Das Rechenzentrum als Produktionsstätte für IT-Dienstleistungen -
Kapazitätswirtschaft in virtualisierten Rechenzentren

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von: Dipl.-Wirtsch.-Inf. Alexander Pinnow

geb. am 11.08.1973 in Staßfurt

Gutachter:

Prof. Dr. Georg Paul

Prof. Dr. Hans-Knud Arndt

Prof. Dr. Klaus Turowski

Magdeburg, den 22.12.2009

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Der Betrieb von Rechenzentren wird heutzutage oft ereignisgetrieben durchgeführt. Als Reaktion auf neue Kundenbedarfe werden weitere Hardwaresysteme in das Rechenzentrum integriert. Auch wenn die Beschaffungskosten für zusätzliche Hardware von den Betreibern der Rechenzentren im Vergleich zu den Personalkosten meist als unkritisch betrachtet werden, führt der Ausbau der Informationsinfrastruktur zu einem erhöhten Administrations-, Wartungs- und damit letztlich auch Personalaufwand.

1.1 Motivation

Für die Ablösung der ereignisgetriebenen Betriebsstrukturen erscheint die Adaption geeigneter Verfahren der operativen Produktionsplanung und -steuerung für die Entwicklung effizienterer Betriebskonzepte sinnvoll. In der Praxis hat der Prozess der Bereitstellung und des Betriebs der Infrastruktur oft den Charakter eines IT-Projekts. Dieser Prozess soll durch erprobte Verfahren der Produktionsplanung und -steuerung standardisiert werden. Zielsetzung ist die Erbringung vorgegebener IT-Dienstleistungen mit möglichst geringen Kosten. Die Umsetzung dieses Wirtschaftlichkeitsprinzips soll durch die Minimierung von Wartezeiten, Stillstandszeiten und Terminüberschreitungen sowie die Maximierung der Kapazitätsauslastung aller Betriebsmittel erreicht werden (vgl. Kurbel, 2005, S. 9 f.).

In der vorliegenden Arbeit wird untersucht, ob und wie betriebswirtschaftliche Methoden auf den Betrieb eines Rechenzentrums als IT-Dienstleister angewendet werden können. Der Schwerpunkt der Arbeit liegt dabei auf der Betrachtung der Kapazitätswirtschaft. Explizit nicht betrachtet werden hierbei Fragen zu den Themen Virtualisierungstechniken, Instandhaltung, und Störungsmanagement. Zu diesen Themen entstehen im Rahmen des Forschungsschwerpunktes weitere Dissertationen.

1.2 Forschungsmethodik

Die Arbeit wird in der Forschungsdisziplin Wirtschaftsinformatik erstellt. Diese versteht sich als eigenständige Subdisziplin der Wirtschaftswissenschaften und der Informatik (vgl. Winter, 2009, S. 231). Forschungsgegenstand der Wirtschaftsinformatik sind Informations- und Kommunikationssysteme in Wirtschaft und Verwaltung (vgl. Alpar, 2000, S. 4). Die in der Arbeit betrachteten Forschungsgegenstände sind somit zum einen IT-Dienstleistungen sowie die Systeme zur Produktionsplanung und -steuerung, in denen die Produktion dieser IT-Dienstleistungen abgebildet werden soll.
Als Forschungsmethodik wird eine ingenieurwissenschaftliche Vorgehensweise gewählt. Diese zeichnet sich durch die Entwicklung von Konzepten und Modellen sowie den Bau von Prototypen aus. Die durch den Bau eines Prototypen neu geschaffene oder veränderte Wirklichkeit bedarf selbst wieder der wissenschaftlichen Untersuchung, um das Wissen über den Forschungsgegenstand zu erweitern (vgl. Heinrich, 2005, S. 107).
Wirtschaftsinformatik ist als Realwissenschaft eine praxisnahe Forschungsdisziplin (vgl. Heinrich, 2005, S. 107). Um diese Praxisnähe zu gewährleisten, wird im Rahmen der Arbeit nicht aus den Anforderungen der Produktion von IT-Dienstleistungen ein neues System zur Produktionsplanung und -steuerung entwickelt. Vielmehr wird versucht, die Anforderungen an die Produktion von IT-Dienstleistungen mit den in realen Systemen existierenden Datenstrukturen und Methoden abzubilden. Die Datenstrukturen

und Methoden etablierter Systeme zur Produktionsplanung und -steuerung werden in dem Standardwerk von Karl Kurbel "Produktionsplanung und -steuerung im Enterprise Resource Planning und Supply Chain Management" (Kurbel, 2005) praxisnah beschrieben. Eine formalere Darstellung dieser Thematik wird in dem Standardwerk "Produktions-Management" (Adam, 1998) von Dietrich Adam gegeben. Diese Werke dienen der Arbeit im Wesentlichen als Grundlage zur Beschreibung der Systeme zur Produktionsplanung und -steuerung als Forschungsgegenstand.
Auf den aktuellen Stand der Forschung im Bereich der Wirtschaftswissenschaften wird dabei bewusst kein Bezug genommen. Als Begründung hierfür sei an dieser Stelle Karl Kurbel zitiert (vgl. Kurbel, 2005, S. 40):
"Zusammenfassend muss man kritisch feststellen, dass die Betriebswirtschaftslehre und insbesondere die Unternehmensforschung ... keinen nennenswerten Beitrag zur Lösung der praktischen Probleme der Produktionsplanung geleistet haben."
Im Rahmen der Arbeit wird die Produktion von IT-Dienstleistungen durch Datenstrukturen und Methoden der Produktionsplanung und -steuerung abgebildet. Als Forschungsergebnis liegt im ersten Schritt ein Referenzmodell zur Umsetzung dieser Anforderungen vor. Bei der Referenzmodellierung wird induktiv und deduktiv eine vereinfachte und optimierte Abbildung eines Systems entwickelt (vgl. Wilde und Hess, 2007, S. 282). Die gewonnenen Erkenntnisse werden als Grundlage für weitere Modellierungsaktivitäten dokumentiert (vgl. Schneider, 1998, S. 714).
Im nächsten Schritt wird das vorliegende Modell anhand eines Prototypen evaluiert. Ein Prototyp ist eine frühe ausführbare Version eines späteren Produkts, die bereits alle relevanten grundlegenden Merkmale aufweist (vgl. Alpar, 2000, S. 218). Die Methodik des Prototyping beschreibt die Entwicklung und Evaluation einer Vorabversion eines Anwendungssystems (vgl. Wilde und Hess, 2007, S. 282). Im Rahmen der Arbeit wird das Prototyping auf die Anpassung eines Informationssystems an die spezifischen Anforderungen eines Unternehmens beschränkt. Diese Aufgabe wird als Customizing bezeichnet (vgl. Kurbel, 2005, S. 414). Zur Erweiterung des Wissens über den Forschungsgegenstand IT-Dienstleistung werden anhand

des Prototypen weitere Forschungsfragen generiert. Die Beantwortung dieser Forschungsfragen liefert als Forschungsergebnis jeweils weitere Modelle der betrachteten Forschungsgegenstände. Diese Modelle werden auf Grundlage von Erkenntnissen der Mathematik als Forschungsdisziplin erstellt.
Forschungsziel der Arbeit ist es, das Einsatzgebiet von Systemen zur Produktionsplanung und -steuerung zu erweitern. Es soll aufgezeigt werden, dass es möglich ist, diese Systeme auch zur Abbildung der Produktion von IT-Dienstleistungen in Rechenzentren einzusetzen. Der wissenschaftliche Mehrwert im Sinne der Wirtschaftsinformatik liegt somit in der Erschließung eines neuen Anwendungsgebiets der Wirtschaft für ein existierendes Informations- und Kommunikationssystem. Dies geschieht insbesondere vor dem Hintergrund, dass in Rechenzentren heute bereits Systeme für das Enterprise Resource Planning (ERP) eingesetzt werden. ERP-Systeme dienen der vollständigen und durchgehenden Abbildung aller betriebswirtschaftlichen Standardprozesse eines Unternehmens (vgl. Lassmann, 2006, S. 489). Dieser Anforderung werden die im Rechenzentrum eingesetzten ERP-Systeme jedoch nicht gerecht, da meist die Komponenten zur Abbildung der Prozesse für Einkauf, Vertrieb, Finanzbuchhaltung, Controlling und Personalwirtschaft eingesetzt werden, eine Abbildung der Produktion der IT-Dienstleistungen bisher jedoch nicht erfolgt.

