Ein Wirtschaftlichkeitsvergleich setzt voraus, dass die Input- und Outputgrößen bekannt sind. Dellmann & Pedell (1994, S. 25) klassifizieren die inputbezogenen Größen als Effizienz, während die Effektivität durch outputbezogene Merkmale beschrieben wird. Dadurch lässt sich eine mengenmäßige Produktivität ableiten, die wertmäßig als Wirtschaftlichkeit bewertet werden kann. Im Sinne dieser inputbezogenen Effizienz sind optimale Systemantwortzeiten dahingehend zu spezifizieren, dass für den gewählten Anwendungskontext keine hardware- und softwaretechnischen Verbesserungen mehr erzielt werden können. Die Effektivität zeigt sich dann sowohl in der Qualität des Systems als auch in der Zufriedenstellung der Benutzer.

Es muss allerdings immer eine Güterabwägung zwischen Kosten einerseits und Nutzen andererseits stattfinden, um ein solches System effizient und effektiv umzusetzen. Die Kosten lassen sich unterscheiden in einmalige und laufende Kosten (vgl. Abb. 29). Der Nutzen lässt sich aufteilen zwischen nicht quantifizierbaren und nicht monetärem Nutzen, wie z.B. Erhöhung der Datenaktualität, sowie quantifizierbarem Nutzen. Der quantifizierbare Nutzen lässt sich zum einen monetär bewerten, z.B. Verkürzung von Arbeitszeiten, und zum anderen nicht monetär bewerten, z.B. höherer Servicegrad (vgl. Stahlknecht & Hasenkamp, 2005, S. 251 f.).

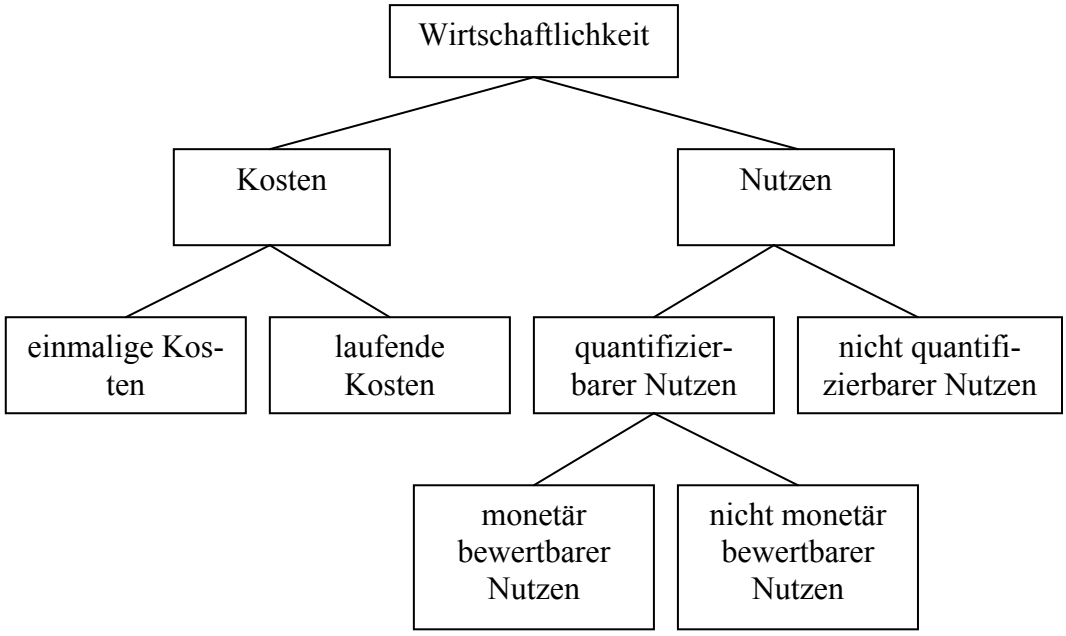


Abbildung 29 Kosten-Nutzen-Vergleich (vgl. Stahlknecht & Hasenkamp, 2005, S. 252)

Nach Heinrich & Lehner (2005, S. 368 ff.) gibt es zur Lösung der Entscheidung unter der Zielvorgabe der Wirtschaftlichkeit folgenden Ablauf:

1. Ermittlung der Kostenarten
2. Ermittlung der Nutzenstruktur
3. Ermittlung der Beziehungszusammenhänge zwischen Kosten und Nutzen
4. Auswahl der optimalen Alternative

Dabei ist immer zu bedenken, dass durch eine Umsetzung Veränderungen entstehen, die sich nicht nur isoliert und technikbezogen auf das Informationssystem beziehen,

sondern auch subsystembezogene, gesamtorganisatorische oder gesellschaftliche Auswirkungen auf die Wirtschaftlichkeit haben können.

Bei der Analyse Systemantwortzeiten konnte Lambert (1984) – mit Bezug auf Studien von Doherty & Kelisky (1979) und Thadhani (1981) – nachweisen, dass die Produktivität signifikant steigt, wenn die Systemantwortzeit reduziert wird. Bei einer technisch bedingten Reduzierung der Antwortzeit von 2,22 Sekunden auf 0,84 Sekunden konnte Lambert eine um 62 % höhere Benutzerproduktivität und eine 40 %ig reduzierte Benutzerantwortzeit feststellen. Als Ergebnis dieser Studie wurden in der IBM Entwicklungsabteilung Hardware-Verbesserungen durchgeführt (Kosten), um den Entwicklern einen besseren Arbeitsplatz zu bieten und somit von deren höherer Produktivität zu profitieren (Nutzen).

Leistungseinbußen durch schlechte Antwortzeiten bedeuten einen nicht zu verachtenden Kostenfaktor. Williams & Smith (o.J.) führen verlorenen Umsatz, beschädigte Kundenbeziehungen, geringere Wettbewerbsfähigkeit, höhere Arbeitskosten sowie Betriebs- und Projektausfälle als Kostenfaktoren an. Dieses ist in der Hauptsache durch die Architektur und Designfaktoren bedingt. Utton & Hill (1997, S. 1 ff.) führen ferner die hohen Kosten der Korrekturen wie Software-Redesign, Weiterentwicklung mit Fehlerbehebung und Erhöhung der Leistung an. Dies führt zu der Forderung, dass die Leistungsparameter schon vor der Entwicklung genau spezifiziert werden. Dadurch lassen sich Kosten und Kapitaleinsatz reduzieren und die Qualität verbessern.

Nunmehr gilt es, im Rahmen der Wirtschaftlichkeitsanalyse einzelne Bewertungsmethoden näher zu betrachten. Hierzu werden im Abschnitt 6.2 die Arten der Investitionsrechnung betrachtet. In dem darauf folgendem Abschnitt 6.3 werden die in der Praxis gern genutzten Bewertungsmethoden des *Return of Investment* (ROI) (Abschnitt 6.3.1) und des *Total Cost of Ownership* (TCO) (Abschnitt 6.3.2) in Bezug auf benutzergerechte Informationssysteme und der Systemantwortzeit im Speziellen (Abschnitt 6.3.3) beschrieben.

6.2 Verfahren der Investitionsrechnung

6.2.1 Investitionsrechnung als Entscheidungsgrundlage

Nachdem die Wirtschaftlichkeitsbetrachtung und eine Kosten-Nutzen-Beurteilung von Informationssystemen analysiert wurden, gilt es den Endscheidern Handlungsmethoden zur Verfügung zu stellen. Es wird von der Annahme ausgegangen, dass als ein langfristiges Ziel das Gewinnstreben unterstellt werden kann. Dieses monetäre Ziel kann mittels Verfahren der Investitionsrechnung erreicht werden. Natürlich gilt es zu beachten, dass nicht nur ein Ziel verfolgt wird, sondern dieses – auch in enger Koppelung an andere nichtmonetäre Ziele – in einem Zielsystem zusammengefasst wird.

In der klassischen Investitionsentscheidung wird unterschieden zwischen statischen Verfahren (Abschnitt 6.2.2), die auf kalkulatorischen Erlös- und Kostengrößen basieren, und dynamischen Verfahren (Abschnitt 6.2.3), die die zeitliche Komponente der Ein- und Auszahlungen inklusive der Zinseszinsrechnung beinhaltet. Durch diese Dynamik sind die dynamischen Verfahren besser als die statischen, aber auch komplexer, so dass die statischen Verfahren wegen Ihrer Einfachheit in der Praxis eher eingesetzt werden (vgl. Strunz, 1998, S. 296). Beide Verfahrensarten sind allerdings eine Vereinfachung, da sie keine Unsicherheiten der Daten behandeln (vgl. Kruschwitz, 2005, S. 27). Daher gilt es abschließend die Beschränkungen, die diese Methoden mit sich führen, zu betrachten (Abschnitt 6.2.4).

6.2.2 Statische Verfahren

Die statischen Verfahren umfassen folgende Methoden (vgl. Kruschwitz, 2005, S. 31 ff.; Strunz, 1998, S. 302 ff.; Wöhe, 1996, S. 748 ff.):

- Gewinnvergleichsrechnung
- Kostenvergleichsrechnung
- Rentabilitätsvergleichsrechnung
- Amortisationsrechnung

Bei der *Gewinnvergleichsrechnung* wird das Projekt ausgewählt, bei dessen Investition der größte Gewinn für einen einperiodischen Bezugszeitraum zu erzielen ist. Somit werden bei der Gewinnermittlung nur die Kosten und Erlöse betrachtet.

$$Gewinn = Erlöse - Kosten \tag{6.1}$$

Das Potential von Fehlentscheidungen liegt bei der Gewinnvergleichsrechnung in der Beschränkung, dass nur Projekte mit gleicher Nutzungsdauer und gleichem Kapitaleinsatz verglichen werden können.

Bei der *Kostenvergleichsrechnung* wird im Gegensatz zu der Gewinnvergleichsrechnung nur die Kosten-Komponente betrachtet. Dies setzt voraus, dass die Erlöse der zu vergleichenden Projekte gleich hoch sind. Dies birgt die große Gefahr, dass bei der kostengünstigsten Alternative nicht zwangsläufig eine Kostendeckung gewährleistet ist, weil die Erlöse nicht betrachtet werden.

Die *Rentabilitätsvergleichsrechnung* berücksichtigt gegenüber den Gewinn- und Kostenvergleichsrechnungen, dass unterschiedlich viel Kapital in Investitionen gebunden werden kann. Hierzu wird eine Renditeziffer bestimmt, die sich aus dem Erlös abzüglich der Kosten und der Abschreibung, dividiert durch das gebundene Kapital ergibt. Investiert wird in das Projekt mit der größten Rendite.

$$Rentabilität = \frac{(Erlös - Kosten) - Abschreibung}{Kapitaleinsatz} \tag{6.2}$$

Die Rentabilität – auch als *return of investment (ROI)* bezeichnet – ist somit eine wichtige Kennzahl, insbesondere bei Projekten mit Informationssystemen (vgl. Abschnitt 6.3.1), weil damit eine Produktivitätssteigerung erzielt werden soll.

Die *Amortisationsrechnung* unterscheidet sich gegenüber den anderen drei genannten einperiodischen Vergleichsrechnungen dahingehend, dass sie die Zeit ermittelt (Amortisationsdauer), in der die Investitionssumme durch Einzahlungsüberschüsse gedeckt und Überschüsse erzielt werden können. Sie ist dahingehend sehr praktisch orientiert, da sie eine längere zeitliche Periode betrachtet.

Zusammenfassend können als Vorteile der statischen Investitionsrechnungen die leichte Handhabbarkeit und der verhältnismäßig geringe Aufwand der Informationsbeschaffung genannt werden. Dem gegenüber sind als Nachteile die zeitliche Struktur mit der lediglich einperiodischen Betrachtung – außer bei der Amortisationsrechnung – und den durchschnittlichen Erlösgrößen anzuführen, so dass diese nicht wirklich aussagekräftig sind.

6.2.3 Dynamische Verfahren

Die dynamischen Verfahren versuchen die Mängel der statischen Verfahren zu überwinden und sind um einiges komplexer. Der Grundgedanke der dynamischen Verfahren liegt in der Erfassung der Zeitstruktur mit den bedingten Ein- und Auszahlungen, die zu den entsprechenden Zeitpunkten mittels Zinseszinsrechung ab- bzw. aufgezinst werden. Es handelt sich hierbei um folgende Methoden (vgl. Kruschwitz, 2005, S. 44 ff.; Strunz, 1998, S. 304 ff.; Wöhe, 1996, S. 754 ff.):

- Kapitalwertmethode
- Annuitätsmethode
- Methode des internen Zinsfußes

Die *Kapitalwertmethode*, auch Nettobarwert genannt (engl. net present value, NPV), ist eine besondere Form des Endwertmodells unter der Annahme eines vollkommenen Kapitalmarktes. Es werden hierbei alle anfallenden Zahlungen auf einen bestimmten Bezugzeitpunkt $t = 0$ diskontiert (abgezinst). Damit ist die Maximierung des Endvermögens mit der des Kapitalwertes gleichzusetzen, so dass sich Investitionsentscheidungen, unter der Bedingung: $NPV \geq 0$, am maximalen Kapitalwert orientieren.

$$NPV = \sum_{t=0}^{T} z_t (1+i)^{-t} \tag{6.3}$$

Beim *Entnahmemodell*, ebenfalls unter der Annahme des vollkommenen Kapitalmarktes, werden wie bei der *Annuitätsmethode* die durchschnittlichen Einnahmen (Einnahmenannuität) und Ausgaben (Ausgabenannuität) einer Investition verglichen. Die positive Differenz der beiden Annuitäten wird neben der Verzinsung als zusätz-

licher Gewinn (Gewinnannuität) bezeichnet. Es gilt somit das Projekt auszuwählen, das positive Gewinnannuitäten und den größten positiven Kapitalwert hat.

Die *Methode des internen Zinsfußes* ist die umstrittenste der drei dynamischen Verfahren. Während die Methode in der Praxis gerne genutzt wird, wird die Anwendung in der Theorie als sehr fragwürdig angesehen. Der Ansatzpunkt der Methode des internen Zinsfußes liegt in der Forderung, dass bei Fremdfinanzierungen nicht nur die Finanzierungskosten gedeckt werden (Kapitalwert = 0), sondern auch ein Vermögenszuwachs stattfinden soll. Daher muss zur Vorteilhaftigkeit einer Investition der interne Zinsfuss r größer gleich dem Kalkulationszins i sein.

6.2.4 Beschränkung der Investitionsverfahren

Angemerkt sei, dass die dargestellten Investitionsverfahren unter den Annahmen einer sehr starken Vereinfachung zu sehen sind. Die Funktionsweise der Modelle konnten aufgrund ihrer Komplexität im Rahmen dieser Arbeit nur skizziert werden. Zur Vertiefung sei auf die angegebene Literatur verwiesen. Ferner wurden keine steuerlichen Aspekte und keine Planungsunsicherheiten berücksichtigt. Letzteres ist dahingehend als Risiko zu betrachten, dass Kosten nicht immer genau spezifiziert werden können und daher geschätzt werden müssen.

Es stellt sich die Frage, welche Verfahren für die Bewertung gebrauchstauglicher Informationssysteme zu verwenden sind. Exemplarisch werden nachfolgend zwei in der Literatur durchgesetzte Ansätze besprochen. Zum einen die statische Methode der Rentabilität (vgl. Abschnitt 6.3.1). Hier hat sich der englische Fachbegriff des *return of investment* (ROI) durchgesetzt. Zum anderen werden die Gesamtkosten einer Investition über den gesamten Lebenszyklus – *total cost of ownership* (TCO) – betrachtet (vgl. Abschnitt 6.3.2).

6.3 Bewertungsmethoden

6.3.1 Return of Investment (ROI)

Die finanzmathematische Berechnung der Rentabilität (vgl. Abschnitt 6.2.2) ermöglicht es, Kosten und Nutzen zum Kapitaleinsatz ins Verhältnis zu setzen und dadurch die Amortisationsdauer zu bestimmen, nach der sich eine Investition bezahlbar macht und Gewinne abwirft. In Bezug auf Informationssysteme kann somit anhand der Investitionen in die Anschaffung und deren erzieltem Nutzen eine Rentabilität ermittelt werden – *return of investment* (ROI). Während die Kostenseite durch die Anschaffungskosten bestimmt wird, ist der erzielte Nutzen näher zu beleuchten.

Betrachten wir den ROI-Ansatz aus der Benutzersicht, so lassen sich nach Wilson & Rosenbaum (2005, S. 216 f.) drei Kategorien bilden:

- *interner ROI* bezieht sich auf die Entwicklung eines Produktes bzw. Systems und gewährleistet eine kostengünstige Entwicklung
- *externer ROI* entsteht durch die Kundenverkäufe, gesteigertem Umsatz, verringerte Support-Kosten etc.
- *sozialer ROI* bezieht sich auf die Auswirkung der Verhältnisse unter den Teammitglieder eines Unternehmens.

All diese Ansätze ermöglichen es zu argumentieren, dass ein benutzergerechtes System bewertbar ist. Während erhöhte Umsätze direkt messbar sind, ist die Benutzerzufriedenstellung unter anderem durch erhöhte Benutzerproduktivität, indirekt auch durch höhere Umsätze messbar (vgl. Wilson & Rosenbaum, 2005, S. 242). Marcus (2005, S. 17 ff.) nennt und belegt eine Reihe von Aspekten die unter der Betrachtung gebrauchstauglicher Systeme einen positiven ROI signifikant beeinflussen. Nachfolgend seinen sie exemplarisch aufgezählt:

- *Reduzierte Kosten in der Entwicklung* durch Einsparungen von Entwicklungskosten und Entwicklungszeit, reduzierte Wartungskosten und eingesparte Redesign-Kosten

- *Erhöhter Umsatz im Verkauf* durch höhere Verkaufsraten und Produktverkäufe, Größe des Kundenkreises, Kundenbindung, mehr Attraktivität für Kunden und höherer Marktanteil
- *Verbesserte Effektivität in der Nutzung* durch erhöhte Erfolgsrate, reduzierte Benutzungsfehler, gesteigerte Produktivität, erhöhte Benutzer- und Arbeitszufriedenheit, Erhöhung der Einfachheit der Benutzung und des Lernens sowie erhöhtes Vertrauen in die Systeme und reduzierte Support-, Trainings- und Dokumentationskosten.