1.3 Aufbau der Arbeit

Im folgenden Kapitel wird das Rechenzentrum als IT-Dienstleister vorgestellt und die IT Infrastructure Library (ITIL) als mögliche Umsetzung eines IT-Service-Managements beschrieben. In Kapitel 3 wird das Rechenzentrum als Produktionsbetrieb für IT-Dienstleistungen dargestellt. Als Forschungsmethode wird die Referenzmodellierung eingesetzt. Es wird ein Referenzmodell der Datenstrukturen und Methoden für die Bereitstellung und den Betrieb der Infrastruktur eines Rechenzentrums durch Methoden der Produktionsplanung und -steuerung erstellt. Das beschriebene Referenzmodell wird durch

ein Prototyping in einem SAP®ERP evaluiert. Anhand des Referenzmodells und des Prototypen werden Probleme identifiziert, die sich nicht durch standardisierte Methoden der Produktionsplanung und -steuerung lösen lassen. Diese Probleme werden in den folgenden Kapiteln behandelt. In Kapitel 4 wird untersucht, welchen Einfluss der Einsatz von Virtualisierungstechniken auf die Planung des Kapazitätsangebots hat. Hierzu wird ein mathematisch-formales Modell (vgl. Wilde und Hess, 2007, S. 282) entwickelt, welches die Inanspruchnahme des Kapazitätsangebots beschreibt. Das Modell wird auf der Grundlage von Markovketten und der Weibullverteilung erstellt. In Kapitel 5 wird ein mathematisch-formales Modell für die Entwicklung des Kapazitätsangebots der Infrastrukturkomponenten eines Rechenzentrums entworfen. Auf Grundlage dieses Modells werden standardisierte Prognoseverfahren angewendet, um die Entwicklung des Kapazitätsangebots in zukünftigen Perioden ermitteln zu können. Kapitel 6 beschäftigt sich mit Fragen der Betriebsmittelzuordnung. Die Frage der Zuordnung von virtuellen zu physischen Betriebsmitteln wird als Bin-Packing-Problem dargestellt. Ansätze zur Lösung des Bin-Packing-Problems werden als Methoden der Betriebsmittelzuordnung in einem Referenzmodell beschrieben. Abschließend wird die Zuordnung virtualisierter Betriebsmittel zu Fertigungsaufträgen dargestellt. Hierzu wird auf Erkenntnisse aus dem mathematisch-formalen Modell zur Planung des Kapazitätsangebots zurückgegriffen.

Kapitel 2

Betrachtung des Rechenzentrums als IT-Dienstleister

In diesem Kapitel wird das Rechenzentrum als IT-Dienstleister dargestellt. Als Grundlage zur Umsetzung der Methoden zur Produktionsplanung und -steuerung wird der Forschungsgegenstand IT-Dienstleistung beschrieben. Weiterer Forschungsgegenstand der Arbeit sind Systeme zur Produktionsplanung und -steuerung. Hierbei handelt es sich um offene Systeme, die im Informationsaustausch mit ihrer Umwelt stehen (vgl. von Bertalanffy, 1980, S. 141). In Abschnitt 2.2 werden Systeme zur Produktionsplanung und -steuerung als betrachteter Forschungsgegenstand zu ihrer Umwelt im Rechenzentrum in Beziehung gesetzt.

2.1 IT-Dienstleistungen

Dienstleistungen sind immaterielle Wirtschaftsgüter, die unter Einsatz externer Produktionsfaktoren für den Bedarf Dritter produziert werden (vgl. Frietzsche und Maleri, 2006, S. 197). IT-Dienstleistungen werden von einem IT-Dienstleistungsanbieter für einen oder mehrere Kunden erbracht.

IT-Dienstleistungen unterstützen die Geschäftsprozesse des Kunden und basieren auf dem Einsatz von Informationstechnologie. Bei der Erbringung von IT-Dienstleistungen werden Personen und Technologien für die Ausführung von Prozessen eingesetzt (vgl. Arbeitskreis Publikation ITIL Version 3 Translation Project, 2007, S. 26).
Der Umfang einer IT-Dienstleistung wird in einer Service-Level-Vereinbarung (engl. Service Level Agreement, SLA) definiert. Service-Level-Vereinbarungen werden zwischen dem Anbieter der IT-Dienstleistung und dem Kunden getroffen. Die Vereinbarungen können für mehrere IT-Dienstleistungen oder mehrere Kunden gelten. Sie beschreiben die IT-Dienstleistung, definieren die Service Level und legen die Verantwortlichkeiten des Anbieters der IT-Dienstleistung und des Kunden fest. Ein Service Level enthält messbare und nachweisbare Ergebnisse, die im Rahmen der Erbringung einer oder mehrerer IT-Dienstleistungen erzielt werden sollen (vgl. Arbeitskreis Publikation ITIL Version 3 Translation Project, 2007, S. 44).
IT-Dienstleistungen werden häufig zu einem Dienstleistungsbündel zusammengefasst (vgl. Corsten und Gössinger, 2007, S. 30). Ein solches Dienstleistungsbündel wird dem Kunden als IT-Produkt angeboten. IT-Dienstleistungen lassen sich unterteilen in (vgl. Zarnekow, 2007, S. 11):

- Infrastrukturdienstleistungen,
- IT-Arbeitsplatzdienstleistungen,
- Geschäftsprozessdienstleistungen und
- Unterstützungs- und Wartungsdienstleistungen.

Das Rechenzentrum stellt dem Kunden IT-Infrastruktur zur Verfügung und erbringt somit eine Infrastrukturdienstleistung. Die Identifikation des Reifegrads eines IT-Dienstleisters kann anhand des fünfstufigen Service Provider Maturity Model vorgenommen werden. Der Fokus des IT-Dienstleisters kann

durch die Reifegradstufe bestimmt werden (vgl. Grawe und Fähnrich, 2008, S. 282 f.):

- **Stufe 1 Infrastruktur:** Der Dienstleistungsanbieter konzentriert sich ausschließlich auf den Betrieb der Infrastruktur. Es existieren keine klar definierten IT-Dienstleistungen oder IT-Produkte. Der Betrieb der Infrastruktur wird ereignisgetrieben durchgeführt.

- **Stufe 2 Prozesse:** Hauptaufgabe ist weiterhin der Betrieb der Infrastruktur. Der Betrieb wird jedoch auf Grundlage definierter Prozesse gesteuert und nach festen Vorgaben durchgeführt.

- **Stufe 3 Anwender:** Es existieren definierte IT-Dienstleistungen. Der Umfang der IT-Dienstleistung wird in Service-Level-Vereinbarungen definiert. Der Anbieter versucht die Service-Level-Vereinbarungen an die Anforderungen des Kunden anzupassen.