Ausgehend von Marcus' Betrachtung der gebrauchstauglichen Systeme und deren Auswirkungen auf den ROI, lassen sich klare Bezüge zu den Systemantwortzeiten herstellen. Schon in der Entwicklung ist es wichtig, dass Systemantwortzeiten beachtet werden. So konnte nachgewiesen werden, dass sich Entwicklungszeiten und damit auch deren Kosten reduzieren und die Produktivität der Programmierer steigern lassen, wenn die Antwortzeiten der zu bedienenden Systeme reduziert wurden (vgl. Abschnitt 6.1.3). Hierzu sei allerdings auch kritisch angemerkt, dass man den Zeitpunkt der damaligen Untersuchungen beachten muss. Des Weiteren wirft es die Frage auf, ob schnelle Antwortzeiten per se der Schlüssel zu mehr Produktivität sind. Barber & Lucas (1983) wiesen in Ihren Studien nach, dass kein linearer Zusammenhang zwischen Antwortzeit und der Produktivität besteht (vgl. Abb. 30). Vielmehr ist das zeitliche Optimum für die Benutzer nicht mit dem technisch möglichen zeitlichen Minimum gleichzusetzen. Eine zeitliche Reduzierung der Systemantwortzeit vom Optimum ausgehend sorgt für eine steigende Anfälligkeit von Flüchtigkeitsfehlern bei den Benutzern. Bei länger andauernden Systemantwortzeiten wird der Arbeitsprozess unterbrochen und die Benutzer vergessen, was sie eigentlich tun wollten.

Es zeigt sich somit, dass bei genau spezifizierten Systemen, sowohl in der Entwicklung als auch der Anwendung, durch optimierte Systemantwortzeiten Kosten reduziert und der Nutzen für die Benutzer erhöht werden können, was sich positiv auf den Gesamt-ROI auswirkt.

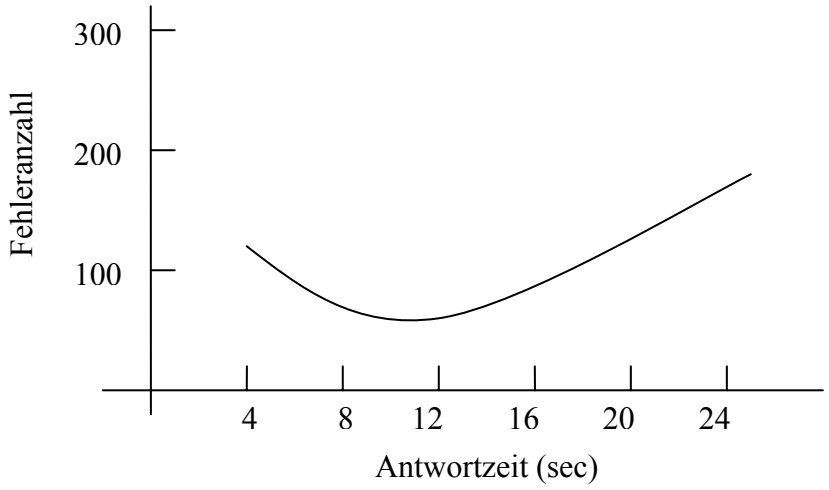


Abbildung 30 Systemantwortzeiten und Fehleranfälligkeit der Benutzer (vgl. Barber & Lucas, 1983, S. 978)

Das Problem, das durch einen solchen Messansatz offenbar wird, ist die schon angesprochene Messbarkeit des Nutzens. Sicherlich lassen sich eingesparte Arbeitszeiten von Benutzern durch reduzierte Systemantwortzeiten bestimmen und mit zugrunde gelegten Stundenvergütungen fakturieren, so dass sich ein geldwerter Vorteil bestimmen lässt. Allerdings muss kritisch hinterfragt werden, ob dieser zeitliche Vorteil wirklich umgesetzt wird, bzw. die erhöhte Benutzerproduktivität weiteren geldwerten Vorteil mit sich bringt. Unbestreitbar ist die gezeigte erhöhte Zufriedenstellung der Benutzer.

Der dargestellte Ansatz geht von der vereinfachten Annahme eines Beschäftigten in einem Unternehmen aus. Lässt sich auch für den Kunden, der beispielsweise über den zeitlich optimierten Webshop einer Unternehmung eine Bestellung aufgibt, ein geldwerter Vorteil wie im betriebswirtschaftlichen Umfeld erzielen? Immerhin ist es ihm möglich, seine Transaktion schneller abzuschließen, dadurch Verbindungskosten zu reduzieren und Zeit für andere Aktivitäten zu gewinnen. Im Gegensatz dazu würde der potentielle Kunde bei langen Antwortzeiten eventuell keine Bestellung oder bei einem Wettbewerber aufgeben. Dadurch muss der Kunde zusätzliche Zeit und Kosten investieren und gleichzeitig werden dem zuerst ausgewählten Unternehmen Umsätze und Gewinne fehlen, wodurch sich der ROI der Unternehmung verschlechtert.

6.3.2 Total cost of ownership (TCO)

Rosenberg (2004, S. 22 ff.) kritisiert den im vorigen Abschnitt behandelten ROI-Ansatz als zeitlich zu kurzsichtig. Er gibt zu bedenken, dass ein entwickeltes Produkt seinen Wert erst langfristig während des gesamten Produktlebenszyklus zeigt. Dieser ganzheitliche Ansatz, die Gesamtkosten einer Investition über den gesamten Lebenszyklus einer IT-Anwendung (vgl. Abb. 31) zu bestimmen, ist der *Total Cost of Ownership*, der von der Gartner Group entwickelt wurde. Die Charakteristik dieses Ansatzes liegt in der Betonung der Gesamtkosten und ist auch ambivalent zu beurteilen. Während die Fokussierung lediglich bei den Kosten liegt und damit den Nutzen von Projekten außer Acht lässt (vgl. Hinderberger, 2003, S. 29), ist gleichzeitig die Betrachtung der Gesamtkosten auch positiv zu beurteilen. Als Grund hierfür sind die Anschaffungsinvestitionen zu nennen, die nicht nur Hard- und Software umfassen, sondern auch die Einführungs-, Wartungs- und Betriebskosten.

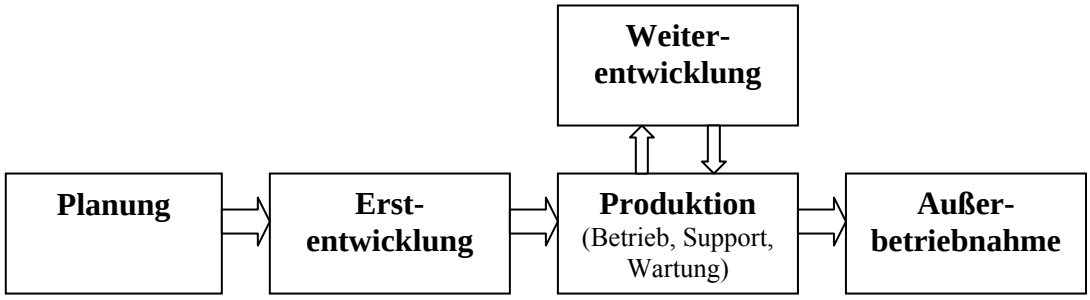


Abbildung 31 Lebenszyklusphasen einer IT-Anwendung (vgl. Zarnekow, Scheeg & Brenner, 2004, S. 182)

Der Artikel von Zarnekow, Scheeg & Brenner (2004) zeigt, dass der ganzheitliche Ansatz des TCO in die richtige Richtung geht. Sie fordern, dass lebenszyklusorientierte Kostenrechnungsmodelle entwickelt werden müssen, um Fehlentscheidungen zu verhindern. In diesem Zusammenhang weisen sie auch auf die Bedeutung der bilanziellen Aktivierung von Software hin, die durch Bilanzierungsvorschriften wie die International Accounting Standards (IAS) möglich ist. Zarnekow et al. kommen in ihrer Untersuchung zu dem Ergebnis, dass die Kostenaspekte der – wie sie es nennen – Produktion (Betrieb, Support, Wartung) und die Weiterentwicklung bei einer Produktdauer von 5 Jahren knapp 80 % der Lebenszykluskosten ausmachen und damit häufig unterschätzt werden.

Insbesondere der Betrieb ist von besonderer Bedeutung. Hierbei gilt es Grenzwerte für einen ordnungsgemäßen Betrieb zu definieren und die Werte zu überwachen. Dies ermöglicht es eine Überlastung bzw. einen Ausfall von Systemkomponenten frühzeitig zu bemerken und gegenzusteuern. Menascé et al. (2004, S.110 ff.) weisen in diesem Zusammenhang auf die Spezifikation von Service Level Agreements (SLA) hin, die zwischen Leistungserbringern und -nehmern geschlossen werden. In diesen SLAs lassen sich dann auch Systemantwortzeiten spezifizieren. Dies ist insbesondere bei unternehmenskritischen Anwendungen wie z.B. Flugbuchungssystemen geboten. Zu bedenken gilt es, dass die Einhaltung hoher Service Levels auch meistens mit hohen Kosten verbunden ist.

Der TCO-Ansatz bietet die Möglichkeit, die Gesamtkosten eines Anwendungssystems zu betrachten. Die Gartner Group (2003) differenziert in ihrem TCO Modell *Distributed Computing – Chart of Accounts* für Client/Server-Umgebungen zwischen direkten und indirekten Kosten (vgl. Tab 11). Während die direkten Kosten direkt messbar sind, werden die indirekten Kosten meist weder beachtet noch gemessen.

Kostenart	**Unterpunkte**
Direkte Kosten	
Hardware und Software	Hardware, Software, IS Hardware, IS Software
Betrieb	Technischer Service (Client, Server, Netzwerk), Planung und Prozessmanagement, Datenbank-Management und Administration, Service Desk
Verwaltung	Finanzen und Verwaltung, IS Training, Endbenutzer-training
Indirekte Kosten	
Endbenutzeroperationen	Gegenseitige Unterstützung, gelegentliches und formales Lernen, Datei- und Daten-Management, Anwendungsentwicklung, Endbenutzer-Zufriedenstellung
Ausfallzeit	Geplante und ungeplante Ausfallzeiten

Tabelle 11 TCO Model Distributed Computing Chart of Accounts (vgl. Gartner, 2003)

Die direkten Kosten umfassen Hard- und Softwarekosten, Betriebskosten und Verwaltungskosten. Die indirekten Kosten, die meist durch die direkten bedingt sind, umfassen Endbenutzeroperationskosten und Ausfallzeiten, die nachfolgend detailliert betrachtet werden sollen.

Unter *Hardware und Software Kosten* werden die Aufwendungen für die Anschaffung von Hardware und Software, Upgrades und die erforderliche Infrastruktur zusammengefasst. *Betriebskosten* beinhalten alle laufenden Kosten für Personal (Technik, Planung, Administration und Servicedesk) und Betriebsausgaben. Der technische Service wird differenziert zwischen Client, Server und Netzwerk. Die V*erwaltungskosten* setzen sich aus Kosten der IT-Leitung und den Trainingskosten (Entwurf und Benutzerschulungen) zusammen.

Die indirekten Kosten der *Endbenutzerkosten* sind meist versteckt und werden in der Kostenrechnung sonst meist nicht mit berücksichtigt. Sie sind gerade daher besonders wichtig, weil sie die wahren Kosten repräsentieren, die es ermöglichen, die Auswirkungen und Produktivität durch die Investitionen in die Informationstechnik zu messen. Hierunter fallen Kostenaspekte wie gegenseitige Hilfestellung, Trainingsstunden, Selbstevaluation des Systems, Management von Daten und Dateien sowie die Entwicklung kleiner Applikationen (Skripte) und die Zufriedenstellung mit der Benutzung der Informationstechnik. Abschließend seien die Ausfallzeiten zu nennen, die sowohl geplant als auch ungeplant an den einzelnen Systemkomponenten auftreten können und als verlorene Produktivität anzusehen sind.

David, Schuff und Louis (2002, S. 101 ff.) zeigen Möglichkeiten zur Reduzierung der TCO auf. Sie argumentieren, dass durch eine Zentralisierung von Software und Netzwerkverwaltung sowie die Standardisierung der Hardware und Softwarekonfigurationen bei den Endnutzern administrative Kosten eingespart werden können. Sie weisen aber auch darauf hin, dass durch die Zentralisierung sich die Netzlast erhöht und der Datendurchsatz im Netz geringer wird und sich damit auf die gesamte Systemleistung auswirkt. Durch den höheren Datenverkehr können Verzögerungen auftreten, die es durch eine optimierte Bandbreitenanforderung zu minimieren gilt.

6.3.3 Implikation für Systemantwortzeiten

Es zeigt sich, dass Systemantwortzeiten in dem TCO-Modell als ein Randaspekt betrachtet werden. Eine differenzierte Betrachtung des Punktes Ausfallzeiten lässt einen Bezug auf die Systemantwortzeiten zu. Dies unter der Betrachtungsweise, dass zu lange Systemantwortzeiten in einem komplexen Anwendungssystem zu Ausfällen führen können und sich dadurch auf die indirekten Kosten auswirken.

Die ROI-Methode bietet dagegen Ansatzpunkte, die Kosten der (Un-)Benutzbarkeit durch nicht optimierte Systemantwortzeiten zu ermitteln. Es zeigt sich eine Parallelität mit dem von Rosenberg (2004, S. 24) als „landmark book“ charakterisierten Buch *Cost Justifying Usability* (Bias & Mayhew, 1994 und 2005). Dort wird der Standpunkt vertreten, dass die geldwerte Beurteilung der Gebrauchstauglichkeit von Software – und damit auch von Informationssystemen allgemein – den meisten Einfluss auf den Entwicklungsprozess nehmen kann (Bias & Karat, 2005, S. 13).

Es lässt sich damit festhalten, dass Systemantwortzeiten eine messbare Größe sind, die als einzuhaltende Grenzwerte zu spezifizieren sind. Eine monetäre Bewertung ist zum einen über die Argumentation der Ausfallzeiten im TCO-Ansatz möglich. Zum anderen – und damit differenzierter – über den ROI-Ansatz. Die Investition wird zielgerichtet auf die Optimierung der Systemantwortzeiten getätigt und ermöglicht dadurch eine reduzierte Ausfallzeit, die eine höhere Produktivität mit sich führt.

7 Schlussbetrachtung

7.1 Zusammenfassung

Im Rahmen dieser Diplomarbeit wurden die Systemantwortzeiten als ein Aspekt der Software-Ergonomie und der Wirtschaftsinformatik dargestellt. Zu Beginn wurde ein Überblick über die Entwicklung der Computer- und Informationstechnologie gegeben, um danach die Wissenschaftsstandpunkte der Software-Ergonomie und der Wirtschaftsinformatik darzulegen und deren Verknüpfungen aufzuzeigen.

Von der ergonomischen Seite wurde das Zeitverhalten in interaktiven Systemen betrachtet. Als erstes wurden die Systemantwortzeiten definiert und der aktuelle Forschungsstand dargestellt. Danach wurde auf die Benutzersicht eingegangen. Angefangen bei den kognitiven Leistungen, dem Zusammenwirken von Beanspruchung und Belastung und den Arbeitsplatztypen, über den Stress am Bildschirmarbeitsplatz bis hin zu den soziographischen Aspekten mit den Merkmalen Benutzergruppen, Alter, besondere Anforderungen und Erfahrungen der Benutzer. Mit einer Analyse von Normen im Bezug auf die Systemantwortzeiten schloss die Betrachtung ab.

Aus der technischen Sicht wurden die Systemantwortzeiten in Anwendungssystemen detailliert betrachtet. Hierzu wurden Grundlagen anhand von Qualitätskriterien und Leistungskenngrößen mit physikalischen Eigenschaften gelegt sowie Modellierungsaspekte auf die Leistungen der Netzknoten behandelt. Nachfolgend wurden sukzessiv erst Einzelsysteme, aufgeteilt nach Hard- und Software, verteilte Systeme mit Client-Server-Architektur, lokalen und Weitverkehrsnetzen, sowie mobile Systeme betrachtet. Abgerundet wurde dies mit einer Betrachtung des Internets.

Im letzten Schritt wurden die wirtschaftlichen Aspekte der Systemantwortzeiten behandelt. Hierzu wurde sich mit der Wirtschaftlichkeit von Informationssystemen und Verfahren der Investitionsrechnung als Entscheidungsgrundlage beschäftigt. Darauf aufbauend wurden die Bewertungsmethoden des ROI und des TCO auf die Systemantwortzeiten angewendet.

7.2 Fazit

In der Arbeit konnte herausgearbeitet werden, dass Systemantwortzeiten eine Problematik darstellen, die interdisziplinär zu bearbeiten ist. Systemantwortzeiten wirken sich direkt auf die Benutzer aus und beeinflussen diese in ihrer Arbeit. Die Systeme sind seitens der Software-Ergonomie daher so zu gestalten, dass sie den Erwartungen der Benutzer entsprechen. Hierzu gilt es im Sinne der Transparenz immer über den aktuellen Systemzustand und die zu erwartende Verzögerung zu informieren, sowie Möglichkeiten der Steuerbarkeit einzuräumen, um Prozesse ggf. zu beschleunigen, zu verlangsamen oder gar abzubrechen. Es muss davon ausgegangen werden, dass der Benutzer sich nicht immer über die komplexe Infrastruktur bewusst ist.

Aus der technischen Sicht gilt es, die Systeme schon während der Entwicklung im Bezug auf ein optimales Systemantwortzeitverhalten zu modellieren und zu analysieren. Aufgrund der Komplexität und Komponentenvielfalt – insbesondere in verteilten Systemen – sind ein umfangreiches Wissen und analytische Werkzeuge erforderlich, um Schwachstellen und Engpässe zu entdecken und zu beheben.

Ferner gilt es eine monetäre Bewertung der Systemantwortzeiten vorzunehmen – und zwar sowohl in der Anschaffung von technischer Infrastruktur, als auch im laufenden Betrieb, da lange Systemantwortzeiten die Produktivität reduzieren und die Kosten steigen lassen. Es gilt zu bedenken, dass neben der Investition in die Optimierung der einzelnen Komponenten auch das Gesamtsystem betrachtet werden muss. Die wirtschaftlichen Überlegungen der Investition in die Optimierung der Systemantwortzeiten sind im Verhältnis zum jeweiligen Anwendungskontext zu betrachten. Die Systemantwortzeiten wirken sich über die Interaktionsschnittstelle direkt auf die Benutzer aus, sodass die Kosten der Unbenutzbarkeit sehr wohl eine Thematik der Wirtschaftsinformatik sind.

Es zeigte sich ferner, dass Systemantwortzeiten nach verschiedenen Anwendungskontexten differenziert werden können und damit verschiedene Leitbilder der Arbeit zu Grunde gelegt werden können:

1. Desktop-Systeme (Einzelsysteme, als auch Netzsysteme)
2. Mobile Systeme
3. Web

Zum einen gibt es Desktop-Systeme, die sowohl als Einzelsysteme als auch im Netzverbund genutzt werden können. Während es beim Einzelsystem für den Benutzer verständlich ist, dass nur seine Hard- und Software die Systemantwortzeit beeinflussen, sind es bei vernetzten Systemen mehrere Komponenten. Hier hat der Benutzer keine direkten Manipulationsmöglichkeiten, da er von der Netzverbindung und -auslastung sowie der technischen Infrastruktur der anderen Rechner abhängig ist.