- **Stufe 4 Produkte:** Der Anbieter entwickelt ein Portfolio von IT-Produkten mit vordefinierten IT-Dienstleistungen. Die Erbringungsprozesse der IT-Dienstleistungen werden für definierte IT-Produkte optimiert. IT-Dienstleistungen und damit auch IT-Produkte lassen sich modularisieren (vgl. Böhmann und Krcmar, 2006, S. 45 ff.). Auf diese Weise können dem Kunden an seine Anforderungen angepasste IT-Produkte aus standardisierten Modulen von IT-Dienstleistungen angeboten werden.

- **Stufe 5 Markt:** Der IT-Dienstleister positioniert sich am Markt oder tritt als interner IT-Dienstleister in Wettbewerb mit externen Anbietern.

Die Hauptaufgabe des Rechenzentrums bleibt unabhängig von der Reifegradstufe der Betrieb der Infrastruktur. Mit steigender Reifegradstufe steigen jedoch die Anforderungen an das Rechenzentrum als IT-Dienstleister. Bereits auf der Prozessstufe werden die Abläufe im Rechenzentrum klar defi-

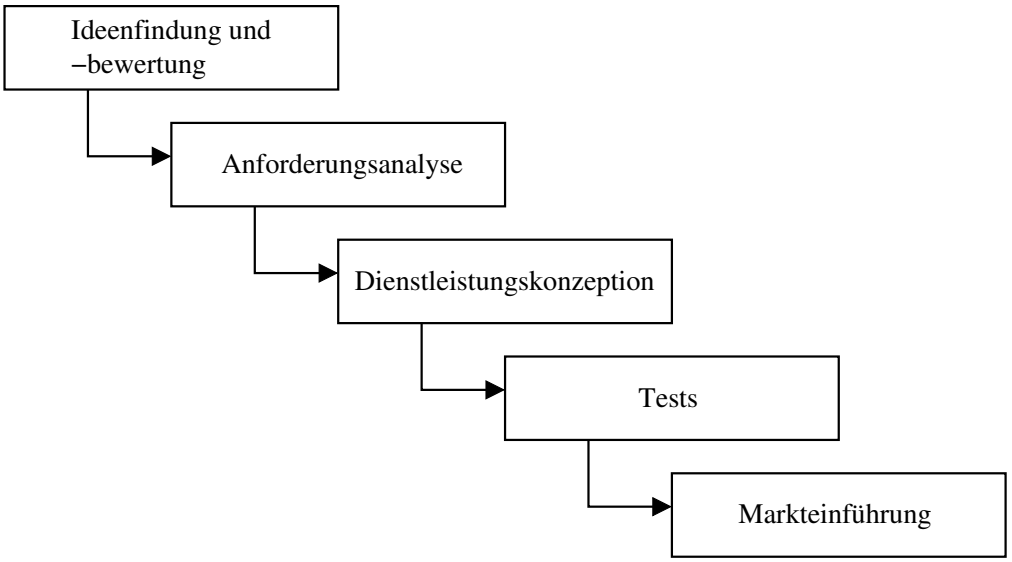

Abbildung 2.1: Fünf-Phasen-Modell der Dienstleistungsentwicklung (nach van Husen et al., 2008, S. 36)

niert. Die Strukturen dieser Abläufe können in Arbeitsplänen erfasst werden. Der Aufbau solcher Arbeitspläne wird in Abschnitt 3.1.1 beschrieben. Für eine Bewertung der Einhaltung der Service-Level-Vereinbarungen auf der Anwenderstufe müssen Bewertungsmodelle entwickelt werden. Die Bewertung der Kapazitätswirtschaft wird in Abschnitt 3.3.2 vorgenommen. Ab der Produktstufe sind für IT-Dienstleistungen die Kapazitätsangebote der Infrastrukturkomponenten eines IT-Produkts festzulegen. Ein Modell zur Kapazitätsangebotsplanung wird in Abschnitt 4 vorgestellt. Weiterhin ist der erwartete Kapazitätsbedarf kommender Planungsperioden zu ermitteln. Die Prognose der Kapazitätsbedarfe wird in Abschnitt 5 beschrieben. Um auf der Marktstufe konkurrenzfähig gegenüber anderen Marktanbietern sein zu können, ist es notwendig, kosteneffizient zu arbeiten. Eine Voraussetzung hierfür ist die effiziente Ausnutzung der Kapazitäten der Infrastrukturkomponenten. Modelle für eine effiziente Betriebsmittelzuordnung werden in Abschnitt 6 vorgestellt.

Die Entwicklung einer neuen IT-Dienstleistung lässt sich in mehrere Pha-

sen unterteilen (siehe Abbildung 2.1). Es wird folgendes Fünf-Phasen-Modell durchlaufen (vgl. van Husen et al., 2008, S. 36):

- **Phase 1: Ideenfindung und -bewertung**
  Ideen werden häufig durch Mitarbeiter formuliert, die in Kontakt zu Kunden stehen. Geeignete Methoden für die Ideenfindung und -bewertung sind Innovationsworkshops und Marktuntersuchungen.
- **Phase 2: Anforderungsanalyse**
  Im Rahmen der Anforderungsanalyse wird eine Klassifizierung und Priorisierung der Anforderungen an die IT-Dienstleistung vorgenommen. Strategische, funktionale, organisatorische und Marktanforderungen sind mögliche Klassifizierungen. Die Anforderungen lassen sich weiterhin in fachliche und technische Anforderungen unterteilen.
- **Phase 3: Dienstleistungskonzeption**
  Bei der Konzeption von IT-Dienstleistungen werden Produktmodelle, Prozessmodelle, Ressourcenkonzepte und Marketingkonzepte eingesetzt.
- **Phase 4: Tests**
  Die Marktreife einer IT-Dienstleistung wird durch Akzeptanz- und Benutzertests bestimmt. Tests werden häufig durch Einbindung eines Pilotkunden durchgeführt.
- **Phase 5: Markteinführung**
  Die Einführung einer neuen IT-Dienstleistung wird durch interne und externe Marketingmaßnahmen begleitet. Es werden Mitarbeiterschulungen durchgeführt, um das Personal mit den neuen IT-Dienstleistungen vertraut zu machen.

Weiterhin ist bei der Einführung von IT-Dienstleistungen im Rahmen der Standortplanung eine geeignete Standortwahl vorzunehmen. Hierbei besteht ein Zielkonflikt zwischen Kosteneffizienz und Kundennähe (vgl. Miklitz et al., 2006, S. 397). Der Anbieter der IT-Dienstleistung ist bestrebt, diese so kostengünstig wie möglich anzubieten und versucht gleichzeitig, eine höhere Akzeptanz durch Kundennähe zu erreichen.

2.2 ITIL

Die IT Infrastructure Library (ITIL) ist ein Best-Practise-Ansatz zur Etablierung eines IT-Service-Managements. Aufgabe des IT-Service-Managements ist die Bereitstellung und Erbringung von IT-Dienstleistungen zur Unterstützung der Geschäftsprozesse einer Betriebswirtschaft (vgl. Buchsein et al., 2008, S. 5). Durch den Einsatz von ITIL sollen folgende Ziele erreicht werden (vgl. Olbrich, 2008, S. 4):

- effektive und zielorientierte Gestaltung von Prozessen, Aufgaben und Rollen,
- größere Flexibilität und Handlungsfreiheit bei veränderten Marktsituationen,
- bessere Umsetzung neuer Anforderungen an eine Betriebswirtschaft,
- Senkung der Kosten zur Erbringung einer IT-Dienstleistung,
- Verbesserung der internen und externen Kommunikation,
- Verbesserung der Kunden- und Mitarbeiterzufriedenheit durch Schaffung transparenter Arbeitsabläufe,
- Verbesserung der Qualität von IT-Dienstleistungen und
- Bildung einer einheitlichen Begriffsverwendung zur Vermeidung von Missverständnissen.