Bei mobilen Systemen stellen sowohl die kabellose Verbindung mit deren Eigenschaften sowie die asynchrone Nutzung einen Schwerpunkt dar. Der Benutzer sollte über den Verbindungszustand und -stärke kontinuierlich informiert werden, damit er hier eine umgehende Rückmeldung über den Systemzustand erhält. Ferner wird die Synchronisation zwischen dem asynchron genutzten mobilen System und einen Server durch den Benutzer nicht zu unterbinden sein, so dass hier – je nach Synchronisationsaufbau und Verbindungsleistung – eine lang anhaltende Systemantwortzeit erforderlich ist, die den Benutzer in seinem direkten Arbeitsfluss bremst.

Im Web gibt es die Besonderheit, dass eine Vielzahl von Komponenten auf die Systemantwortzeit einwirken, so dass es hier z.B. schon nur durch eine Serverüberlastung zu sehr langen Systemantwortzeiten kommt, die einen Gesamtsystemausfall bewirken können. Durch die Nutzung des Webs im e-Commerce Bereich kommt der monetären Bewertung eine besondere Bedeutung zu, da lange Systemantwortzeiten verminderte Umsätze bedeuten. Der Benutzer ist immer über die (zu erwartende) Systemantwortzeit prospektiv zu informieren, damit er sich darauf vorbereiten kann.

Die vorhandenen Forschungsergebnisse ermöglichen noch keine generalisierten Aussagen im Bezug auf optimale Systemantwortzeiten. Sie sind noch zu variabel und nicht auf den jeweiligen Anwendungskontext bezogen. Primär gilt es die Systemantwortzeiten und deren Varianz zu minimieren, um sich dem Optimum zu nähern.

7.3 Ausblick

Als ein wesentliches Ergebnis dieser Arbeit kann festgehalten werden, dass es trotz vieler Studien und Untersuchungen noch immer viele Fragen im Bezug auf die Systemantwortzeiten gibt. Lassen sich die Befunde aus den 1960er und 1980er Jahren trotz der technischen Entwicklung in die heutige Zeit übertragen? Die Ergebnisse von Nah (2004) bestätigen die Thesen von Miller (1968). Es gilt allerdings zu bedenken, dass die technologische Entwicklung immer komplexere Systeme ermöglicht und die Benutzer erfahrener werden und andere Erwartungen haben. Dies wirft die Frage auf, wie sich die Erwartungen der Benutzer verändert haben. Sind Übertragungen der Zeiterwartungen von Desktop Systeme auf verteilte Systeme oder mobile Systeme möglich? Sind die Toleranzschwellen bei erfahrenen Benutzern anders als bei unerfahrenen? Und falls ja, ergibt sich die Frage, durch welche Erfahrungen diese Toleranzschwellen entstehen und sich verschieben lassen.

Für die Software-Ergonomie kommt daher die besondere Bedeutung zu, dass sie die Interaktionsschnittstelle für das (komplexe) System und den Benutzer darstellt und entsprechend modelliert werden muss. Zwar lassen sich in Meyer, Vogt & Glier (2005a,b) in Anlehnung an Hüttner et al. (1995) Empfehlungen finden, die Rückmeldungen über die Systemantwortzeiten fordern, doch werfen diese Empfehlungen Fragen der empirischen Belegbarkeit auf. So soll nach spätestens 10 Sekunden Systemantwortzeit eine Anzeige über den Systemzustand und nach spätestens 30 Sekunden eine Anzeige über die verbleibende Wartezeit erscheinen.

Es gilt herauszuarbeiten, ab welchem Zeitpunkt und in welchem Anwendungskontext die Benutzer über ausstehende Systemantwortzeiten zu informieren sind und ob sich Maximalwerte manifestieren lassen. Interessant wäre es zu untersuchen, ob durch Wartezeitanzeigen die maximale Wartebereitschaftstoleranz erhöht werden kann. Des Weiteren muss untersucht werden ob die Informationen über die Systemantwortzeit überhaupt als hilfreich eingestuft werden und sie nicht noch zusätzlich verwirren. In diesem Zusammenhang sei zu bedenken, dass Verarbeitungsprozesse län-

ger oder kürzer dauern können, als ursprünglich angenommen wurde. Welche Möglichkeiten bieten sich, dies den Benutzern glaubhaft mitzuteilen?

Ebenso ist auch die schon erwähnte Studie von Barber & Lucas (1983) mit dem Ergebnis des U-förmigen Verlaufes in der heutigen Zeit kritisch zu hinterfragen. Ist die damals gezeigte Verschiebung von optimaler und minimaler Systemantwortzeit heute noch gegeben? Es müsste betrachtet werden, ob diese zweiseitige Fehleranfälligkeit – Flüchtigkeitsfehler bei schnellen und Gedächtnisproblemen bei langsamen Systemantwortzeiten – heute noch zutrifft, oder ob es eine Verschiebung der optimalen Systemantwortzeit hin zu der technisch minimal möglichen Systemantwortzeit gibt, weil die Systeme doch schon so schnell sein könnten, wie die Benutzer es erwarten.

Neuere Studien beschäftigen sich mit den Systemantwortzeiten im Web; ältere Studien mit Mainframe-Systemen. Eine Gruppe von Systemen, die insbesondere im betrieblichen Umfeld sehr wichtig ist, wird fast kaum erwähnt – Client-Server-Systeme. Da stellt sich die Frage, warum es zu diesem Bereich kaum Studien und Ergebnisse gibt. Schließlich sind Client-Server-Architekturen wesentlich einfacher zu überwachen als Web-Applikationen. Zwar sind Web-Systeme eigentlich auch nichts anderes als Client-Server-Systeme, nur ist die Systemantwortzeit aufgrund der Netzarchitektur nicht so leicht zu überwachen, wie im internen betrieblichen Kontext. Über die Gründe des Nichtvorhandenseins von Client-Server-Systemantwortzeiten lässt sich nur orakeln. Begnügen sich Anbieter und Kunden lediglich mit Richtwerten oder passen sie sogar – was eigentlich verwerflich wäre – ihre Geschäftsprozesse den längeren Systemantwortzeiten an, anstatt diese zu reduzieren? Welche Gründe, besondere Interessen oder sogar Desinteressen gibt es, dass es hierzu kaum Veröffentlichungen gibt?

Die Verknüpfung der Software-Ergonomie mit der Wirtschafsinformatik wurde aufgezeigt und Ansätze zur monetären Bewertung besprochen. Dass damit die Kosten der Unbenutzbarkeit ein Thema der Wirtschaftsinformatik sind, ist evident. Allerdings fehlt es an spezifischen Bewertungsinstrumenten, da eine monetäre Bewertung

nur über die Hilfskonstruktion von Systemausfällen als Schaden zu klassifizieren ist bzw. über die Rückzahlungen von Investitionen. Einen Ansatzpunkt bieten die Service Level Agreements, in denen Leistungsparameter – und damit auch Systemantwortzeiten zwischen Leistungserbringer und -nehmer – festgeschrieben werden. Für den Leistungsnehmer bietet sich dann im Falle des Ausfalles die Möglichkeit den entstandenen Schaden mittels Vertragsstrafenzahlungen durch den Leistungserbringer zu reduzieren.

Somit zeigt sich, dass es noch sehr viele offene Fragen im Bezug auf die Systemantwortzeiten gibt, die einer weiteren Klärung und Untersuchung bedürfen. Es gilt Lösungsansätze herauszuarbeiten, die die Systemantwortzeiten auf ein kontinuierliches Optimum – dem technischen Minimum? – reduzieren, um den Benutzern ein beeinträchtigungsfreies Arbeiten zu ermöglichen. Im Falle von länger andauernden Operationen gilt es, den Benutzern transparent und kontrollierbar über den aktuellen Systemzustand und die noch abzuwartende Systemantwortzeit zu informieren und ihnen Steuerungsmöglichkeiten anzubieten.

Literaturverzeichnis

Abts, D., Mülder, W. (2004): Grundkurs Wirtschaftsinformatik, 5. Auflage, Vieweg, Wiesbaden

Alexander, J.-M. (1986): Psychologische Beanspruchung und Leistung in Abhängigkeit von Systemresponsezeit am Bildschirmarbeitsplatz, Bergische Universität - GH - Wuppertal

ANSI X3.102 (1992): Data Communication systems and services – user-oriented performance parameters

Arbeitsschutzgesetz (2004): Gesetz über die Durchführung von Maßnahmen des Arbeitsschutzes zur Verbesserung der Sicherheit und des Gesundheitsschutzes der Beschäftigten bei der Arbeit, BGBl I 1996, 1246 vom 07.08.1996, Stand: Zuletzt geändert durch Art. 11 Nr. 20 G vom 30.07.2004 I 1950
http://bundesrecht.juris.de/bundesrecht/arbschg/
(zuletzt besucht am 16.11.2005)

Barber, R.E., Lucas, H.C. (1983): System response time, operator productivity, and job satisfaction. In: Communications of the ACM, 26, S. 972-986

Bauknecht, K., Tjoa, A.M., Draxler, C. (1991): Informationssysteme. Institutsbericht Nr. 91.07, Institut für Informatik, Universität Zürich, Zürich

Berbner, R., Heckmann, O., Mauthe, A., Steinmetz, R. (2005): Eine Dienstgüte unterstützende Webservice-Architektur für flexible Geschäftsprozesse. In: Wirtschaftsinformatik, 47 (4), S. 268-277

Bhalekar, A., Baras, J. (2004): Cumulative Caching for reduced user-perceive latency for www transfers on networks with satellite links. In: Dini, P., Lorenz, P., de Souza, J.N. (Eds.) First International Workshop, SAPIR 2004, Proceedings, Springer, Berlin, S. 179-186

Bias, R.G., Karat, C.M.. (2005): Justifying Cost-Justifying Usability. In: Bias, R.G., Mayhew, D.J. (Eds.): Cost-Justifying Usability. An Update for the Internet Age, Morgan Kaufmann, San Francisco, S. 1-16

Bildschirmarbeitsverordnung (2003): Verordnung über Sicherheit und Gesundheitsschutz bei der Arbeit an Bildschirmgeräten, BGBl I 1996, 1843 vom 04.12.1996, Stand: Zuletzt geändert durch Art. 304 V vom 25.11.2003 I 2304 http://bundesrecht.juris.de/bundesrecht/bildscharbv/ (zuletzt besucht am 16.11.2005)

Bödeker, W. (2003): Psychische Belastungen in der Arbeitswelt – Ergebnisse internationaler Studien. In: Schriftenreihe der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin: Psychische Belastung am Arbeitsplatz, Wirtschaftsverlag NW, Bremerhaven, S. 122-134

Bolch, G., Greiner, S., de Meer, H., Trivedi, K.S. (1998): Queueing Networks and Markov Chains. Modelling and Performance Evaluation with Computer Science Applications, John Wiley, New York

Bouch, A., Kuchinsky, A., Bhatti, N. (2000a): Quality is in the Eye of the Beholder: meeting Users' Requirements for Internet Quality of Service. Proceedings of the CHI 2000 Conference on Human factors in computing systems, ACM, S. 297-304

Bouch, A., Kuchinsky, A., Bhatti, N. (2000b): Integrating User-Perceived Quality into Web Server Design. In: Computer Networks, 33, S.1-16

Boucsein, W., Greif, S., Wittekamp, J. (1984): Systemresponsezeiten als Belastungsfaktor bei Bildschirm-Dialogtätigkeiten. In: Zeitschrift für Arbeitswissenschaft, 38, S.113-122

Boucsein, W. (1987): Psychophysiological investigation of stress induced by temporal factors in Human-Computer Interaction. In: Frese, M., Ulich, E, Dzida, W. (Eds.): Psychological issues of Human Computer Interaction in the work place, Elsevier Science Publishers B.V., North-Holland, S.163-181

Bräutigam, L., Schneider, W. (2003): Projektleitfaden Software-Ergonomie, InvestitionsBank Hessen, Wiesbaden

Brynjolfsson, E., Hitt, L.M. (2003): Computing Productivity: Firm-Level Evidence, eBusiness@MIT Working Paper 139, http://ssrn.com/abstract=290325 (zuletzt besucht am 15.11.2005)

Bubb, H. (1993): Ergonomie. In: Schorr, A (Hrsg.): Handwörterbuch der angewandten Psychologie, Deutscher Psychologen Verlag, Bonn, S.194-198

Burmester, M. (2001): Optimierung der Erlern- und Benutzbarkeit von Benutzungsschnittstellen interaktiver Hausgeräte auf der Basis der speziellen Anforderung älterer Menschen, VDI Verlag, Düsseldorf

Bush, V. (1945): As we may think. In: Atlantic Monthly, 176, S. 101-108

Bux, W. (1984): Performance issues in local-area networks. In: IBM Systems Journal 23 (4), S. 351-374

Çakir, A., Reuter, H-J., von Schmude, L., Armbruster, A. (1978): Anpassung von Bildschirmarbeitsplätzen an die physische und psychische Funktionsweise des Menschen, Der Bundesminister für Arbeit und Sozialordnung, Bonn

Carbonell, J.R., Elkin, J.I., Nicherson, R.S. (1968): On the psychological importance of time in time sharing system. In: Human Factory, 10, S. 135-142

Card, S.K., Moran, T.P., Newell, A. (1983): The psychology of human-computer interaction, Lawrence Erlbaum Associates, Hillsdale

Carr, N.G. (2003): IT doesn't matter. In: Harvard Business Review, 81 (5), S. 41-49

Carr, N.G. (2004): Does IT matter? Information technology and the corrosion of competitive advantage, Harvard Business School Publishing Corporation, Boston

Ceruzzi, P. E. (2003): A History of modern computing, MIT Press, Landsberg

Coulouris, G., Dollimore, J., Kindberg, T. (2005): Distributed systems. Concepts and design, 4th edition, Addison-Wesley, Harlow

Cremonesi, P., Serazzi, G. (2002): End-to-End Performance of Web Services. In: Calzarossa, M., Tucci, S. (Eds.): Performance 2002, Springer, Berlin, S. 158-178

Czaja, S.J., Sharit, J. (1993): Stress Reactions to Computer-Interactive Tasks as a Function of Task Structure and Individual Differences. In: International Journal of Human-Computer Interaction, 5 (1), S.1-22

David, J. S., Schuff, D., Louis, R.S. (2002): Managing your IT total cost of ownership. In: Communications of the ACM, 45 (1), S. 101-106

Dehning, W., Essig, H., Maass, S. (1978): Zur Anpassung virtueller Mensch-Rechner-Schnittstellen an Benutzererfordernisse im Dialog, Universität Hamburg, Fachbereich Informatik, Bericht IFI-HH-B-50/78

Dellmann, K., Pedell, K.L. (1994): Controlling von Produktivität, Wirtschaftlichkeit und Ergebnis, Schäffer-Poeschel, Stuttgart

DIN EN ISO 9241 (1996): Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten, Teil 10: Grundsätze der Dialoggestaltung

DIN EN ISO 9241 (1998): Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten, Teil 11: Anforderungen an die Gebrauchstauglichkeit – Leitsätze

DIN EN ISO 10075 (2000): Ergonomische Grundlagen bezüglich psychischer Arbeitsbelastung, Teil 1: Allgemeines und Begriffe

DIN EN ISO 10075 (2000): Ergonomische Grundlagen bezüglich psychischer Arbeitsbelastung, Teil 2: Gestaltungsgrundsätze

Dix, A. J. (1987): The myth of the infinitely fast machine. In: Diaper, D., Winder, R. (Eds.): People and Computers, Volume III – Proceedings of HCI'87, Cambridge University Press, Cambridge, S. 215-228

Dix, A. J. (2003): Network-based interaction. In: Jacko, J.A., Sears, A. (Eds.): The human-computer interaction handbook, Erlbaum, Mahwah, S. 331-357

Dix, A. J., Finlay, J., Abowd, G.D., Beale, R. (2004): Human-computer interaction, 3rd edition, Pearson, Harlow

Doherty, W.J., Kelisky, R.P. (1979): Managing VM/CMS systems for user effectiveness. In: IBM Systems Journal, 18 (1), S. 143-163

Doherty, W.J., Thadani, A.J. (1982): The economic value of rapid response time. http://www.vm.ibm.com/devpages/JELLIOTT/evrrt.html (zuletzt besucht am 27.10.2005)

Dyson, P., Longshaw, A. (2004): Architecting enterprise solutions. Patterns for high-capability internet-based systems, Wiley, Chichester

Eversmann, L. (2002): Erkenntnisziele der Wissenschaft „Wirtschaftsinformatik“. In: Wirtschaftsinformatik, 44, S. 91-96

Fehling, G., Jahnke, B. (1999): Wirtschaftsinformatik und Ethik – Komplementarität oder Konkurrenz? In: Informatik-Spektrum, 22, S. 197-205

Ferstl, O.K., Sinz, E.J. (2001): Grundlagen der Wirtschaftsinformatik, 4. Auflage, Band I, Oldenbourg, München [u.a.]

Field, A.J., Harrison, P.G. & Parry, J. (1998): Response Time in Client-Server Systems. In: Puigjaner, R., Savino, N.N., Serra, B. (Eds.): Computer Performance Evaluation. Modelling Techniques and Tools, Springer, Berlin

Fink, A., Schneidereit, G., Voß, S. (2001): Grundlagen der Wirtschaftsinformatik, Physica, Heidelberg

Frese, M. (1983): Der Einfluss der Arbeit auf die Persönlichkeit. Zum Konzept des Handlungsstils in der beruflichen Sozialisation. In: Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 3 (1/83), S. 11-28

Frese, M., Brodbeck, F.C. (1989): Computer in Büro und Verwaltung. Psychologisches Wissen für die Praxis, Springer, Berlin

Garg, P.K., Eshghi, K., Gschwind, T., Haverkort, B., Wolter, K. (2002): Enabling Network Caching of dynamic web objects. In: Field, T. (Eds.) Tools 2002, S. 329-338

Gartner (2003) Distributed Computing. Chart of Accounts. http://www.gartner.com/ 4_decision_tools/modeling_tools/costcat.pdf (zuletzt besucht am 16.09.2005)

Geis, T., Dzida, W., Redtenbacher, W. (2004): Specifying usability requirements and test criteria for interactive systems, Wirtschaftsverlag NW, Bremerhaven

Greif, S. (1991): Stress in der Arbeit. Einführung und Grundbegriffe. In: Greif, S., Bamberg, E. und Semmer, N. (Hrsg.): Psychischer Stress am Arbeitsplatz, Hogrefe, Göttingen, S. 1 – 28

Gomez, G., Sanchez, R., Cuny, R. Kuure, P., Paavonen, T. (2003): Packet Data Services and End-user Performance. In: Halonen, T., Romero, J, Melero, J. (Eds.): GSM, GPRS and EDGE performance, 2nd edition, Wiley, Chichester, S. 307-350