In der Version 3 besteht ITIL aus fünf Phasen eines iterativen mehrdimensionalen Lebenszyklus (vgl. Buchsein et al., 2008, S. 15). Es werden die Phasen Servicestrategie (engl. Service Strategy), Service Design, Serviceüberführung (engl. Service Transition), Servicebetrieb (engl. Service Operation) und Kontinuierliche Serviceverbesserung (engl. Continual Service Improvement) durchlaufen (vgl. OGC, 2007b, S. 6). Abbildung 2.2 veranschaulicht die Beziehung der Phasen zueinander.

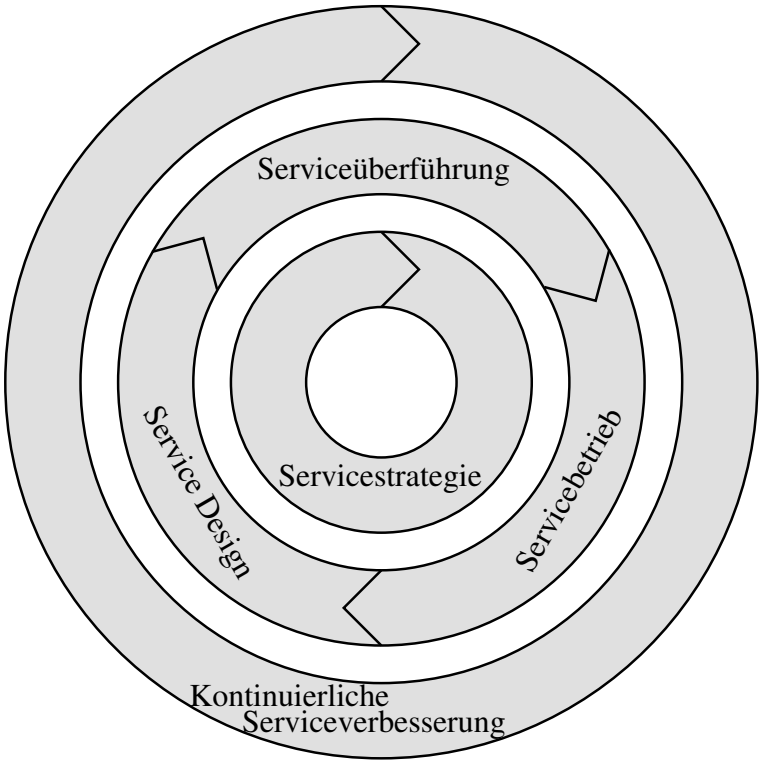

Abbildung 2.2: ITIL Service Lebenszyklus (nach OGC, 2007b, S. 19)

2.2.1 Servicestrategie

Im Mittelpunkt des Lebenszyklus steht die Phase Servicestrategie. Im Rahmen der Servicestrategie werden strategische Entscheidungen für das Design, die Entwicklung und die Implementierung des Service Managements getroffen (vgl. Buchsein et al., 2008, S. 15).
Aufgabe dieser Phase ist die Erschließung neuer Geschäftsmöglichkeiten. Hierzu ist es notwendig die Marktsituation zu erfassen, richtig einzuschätzen und die Bedürfnisse der Kunden zu verstehen. Auf dieser Grundlage können neue, auf die Kundenbedürfnisse angepasste Serviceangebote bereitgestellt werden. Das Serviceangebot wird in einem Serviceportfolio erfasst (vgl. Olbrich, 2008, S. 147). Ein Serviceportfolio enthält alle Dienstleistungen,

die von einem Leistungsanbieter verwaltet werden. Hierzu zählen (vgl. Arbeitskreis Publikation ITIL Version 3 Translation Project, 2007, S. 48):

- beantragte oder in der Entwicklung befindliche Dienstleistungen,
- der Servicekatalog mit allen Dienstleistungen, die sich in Produktion befinden oder bereit zur Produktion sind und
- alle außer Kraft gesetzten Dienstleistungen.

Im Rahmen der Servicestrategie werden kritische Erfolgsfaktoren identifiziert. Weiterhin wird die Wettbewerbssituation am Markt analysiert. Es werden Methoden, Modelle, Techniken, Analysen und Aktivitäten zur Budgetierung und Kostenrechnung entwickelt und angewendet (vgl. Olbrich, 2008, S. 147). Die Sericestrategie bildet die Grundlage der Phasen Service Design, Serviceüberführung und Servicebetrieb.

2.2.2 Service Design

Aufgabe des Service Design ist die Entwicklung innovativer IT-Dienstleistungen auf Grundlage der Servicestrategie. Im Rahmen des Service Designs wird der Servicekatalog erstellt. Es werden Serviceziele mit den Geschäftspartnern definiert und in Service Level Vereinbarungen festgehalten (vgl. Olbrich, 2008, S. 149 f.). Um die Service Level Vereinbarungen wirtschaftlich und zeitlich einhalten zu können, muss sichergestellt werden, dass ausreichende Kapazitäten der IT-Dienstleistungen und der IT-Infrastruktur bereitgestellt werden. (vgl. Arbeitskreis Publikation ITIL Version 3 Translation Project, 2007, S. 11). Die in dieser Arbeit betrachtete Kapazitätswirtschaft ist somit Teil der Phase Service Design des ITIL Lebenszyklus.
Im Rahmen des Service Design ist weiterhin sicherzustellen, dass die bereitgestellten Kapazitäten auch verfügbar sind und bei auftretenden Störungen innerhalb eines definierten Zeitintervalls wieder in Betrieb gesetzt werden

können (vgl. Olbrich, 2008, S. 150). Eine weitere Aufgabe ist die Sicherstellung der Vertraulichkeit, Integrität und Verfügbarkeit der Ressourcen, Informationen, Daten und IT-Dienstleistungen (vgl. Arbeitskreis Publikation ITIL Version 3 Translation Project, 2007, S. 23). Im Rahmen des Service Design ist weiterhin sicherzustellen, dass alle Zulieferer ihre vertraglichen Vereinbarungen erfüllen.

2.2.3 Serviceüberführung

Aufgabe der Serviceüberführung ist die Ausführung und Koordination von Prozessen und Funktionen zur Erstellung, Testung, Paketierung und Auslieferung neuer Versionen von IT-Dienstleistungen und technischer Komponenten in die Produktion. Die Serviceüberführung wird auf Grundlage der Servicestrategie und der Anforderungen des Service Design vorgenommen (vgl. Buchsein et al., 2008, S. 19 f.).
Die Umsetzung der Serviceüberführung erfordert eine strukturierte Planung. Bei weitreichenden Veränderungen ist eine projektübergreifende Koordination sicherzustellen (vgl. Olbrich, 2008, S. 152). Für die Ablösung alter Versionen von IT-Dienstleistungen und technischer Komponenten ist eine zeitlich minimale Unterbrechung anzustreben. Sämtliche Konfigurationselemente von IT-Dienstleistungen und technischen Komponenten werden in einem zentralen Datenbankmanagementsystem verwaltet. Dieses System wird als Configuration Management Database (CMDB) bezeichnet. Konfigurationselemente sind alle Komponenten, die eingesetzt werden, um eine IT-Dienstleistung bereitzustellen (vgl. Arbeitskreis Publikation ITIL Version 3 Translation Project, 2007, S. 14). Um gewährleisten zu können, dass neue oder geänderte Dienstleistungen die Service Level Vereinbarungen erfüllen, ist ein Qualitätssicherungsprozess zu etablieren. Leistungsmerkmale von IT-Dienstleistungen und technischen Komponenten sind kontinuierlich zu evaluieren, um Abweichungen von Vorgabewerten zu erkennen. Eine weitere Aufgabe der Serviceüberführung besteht darin, neu gewonnenes Wissen per-

sistent zu machen und sicherzustellen, dass dieses Wissen allen Interessenten zugänglich ist (vgl. Olbrich, 2008, S. 153).