Granit, R. (1985): Comments on time in action and perception. In: Human Neurobiology, 4, S. 61-62

Gray/Reuther (1993): Transaction processing: Concepts and techniques, Morgan Kaufmann, San Mateo

Griese, J. (1982): Software-Ergonomie. Das Aktuelle Schlagwort. In: Informatik-Spektrum, 5, S. 124-125

Gros, E. (1994): Analyse von Arbeitstätigkeiten: Ermittlung von Belastung und Beanspruchung am Arbeitsplatz. In: Gros, E.: Anwendungsbezogene Arbeits-, Betriebs- und Organisationspsychologie: eine Einführung, Verlag für Angewandte Psychologie, Göttingen, S. 95-122

Gründler, A. (1997): Computer und Produktivität: Das Produktivitätsparadoxon der Informationstechnologie, Gabler, Wiesbaden

Hacker, W. (1998): Allgemeine Arbeitspsychologie: psychische Regulation von Arbeitstätigkeiten, Huber, Bern

Halonen, T., Romero, J, Melero, J. (2003): GSM, GPRS and EDGE performance, 2nd edition, Wiley, Chichester

Hardenacke, H., Peetz, W., Wichardt, G. (1985): Arbeitswissenschaft, Hanser, München

Heinecke, A.M. (2004): Mensch-Computer-Interaktion, Fachbuchverlag Leipzig, München

Heinrich, L.J. (2001): Wirtschaftsinformatik: Einführung und Grundlegung, 2. Auflage, Oldenbourg, München

Heinrich, L.J., Lehner, F. (2002): Informationsmanagement, 8. Auflage, Oldenbourg, München

Heinzl, A., König, W., Hack, J. (2001): Erkenntnisziele der Wirtschaftsinformatik in den nächsten drei und zehn Jahren. In: Wirtschaftsinformatik, 43, S. 223-233

Herczeg, M. (2005): Softwareergonomie, 2. Auflage, Oldenbourg, München

Hinderberger, R. (2003): Usability als Investition. In: Heinsen, S., Vogt, P. (Hrsg.): Usability praktisch umsetzen, Hanser, München

Hoch, D. J. (1995): Die Wirtschaftsinformatik muß sich mehr vornehmen. In: Wirtschaftsinformatik, 39, S. 328-329

Holling, H. (1989): Psychische Beanspruchung durch Wartezeiten in der Mensch-Computer Interaktion, Springer, Berlin

Hübner, G. (2002): Stochastik, 3. Auflage, Vieweg, Braunschweig

Hui, M.K., Zhou, L. (1996): How does waiting duration information influence customers' reactions to waiting for services? In: Journal of Applied Social Psychology, 26, S. 1702-1717

Hüttner, J., Wandke, H., Rätz, A. (1995): Benutzerfreundliche Software. Psychologisches Wissen für die ergonomische Schnittstellengestaltung. Paschke, Berlin. http://www.bmp.de/ISBN/3-929711-06-0/ (zuletzt besucht am 14.10.2005)

ISC (2005): ISC Domain Survey: Number of Internet Hosts. http://www.isc.org/index.pl?/ops/ds/host-count-history.php (zuletzt besucht am 14.10.2005)

ISO/IEC 9126 (1991): Information technology – Software product evaluation – Quality characteristics and guidelines for their use

ISO/TS 16071 (2003): Ergonomics of human-system interaction – Guidance on accessibility for human-computer interfaces

Jorgenson, D.W., Stiroh, K.J. (2000): Raising the speed limit: US economic growth in the information age. http://www.olis.oecd.org/olis/2000doc.nsf/linkto/eco-wkp(2000)34 (zuletzt besucht am 23.09.2005)

Jovanovic, R., Rousseau, P.L. (2003): General purpose technologies. http://www.econ.nyu.edu/user/jovanovi/GPT.pdf (zuletzt besucht am 23.09.2005)

König, W., Heinzl, A. (2002): Die Wirtschaftsinformatik als Eckwissenschaft der Informationsgesellschaft. In: Wirtschaftsinformatik, 44, S.508-511

Kruschwitz, L. (2005): Investitionsrechnung, 10. Auflage, Oldenbourg, München

Kühlmann, T. (1993): Stressbewältigung bei computerunterstützter Arbeit: Ein prozessorientierter Ansatz. In: Zeitschrift für Arbeitswissenschaft, 47, S. 233-238

Lambert (1984): A comparative study of system response time on program developer productivity. In: IBM Systems journal, 23 (1), S. 36-43

Lamport, L. (1978): Time, clocks and the ordering of events in a distributed system. In: Communication of the ACM, 21, S. 558-565

Lenk, M. (2005): Durchsatzmaximierung von Wireless-LAN-Netzen: Parameteranalyse des Standards IEEE 802.11b. In: Müller, P., Gotzhein, R., Schmitt, J.B. (Hrsg.): Kommunikation in verteilten Systemen (KIVS): 14. Fachtagung Kommunikation in Verteilten Systemen (KIVS 2005), Springer, Berlin, S. 309-321

Licklider, J.C.R. (1960): Man-Computer Symbiosis. In: IRE Transactions on Human Factors in Electronics HFE-1, S. 4-11

Maass, S. (1993): Software-Ergonomie. Benutzer- und aufgabenorientierte Systemgestaltung. In: Informatik-Spektrum, 16, S. 191-205

Maccabee, M. (1996): Client/Server end-to-end response time: real life experience, IBM Research report RC 20483

Marcus, A. (2005): User Interface Design's Return on Investment: Examples and Statistics. In: Bias, R.G., Mayhew, D.J. (Eds.): Cost-Justifying Usability. An Update for the Internet Age, Morgan Kaufmann, San Francisco, S. 17-40

Martin, J. (1973): Design of man-computer dialogues, Prentice-Hall, Englewood

Matis, H. (2002): Die Wundermaschine, mitp, Bonn

Meinel, Ch., Sack, H. (2004): WWW: Kommunikation, Internetworking, Web-Technologien, Springer, Berlin

Melero, J., Toskala, A., Hakalin, P., Tolli, A. (2003): IMT-2000 3G Radio Access Technologies. In: Halonen, T., Romero, J, Melero, J. (Eds.): GSM, GPRS and EDGE performance, 2nd edition, Wiley, Chichester, S. 515-541

Menascé, D.A., Almeida, V.A.F. (2002): Capacity Planning for web services. Metrics, Models, and Methods, Prentice Hall, Upper Saddle River

Menascé, D.A., Almeida, V.A.F., Dowdy, L.W. (2004): Performance by design. Computer capacity planning by example, Prentice Hall, Upper Saddle River

Mertens, P. (1995): Wirtschaftsinformatik – Von der Mode zum Trend. In: König, W. (Hrsg.): Wirtschaftsinformatik ’95: Wettbewerbsfähigkeit, Innovation, Wirtschaftlichkeit, Physica, Heidelberg, S. 25-64

Mertens, P., Bodendorf, F., König, W., Picot, A., Schumann, M. & Hess, T. (2005): Grundzüge der Wirtschaftsinformatik, 9. Auflage, Springer, Berlin

Metcalf, R. (2004): Und wir brauchen sie doch. In: Technologie review, 7, S. 100-102

Meyer, J., Shinar, D., Bitan, Y., Leiser, D. (1996): Duration estimates and users’ preferences in human-computer interaction. In: Ergonomics, 39 (1), S. 46-60

Meyer; H.A., Hänze, M., Hildebrandt, M. (1999): Das Zusammenwirken von Systemresponsezeiten und Verweilzeiten beim Explorieren von Hypertextstrukturen. In: Wachsmuth, I., Jung, B. (Hrsg.): KogWis99: Proceedings der 4. Fachtagung der Gesellschaft für Kognitionswissenschaft, Infix, St. Augustin, S. 86-91

Meyer, H.A., Vogt, P. & Glier, M. (2005a): Performance – (k)ein Thema für Usability Professionals? In: Hassenzahl, M., Peissner M. (Hrsg.): Usability Professionals 2005 (Supplement, S. I-V), German Chapter der Usability Professionals Association, Stuttgart

Meyer, H.A., Vogt, P. & Glier, M. (2005b): Performance und Usability [im Druck] In: i-com, Zeitschrift für interaktive und kooperative Medien

Miller, G.A. (1956). The magic number seven plus or minus two: Some limits on our capacity for processing information. In: Psychological Review, 63, S. 81-97

Miller, R.B. (1968): Response time in man-computer conversational transactions. In: Proceedings of the Spring Joint Computer Conference, 33, S. 267-277

Mißbach, M.; Gibbels, P., Stelzel, J., Wagenblast, T. (2005): Adaptive Hardware-Infrastrukturen für SAP, Galileo Press, Bonn

Mons, W. (2000): Konrad Zuse – Persönlichkeit und Werdegang. In: Alex, J., Flessner, H., Mons, W., Pauli, K., Zuse, H. (Hrsg.): Konrad Zuse – Der Vater des Computers, Verlag Parzeller, Fulda, S. 15-60

Moore, G.E. (1965): Cramming more components onto integrated circuits. In: Electronics, 38 (8), 114-117

Mühlhäuser, M. (2002): Multimedia. In: Rechenberg, P., Pomberg, G. (Hrsg.): Informatik-Handbuch, 3. Auflage, Hanser, München

Müller-Merbach, H. (2002): Die Brückenaufgabe der Wirtschaftsinformatik. In: Wirtschaftsinformatik, 44, S. 300-301