2.2.4 Servicebetrieb

Aufgabe des Servicebetriebs ist die Sicherstellung der Erbringung der IT-Dienstleistungen. In dieser Phase erfolgt die eigentliche Realisierung der strategischen Ziele (vgl. Buchsein et al., 2008, S. 21).
Im Rahmen des Servicebetriebs sind alle IT-Dienstleistungen und technischen Komponenten zu überwachen. Beobachtete Ereignisse und Auffälligkeiten werden zur Steuerung des Regelbetriebs und zur Erkennung und Eskalation von Fehlern und Ausnahmesituationen herangezogen (vgl. Olbrich, 2008, S. 155).
Bei Ausfällen von IT-Dienstleistungen oder technischen Komponenten ist eine schnellstmögliche Wiederherstellung des Sollzustands zu gewährleisten (vgl. Arbeitskreis Publikation ITIL Version 3 Translation Project, 2007, S. 23). Die Ursachen auftretender Störungen sind zu bestimmen, um geeignete Lösungsmöglichkeiten zu finden und proaktiv auf zukünftige potentielle Störungen reagieren zu können.
Weiterhin sind im Rahmen des Servicebetriebs die Berechtigungen für den Zugriff auf eine IT-Dienstleistung zu regeln. Es wird sichergestellt, dass berechtigte Personen oder Gruppen die IT-Dienstleistung in Anspruch nehmen können (vgl. Olbrich, 2008, S. 155).

2.2.5 Kontinuierliche Serviceverbesserung

IT-Dienstleistungen müssen ständig an sich ändernde Anforderungen der Geschäftsprozesse angepasst werden. Aufgabe der kontinuierliche Serviceverbesserung ist eine stetige Optimierung aller Phasen des ITIL Lebenszyklus (vgl. Buchsein et al., 2008, S. 22). Hierzu sind sämtliche gewonnenen Infor-

mationen in einem Berichtswesen zu erfassen und in geeigneter Form für den Informationsempfänger aufzubereiten. Für die Datenerhebung sind Methoden zu definieren, die eine Messung von Merkmalen der IT-Dienstleistungen und technischen Komponenten ermöglichen. Durch Ermittlung des Investitionsertrags (engl. Return on Invest) kann der Mehrwert der Verbesserung gegenüber einem Ausgangszustand ermittelt werden (vgl. Olbrich, 2008, S. 157 f.).

2.3 Zusammenfassung

Das Rechenzentrum ist ein IT-Dienstleister. Es erbringt Infrastrukturdienstleistungen für Kunden. Diese Infrastrukturdienstleistungen werden häufig mit weiteren IT-Dienstleistungen zu einem Dienstleistungsbündel zusammengefasst und als IT-Produkte angeboten. Die Bereitstellung und Erbringung von IT-Dienstleistungen ist Aufgabe des IT-Servicemanagements.
ITIL ist ein Rahmenwerk zur Umsetzung des IT-Service-Managements. Es beschreibt Ziele, Aufgaben, Rollen und Prozesse jedoch nicht deren konkrete Umsetzung. Der ITIL Lebenszyklus besteht aus den Phasen Servicestrategie, Service Design, Serviceüberführung, Servicebetrieb und Kontinuierliche Serviceverbesserung. Die Kapazitätswirtschaft ist Bestandteil der Phase Service Design des ITIL Lebenszyklus. Im weiteren Verlauf der Arbeit wird untersucht, wie sich die Aufgaben der Kapazitätswirtschaft der Phase Service Design durch Methoden der Produktionsplanung und -steuerung umsetzen lassen.

Kapitel 3

Abbildung des Rechenzentrums als Produktionsbetrieb

Im vorigen Kapitel wurde das Rechenzentrum als IT-Dienstleister beschrieben. Zur Ablösung der ereignisgetriebenen Betriebsstrukturen sollen geeignete Verfahren der operativen Produktionsplanung und -steuerung für den Betrieb eines Rechenzentrums adaptiert werden. In diesem Kapitel soll folgende Hypothese überprüft werden:

> IT-Dienstleistungen lassen sich in einem System zur Produktionsplanung und -steuerung abbilden.

Bei der Bestätigung der Hypothese wird von der Bedingung ausgegangen, dass der Einsatz von Virtualisierungstechniken aus Sicht der Kapazitätswirtschaft sinnvoll ist. Zielsetzung ist die Abbildung des Prozesses zur Erbringung einer IT-Dienstleistung auf einen Produktionsprozess. Als Ergebnis liegt ein Referenzmodell für die Umsetzung dieser Anforderung vor. Das Referenzmodell soll anhand eines Prototypen evaluiert werden. Am Beispiel des Enterprise Resource Planning Systems SAP ERP wird aufgezeigt, wie sich diese Prozesse in der Praxis umsetzen lassen. Weiterhin wird ein Modell erstellt, das die Bewertung der Qualität der erbrachten IT-Dienstleistung aus Sicht der Kapazitätswirtschaft ermöglicht. Als Bewertungsgrundlage dienen die Bewegungsdaten (vgl. Schuh, 2006, S. 77) des Systems zur Produktionsplanung und -steuerung.
Das in Abschnitt 3.1 beschriebene Referenzmodell wurde bereits in der Fachzeitschrift für Information Management & Consulting (vgl. Osterburg et al.,

2009b, S. 65 ff.) publiziert. Der in Abschnitt 3.2 beschriebene Implementierungsansatz für SAP ERP wurde in Auszügen bereits in der Fachzeitschrift PPS-Management veröffentlicht (vgl. Osterburg und Pinnow, 2009, S. 27 ff.).

3.1 Produktionsplanung und -steuerung im Rechenzentrum

Der folgende Abschnitt zeigt auf, wie der Prozess der Bereitstellung und des Betriebs der Informationsinfrastruktur zur Erbringung einer IT-Dienstleistung durch Methoden der Produktionsplanung und -steuerung abgebildet werden kann. Hierzu werden im ersten Schritt die Datenstrukturen der Produktionsplanung und -steuerung beschrieben.

3.1.1 Datenstrukturen der Produktionsplanung und -steuerung

Die Aufgaben einer Betriebswirtschaft, die sich mit Information und Kommunikation als wirtschaftlichem Gut befassen, werden als Informationsfunktion bezeichnet (vgl. Heinrich und Lehner, 2005, S. 19).
Aus Sicht des Betreibers eines Rechenzentrums werden dem Kunden als Konsumenten physische und immaterielle Potentialfaktoren zur Ausführung der Informationsfunktion zur Verfügung gestellt. Potentialfaktoren stellen ihre Leistungspotentiale dem Produktionsprozess zur Verfügung, ohne ihre produktive Wirksamkeit innerhalb einer abgegrenzten Periode zu verlieren. Potentialfaktoren sind die Arbeitsleistungen der Arbeitskräfte sowie Betriebsmittel (vgl. Zäpfel, 2001, S. 16 ff.). Der Betrieb dieser Potentialfaktoren ist die zu erbringende IT-Dienstleistung und lässt sich als diskretes Fertigungsprodukt betrachten. Der Zeitraum, für den die IT-Dienstleistung dem Kon-

sumenten erbracht wird, entspricht der Produktionszeit des Fertigungsprodukts. Die vom Konsumenten beziehbaren Zeiteinheiten werden als Produktionszeit einer Mengeneinheit des Fertigungsprodukts definiert. Dies führt zu einer Auslastung der zur Erbringung der IT-Dienstleistung verwendeten Potentialfaktoren. Der Aufwand für die Bereitstellung der Potentialfaktoren wird als Rüstvorgang betrachtet. Vom Fertigungsprodukt IT-Dienstleistung wird für den Konsumenten eine festgelegte Menge als zeitlich geschlossener Posten hergestellt. Ein solcher Produktionsprozess ist als Form der Serienfertigung definiert (vgl. Gutenberg, 1983, S. 109). Jede produzierte Serie ist genau einem Kundenauftrag zugeordnet. Bei Kundenaufträgen mit einer langen Laufzeit ist der Übergang zur Massenfertigung fließend. Jede zu produzierende Serie erfordert ein Umrüsten der Betriebsmittel, da die Potentialfaktoren entsprechend den Kundenaufträgen konfiguriert werden müssen.

Endprodukt IT-Dienstleistung

Das Rechenzentrum als Produktionssystem erzeugt IT-Dienstleistungen als Ausbringung der Informationsfunktion. Diese IT-Dienstleistungen werden den Konsumenten als IT-Produkte zur Verfügung gestellt (vgl. Zarnekow et al., 2006, S. 16 ff.). IT-Dienstleistungen sind somit die Endprodukte der Produktion. Für jede Art von IT-Dienstleistung ist ein eigenes Endprodukt zu spezifizieren. Die gleiche Art von IT-Dienstleistungen kann mit unterschiedlichen Ausprägungen des Leistungsvermögens als eigenständiges IT-Produkt angeboten werden. Das Leistungsvermögen des IT-Produkts wird als Kapazität bezeichnet (vgl. Kern, 1992, S. 21). Aus dem Leistungsvermögen des IT-Produkts ergeben sich Anforderungen an das Leistungsvermögen der zur Produktion benötigten Potentialfaktoren. Aus diesem Grund sind die Endprodukte nach ihrem Leistungsvermögen zu differenzieren.
IT-Dienstleistungen werden als Endprodukte in Eigenfertigung vom Rechenzentrum erstellt. Die Mengeneinheit des Endprodukts wird in Zeiteinheiten angegeben und soll der Zeiteinheit des angebotenen IT-Produkts entsprechen. Kann ein Konsument beispielsweise eine IT-Dienstleistung für eine

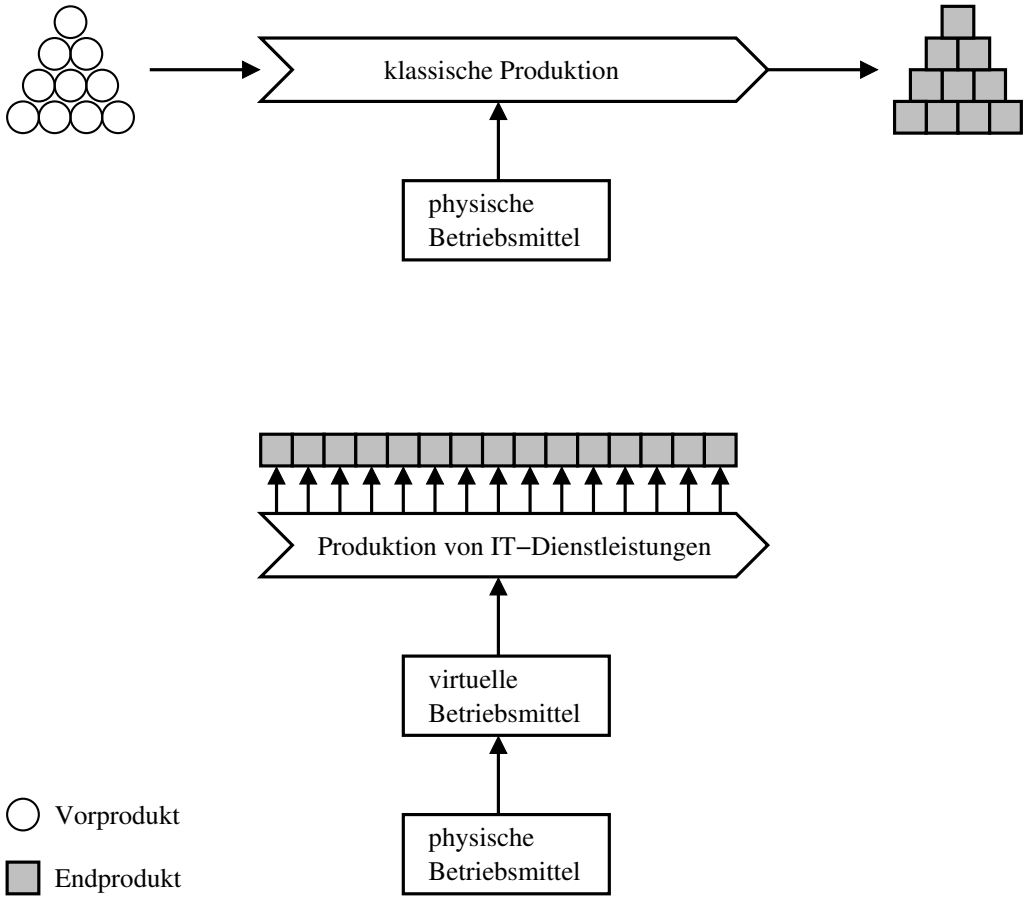

Abbildung 3.1: Produktion von IT-Dienstleistungen

bestimmte Menge von Tagen in Anspruch nehmen, ist auch die Mengeneinheit des Endprodukts in Tagen zu dimensionieren. Die beschriebenen Eigenschaften des Endprodukts sind in den Teilestammdaten des jeweiligen IT-Produkts zu spezifizieren.

Dienstleistungen sind für den Absatz produzierte immaterielle Wirtschaftsgüter (vgl. Malerie und Frietzsche, 2008, S. 5). Zur Erstellung einer IT-Dienstleistung als Endprodukt werden somit keine materiellen Vorprodukte oder Rohstoffe benötigt (vgl. Malerie und Frietzsche, 2008, S. 100). Die Erzeugnisstruktur des IT-Produkts besteht deshalb nur aus einem Knoten,

der das Endprodukt selbst repräsentiert.
Abbildung 3.1 stellt die Produktion von IT-Dienstleistungen aus Anbietersicht der klassischen Produktion gegenüber. In der klassischen Produktion werden Vorprodukte durch Einsatz von Potentialfaktoren in Endprodukte transformiert, die am Ende des Produktionsprozesses dem Konsumenten zur Verfügung stehen. Bei der Produktion von IT-Dienstleistungen kann aus Anbietersicht der Produktionsprozess selbst als Endprodukt betrachtet werden, da dieser zur Auslastung der Potentialfaktoren führt. Aus Sicht des Konsumenten kann die Ausführung der Informationsfunktion innerhalb dieses Produktionsprozesses unter Einsatz der Potentialfaktoren realisiert werden. Im weiteren Verlauf der Arbeit wird die Ausführung der Informationsfunktion aus Sicht des Konsumenten nicht näher betrachtet, da diese für die Kapazitätsplanung aus Sicht des Betreibers des Rechenzentrums nicht relevant ist. Zur Herstellung des Endprodukts IT-Dienstleistung werden Betriebsmittel verwendet.