Münzel, K. (1993): Depression und Erleben von Dauer. Zeitpsychologische Grundlagen und Ergebnisse klinischer Studien, Springer, Berlin

Nachreiner, F., Meyer, I., Schomann, C. und Hildebrandt, M. (1998): Überprüfung der Umsetzbarkeit der Empfehlungen der ISO 10075-2 in ein Beurteilungsverfahren zur Erfassung der psychischen Belastung, Wirtschaftsverlag NW, Bremerhaven

Nah, F. (2004): A study on tolerable waiting time: how long are Web users willing to wait? In: Behaviour & Information Technology, 23 (3)
http://sigs.aisnet.org/SIGHCI/bit04/BIT_Nah.pdf
(zuletzt besucht am 25.10.2005)

Nelson, R., Tantawi, A.N. (1989): Comparison of task response times in parallel systems, IBM Research report RC 15282

Nielsen, J. (1993): Usability Engineering, Academic Press, Cambridge

Nielsen, J. (2001): Designing Web Usability, New Riders, Indianapolis

Oberquelle, H. (1991): MCI – quo vadis? Perspektiven für die Gestaltung und Entwicklung der Mensch-Computer-Interaktion. In: Ackermann, D.; Ulich, E. (Hrsg.): Software-Ergonomie `91. Benutzerorientierte Software-Entwicklung, Teubner, Stuttgart, S. 9-24

Oberquelle, H. (2000): Kosten der (Un-)Benutzbarkeit – (k)ein Thema für die Wirtschaftsinformatik? In: HMD – Praxis der Wirtschaftsinformatik, 212, S. 4-6

Pearrow, M. (2002): The wireless web usability handbook, Charles River, Hingham

Peterson, L.R., Peterson, M.J. (1959): Short-term retention of individual verbal items. In: Journal of Experimental Psychology, 58, S. 193-198

Pfaff, H. und Mitarbeiter (2004): Lebenslagen der behinderten Menschen. Ergebnis des Mikrozensus 2003. In: Statistisches Bundesamt: Wirtschaft und Statistik, SFG Servicecenter Fachverlage, Berlin, S. 1181-1194

Pfleeger, C.P., Pfleeger, S.L. (2003): Security in computing, 3. Auflage, Prentice Hall, Upper Saddle River

Potthoff, I. (1998): Empirische Studien zum wirtschaftlichen Erfolg der Informationsverarbeitung. In: Wirtschaftsinformatik, 40, S. 54-65

Raskin, J. (2000): The human interface. New directions for designing interactive systems, Addison-Wesley, Reading

RFC 2616 (1999): Hypertext Transfer Protocol – http/ 1.1. http://www.ietf.org/rfc/rfc2616.txt (zuletzt besucht am 14.11.2005)

Richter, G. (2000): Psychische Belastung und Beanspruchung – Stress, psychische Ermüdung, Monotonie, psychische Sättigung, Wirtschaftsverlag NW, Bremerhaven

Riemann, W.O. (2001): Wirtschaftsinformatik, 3. Auflage, Oldenbourg, München

Rolf, A. (1998): Herausforderungen für die Wirtschaftsinformatik. In: Informatik-Spektrum, 21, S. 259-264

Rolf, A. (2004): Informatiksysteme in Organisation und Gesellschaft. Teil A: „Informatiksysteme in Organisationen und globaler Ökonomie – Ein Orientierungsrahmen“, Universität Hamburg, Fachbereich Informatik, Mitteilung 330

Rosenberg, D. (2004): The myths of usability ROI. In: Interactions, 11 (5), S. 22-29

Schneider, W. (1998a): Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten – Grundsätze der Dialoggestaltung. Kommentar zu DIN EN ISO 9241-10, Beuth, Berlin

Schneider, F. B. (1998b): Toward Trustworthy Networked Information Systems. In: Communications of the ACM, 40 (11), S. 144

Schwartz, M. (2001): Client/Server-Architektur. In: Mertens, P.: Lexikon der Wirtschaftsinformatik, Springer, Berlin, S. 96-97

Schwarze, J. (2000): Einführung in die Wirtschaftsinformatik, 5. Auflage. Verlag Neue Wirtschafts-Briefe, Herne/Berlin

Shneiderman, B. (1984): Response Time und Display Rate in Human Performance with Computers. In: Computing Surveys, 16 (3), S. 265-285

Shneiderman, B. (1998): Designing the user interface: strategies for effective human-computer interaction, 3rd edition, Addison-Wesley, Reading

Shneiderman, B., Plaisant, C. (2005): Designing the user interface: strategies for effective human-computer interaction, 4. Auflage, Addison-Wesley, Boston

Smith, C. U. (1993): Software Performance Engineering. In: Donatiello, L., Nelson, R. (Eds.): Performance Evaluation of Computer and Communication Systems, Springer, Berlin, S. 509-536

Smith, C.U., Williams, L.G. (2002): Performance Solutions. A practical guide to creating responsive, scalable software, Addison-Wesley, Boston

Städler, T. (2003): Lexikon der Psychologie, Alfred Kröner Verlag, Stuttgart

Stahlknecht, P., Hasenkamp, U. (2005): Einführung in die Wirtschaftsinformatik, 11. Auflage, Springer, Berlin

Stein, E. (2004): Taschenbuch Rechnernetze und Internet, 2. Auflage, Fachbuchverlag Leipzig, München

Strunz, H. (1998): Investition. In: Krabbe. E. (Hrsg.): Leitfaden zum Grundstudium der Betriebswirtschaftslehre, 6. Auflage, dbv, Gernsbach, S. 287-362

Thadhani, A.J. (1981): Interactive user productivity. In: IBM Systems Journal, 20 (4), S. 407-423

Tanenbaum, A., van Stehen, M. (2003): Verteilte Systeme. Grundlagen und Paradigmen, Pearson Studium, München

Titchkosky, L., Arlitt, M., Williamson, C. (2003): A performance comparison of dynamic web technologies. In: ACM Sigmetrics, 31 (3), S. 2-11

Unland, R. (2001): Timesharing. In: Mertens, P.: Lexikon der Wirtschaftsinformatik, Springer, Berlin, S. 475

Utton, P., Hill, B. (1997): Performance Prediction: an Industry Perspective. In: Raymond, M. (Ed.): Computer performance evaluation, proceedings 9th international conference, Springer, Berlin, S. 1-5

Weist, D. (2004): Accessibility – Barrierefreies Internet, VDM, Berlin

Weizenbaum, J. (1984): Kurs auf den Eisberg, Pendo-Verlag, Zürich

Weizenbaum, J. (1993): Wer erfindet die Computermythen? Der Fortschritt in einem großen Irrtum, Herder, Freiburg im Breisgau

Williams, L.G., Smith, C.U. (o.J.): Coping with Anxiety. http://www.perfeng.com/perfanx.htm (zuletzt besucht am 25.10.2005)

Wilson, C.E., Rosenbaum, S. (2005): Categories of Return on Investment and their practical implications. In: Bias, R.G., Mayhew, D.J. (Eds.): Cost-Justifying Usability. An Update for the Internet Age, Morgan Kaufmann, San Francisco, S. 215-264

Winzerling, W. (2001): E-Business-Technik. Grundlagen, Anwendungen, Perspektiven, VDE Verlag, Berlin

WKWI (1994): Wissenschaftliche Kommission Wirtschaftsinformatik im Verband der Hochschullehrer für Betriebswirtschaftslehre. In: Wirtschaftsinformatik, 36, S. 80-81

Wöhe, G. (1996): Einführung in die Allgemeine Betriebswirtschaftslehre, 19. Auflage, Vahlen, München

Woodside, C.M. (1993): Performance Engineering of Client-Server Systems. In: Donatiello, L., Nelson, R. (Eds.) Performance Evaluation of Computer and Communication Systems. Joint Tutorial Papers of Performance ’93 and Sigmetric ’93, Springer, Berlin

Yuan, J-L, Chi, C-H. (2003): Web Caching Performance: How much is lost unwarily? In: Chung, C.W. et al. (Eds.): Web and Communication technologies and Internet-Related Social issues – HIS 2003, S. 23-33

Zapf, D.; Frese, M. (1993): Stress. In: Schorr, A.: Handwörterbuch der angewandten Psychologie, Deutscher Psychologen Verlag, Bonn, S. 658–660

Zarnekow, R., Scheeg, J., Brenner, W. (2004): Untersuchung der Lebenszykluskosten von IT-Anwendungen. In: Wirtschaftsinformatik, 46, S. 181-187

Zeidler, A., Zellner, R. (1994): Software-Ergonomie. Techniken der Dialoggestaltung, 2. Auflage, Oldenbourg, München

Zhi, J. (2001): Web page design and download time. In: CMG Journal of Computer Management, 102, S. 40-55

Zimbardo, P.G. (1995): Psychologie, 6. Auflage, Springer, Berlin

Züllighoven, H. (1998): Das objektorientierte Konstruktionshandbuch nach dem Werkzeug & Material-Ansatz, dpunkt, Heidelberg

Züllighoven, H. (2005): Object-oriented construction handbook: developing application-oriented software with tools and materials approach, dpunkt, Amsterdam

Erklärung

Ich versichere, dass ich die vorstehende Arbeit selbstständig und ohne fremde Hilfe angefertigt und mich anderer als der im beigefügten Verzeichnis angegebenen Hilfsmittel nicht bedient habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Alle Quellen, die dem World Wide Web entnommen oder in einer sonstigen digitalen Form verwendet wurden, sind der Arbeit beigefügt.

Hamburg, im November 2005 _____

Marco Glier