Betriebsmittel im Rechenzentrum

Betriebsmittel sind technische Mittel, die zur Leistungserstellung benötigt werden und die ihr Nutzungspotential über längere Zeiträume abgeben. Führt der planmäßige Einsatz der technischen Mittel zu einem Fertigungsfortschritt, werden diese als Betriebsmittel mit Werkverrichtungen bezeichnet. Das Rechenzentrum als Gebäude kann als Betriebsmittel ohne Werkverrichtung angesehen werden, da es die Fertigungsvorbedingung darstellt (vgl. Zäpfel, 2001, S. 18).
Der Aufbau eines Rechenzentrums kann anhand seiner Informationsinfrastruktur beschrieben werden. Die Informationsinfrastruktur als Erkenntnisobjekt des Informationsmanagements umfasst die Einrichtungen, Mittel und Maßnahmen, die zur Produktion, Verbreitung und Nutzung von Informationen im Unternehmen benötigt werden (vgl. Heinrich und Lehner, 2005, S. 19).
Anhand dieser Definition lassen sich die in Abbildung 3.2 dargestellten

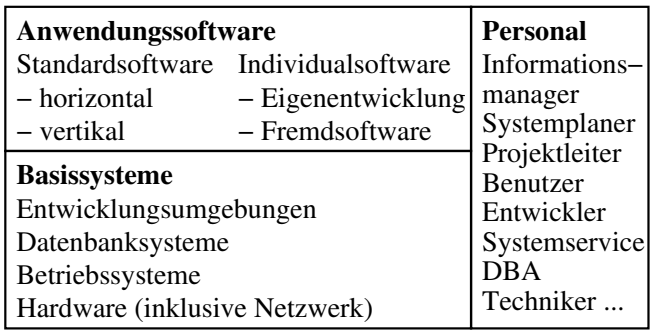

Abbildung 3.2: Informationsinfrastruktur (nach Rautenstrauch, 1997, S. 13)

Hauptkomponenten Anwendungssoftware, Basissysteme und Personal identifizieren. Anwendungssoftware und Basissysteme stellen die Betriebsmittel mit Werkverrichtung dar. Für jede Infrastrukturkomponente sind Betriebsmitteldaten zu definieren. Betriebsmittel können auf mehrere Rechenzentren verteilt sein und sind deshalb eindeutig einem Standort zuzuordnen.
Das Kapazitätsangebot einer Infrastrukturkomponente lässt sich zum einen über die technische Leistungsfähigkeit und zum anderen über die zeitliche Verfügbarkeit beschreiben. Das zeitliche Kapazitätsangebot eines Betriebsmittels im Rechenzentrum beträgt üblicherweise 24 Stunden pro Tag. Das technische Leistungsvermögen als Kapazitätsangebot eines Betriebsmittels ergibt sich aus seinen technischen Spezifikationen.
Rechnersysteme als Betriebsmittel lassen sich als System aus den Hardwarekomponenten Prozessor, Speicher und Ein-/Ausgabegeräten betrachten. Ein-/Ausgabegeräte werden zur Kommunikation mit dem Rechnersystem verwendet. Der Prozessor ruft Instruktionen aus dem Speicher ab, dekodiert diese Instruktionen und führt sie anschließend aus. Die Hardwarekomponente Speicher lässt sich in Primär- und Sekundärspeicher unterteilen. Primärspeicher verwaltet die Daten während des unmittelbaren Betriebs von Anwendungs- und Basissystemen (vgl. Tanenbaum, 1999, S. 113). Sekundärspeicher ist nichtflüchtiger Speicher und meist ein als Plattenspeicher ausgebildetes Medium. Die Kapazität des Sekundärspeichers ist üblicherweise ein Vielfaches größer als die Kapazität des Primärspeichers (vgl. Saake und Heuer, 1999, S. 44). Jede Hardwarekomponente verfügt über ein technisches Leistungsvermögen, das sich durch ein geeignetes Maß beschreiben lässt.

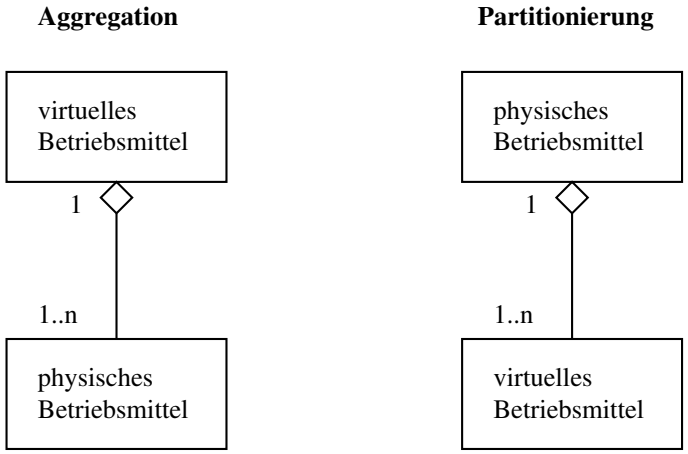

Abbildung 3.3: Virtualisierung von Betriebsmitteln

Eine mögliche Kenngröße zur Beschreibung der Prozessorkapazität ist die Anzahl an Operationen, die von einem Prozessor pro Sekunde ausgeführt werden kann (vgl. Langendörfer, 1992, S. 10). Die Kapazitätsbewertung erfolgt in Millionen Instruktionen pro Sekunde (MIPS). Die Kommunikation der Ein-/Ausgabegeräte erfolgt über Kanäle. Ein Kanal ist die Verbindung zwischen dem Sender der Information und dem Empfänger. Als Maß zur Bestimmung der maximalen Übertragungsgeschwindigkeit eines Kanals wird die Bandbreite verwendet (vgl. Klimant et al., 2006, S. 75 f.). Die Bewertung der Übertragungsgeschwindigkeit erfolgt in bit/s (vgl. Klimant et al., 2006, S. 86). Die Speicherkapazität ist eine Maßzahl, die definiert, wie viele Speicherzellen zur Speicherung von Binärzeichen auf dem Medium enthalten sind. Die Kapazitätsbewertung erfolgt in Bit oder Byte (vgl. Schneider, 1998, S. 802). Die Kommunikation mit externen Speichersystemen erfolgt über Ein-/Ausgabegeräte. Neben der Speicherkapazität ist zusätzlich die Kanalbandbreite der Kommunikationskapazität zu berücksichtigen.

Für eine effiziente Ausnutzung der vorhandenen Betriebsmittel stehen heute Konzepte wie Virtual und Adaptive Computing zur logischen Trennung von Hard- und Software zur Verfügung. Virtualisierungskonzepte erlauben die Aggregation und Partitionierung physischer Betriebsmittel in virtuelle Betriebsmittel (vgl. Osterburg et al., 2009a, S. 120 f.). Die Beziehung zwischen virtuellen und physischen Betriebsmitteln lässt sich über Betriebsmittelhierarchien abbilden. Abbildung 3.3 veranschaulicht die beiden

Virtualisierungskonzepte. Das zeitliche Kapazitätsangebot eines physischen Betriebsmittels lässt sich durch Partitionierung in n virtuelle Betriebsmittel um den Faktor n vervielfachen. Das technische Leistungsvermögen physischer Betriebsmittel wird durch Aggregation summiert oder durch Partitionierung auf die virtuellen Betriebsmittel aufgeteilt. Für den Betrieb virtueller Betriebsmittel wird ebenfalls technische Kapazität des physischen Betriebsmittels in Anspruch genommen. Dieser Kapazitätsbedarf wird als Virtualisierungsoverhead bezeichnet (vgl. Smith und Nair, 2005, S. 166 f.) und limitiert den technischen Nutzungsgrad des physischen Betriebsmittels. Der Einsatz von Virtualisierungstechniken beeinflusst den Kapazitätsbedarf und somit auch das benötigte Kapazitätsangebot. Die Auswirkungen der Virtualisierung auf das Kapazitätsangebot werden in Kapitel 4 untersucht. Im Gegensatz zur konventionellen Produktion können durch Virtualisierung Betriebsmittel kurzfristig entsprechend einem Fertigungsauftrag angelegt oder bezüglich des technischen Leistungsvermögens angepasst werden. Diese Möglichkeiten sind bei der Erstellung der Arbeitspläne zu berücksichtigen.

Arbeitspläne

In Arbeitsplänen werden die für die Produktion eines Auftrags vorgesehenen Arbeitsoperationen in den jeweiligen Bearbeitungsstationen in ihrer logischen Abfolge erfasst. Es wird festgelegt, welche Arbeitsoperationen vor Beginn des nächsten Arbeitsschritts abgeschlossen sein müssen (vgl. Adam, 1998, S. 556). Die Arbeitsschritte zur Erstellung des Endprodukts IT-Dienstleistung lassen sich in die Phasen Bereitstellung und Betrieb der Informationsinfrastruktur unterteilen.
Die Phase des Betriebs der Informationsinfrastruktur wird in einem Arbeitsschritt abgebildet. Die Bearbeitungszeit zur Erstellung einer Mengeneinheit entspricht der definierten Mengeneinheit des Endprodukts IT-Dienstleistung. Werden beispielsweise als Mengeneinheit für das Endprodukt Tage definiert, beträgt die Bearbeitungszeit zur Erstellung eines IT-Dienstleistungstages 24 Stunden. Für die Phase des Betriebs werden Anwendungssoftware und

Basissysteme als Infrastrukturkomponenten eingesetzt. Die benötigten Infrastrukturkomponenten werden als Betriebsmittel dem Arbeitsschritt des Betriebs der Informationsinfrastruktur zugeordnet. Im Arbeitsplan ist festzulegen, dass eine Aufspaltung der zu produzierenden Gesamtmenge in mehrere Teilmengen nicht zulässig ist. Das Splitten von Fertigungsaufträgen wird üblicherweise zur Verkürzung der Durchlaufzeiten verwendet (vgl. Kurbel, 2005, S. 148). Da über die Bearbeitungszeit des Arbeitsschritts zum Betrieb der Informationsinfrastruktur der Zeitraum der Leistungserstellung abgebildet wird, kann diese Methode jedoch nicht eingesetzt werden.
Die Ausführung dieses Arbeitsschritts führt zu einer Reduzierung des Kapazitätsangebots von Anwendungssoftware und Basissystemen für einen definierten Zeitraum in Abhängigkeit von der produzierten Menge IT-Dienstleistung in Zeiteinheiten. In der Phase des Betriebs wird die Infrastrukturkomponente Personal üblicherweise nur für Aufgaben der Wartung und Instandhaltung benötigt. Diese Tätigkeiten sind jedoch nicht Bestandteil von Arbeitsplänen sondern werden in Wartungs- und Instandhaltungsplänen erfasst.
Die Vorbereitung der Betriebsmittel zur Erfüllung einer Arbeitsaufgabe wird als Rüsten bezeichnet (vgl. Zäpfel, 2001, S. 71). Die Phase der Bereitstellung der Informationsinfrastruktur ist somit die Rüstzeit zur Erstellung des Endprodukts IT-Dienstleistung. Im Arbeitsschritt Betrieb der Informationsinfrastruktur werden verschiedene Infrastrukturkomponenten verwendet, deren Arbeitsvorbereitung in mehreren aufeinander abgestimmten Arbeitsschritten erfolgt.
Als erstes werden die Basissysteme gerüstet. Die Installation und Konfiguration jeder physischen und virtuellen Hardwarekomponente wird in einem eigenen Arbeitsschritt abgebildet. Die Ausführung dieser Arbeitsschritte kann parallel erfolgen. Im nächsten Arbeitsschritt wird auf der Hardware das Betriebssystem installiert. Im Folgeschritt wird ein Datenbankmanagementsystem auf dem Betriebssystem eingerichtet, sofern dies zur Erstellung der IT-Dienstleistung benötigt wird.
Nachdem die Arbeitsvorbereitung der Basissysteme abgeschlossen ist, wird die Anwendungssoftware auf die Basissysteme aufgesetzt. Die Installation

und Konfiguration jeder Anwendungssoftware wird in einem eigenen Arbeitsschritt abgebildet. Solange keine Abhängigkeitsbeziehungen zwischen den Infrastrukturkomponenten Anwendungssoftware bestehen, können diese Arbeitsschritte parallel durchgeführt werden. Die Ausführung der Arbeitsschritte zum Rüsten der Betriebsmittel ist teilweise automatisierbar. Zur Durchführung nichtautomatisierter Arbeitsschritte wird Personal eingesetzt. Für manuell auszuführende Rüsttätigkeiten ist die erforderliche Qualifikation des Personals in den Arbeitsplänen zu spezifizieren.
Der Einsatz von Virtualisierungstechniken hat zur Folge, dass zum Zeitpunkt der Erstellung der Arbeitspläne die eingesetzten Betriebsmittel nicht vollständig bekannt sind. Virtuelle Betriebsmittel werden im Arbeitsplan durch Planbetriebsmittel beschrieben. Ein Planbetriebsmittel wird durch sein technisches Leistungsvermögen und die Betriebsmittelart spezifiziert und im Fertigungsauftrag durch ein virtuelles Betriebsmittel substituiert.
Die beschriebenen Datenstrukturen bilden die Grundlage zur Umsetzung der Methoden der Produktionsplanung und -steuerung.

3.1.2 Methoden der Produktionsplanung und -steuerung

Ein Produktionsbetrieb erzeugt Zwischenprodukte oder Endprodukte für Konsumenten. Der Primärbedarf an Zwischen- und Endprodukten wird als Produktionsprogramm bezeichnet. Eine Methode zur Ermittlung des Produktionsprogramms ist die Produktionsprogrammplanung. Aufgabe der Produktionsprogrammplanung ist die Definition der Mengen an Produkten, die in einer gegebenen Periode produziert werden.
Ausgangspunkt der Produktionsprogrammplanung ist eine vorgegebene Menge an IT-Produktarten sowie die gegebenen Kapazitäten der Infrastrukturkomponenten Anwendungssoftware, Basissysteme und Personal. Das Produktionsprogramm bildet die Grundlage der Material- und Kapazitätsbedarfsplanung. Da Dienstleistungen immaterielle Wirtschaftsgüter

darstellen, ist die Durchführung einer Materialbedarfsplanung nicht notwendig.

Kapazitätsbedarfsplanung

Die Planung des Kapazitätsbedarfs kann nicht ausschließlich auf Basis der Produktionsprogrammplanung durchgeführt werden. So führt zum Beispiel der Release-Wechsel von Softwareprodukten, die zur Erbringung von IT-Dienstleistungen benötigt werden, zu einer Änderung des Kapazitätsbedarfs an Basissystemen. Zur Vorhersage des erforderlichen Kapazitätsbedarfs kann auf Basis des in der Vergangenheit beobachteten Bedarfs deren Entwicklung in die Zukunft extrapoliert werden. Der Kapazitätsbedarf an Basissystemen kann durch entsprechende Monitoringwerkzeuge ermittelt werden. Zur Prognose des zukünftigen Kapazitätsbedarfs stehen Prognoseverfahren wie die lineare Regressionsrechnung (vgl. Tempelmeier, 2006, S. 51 ff.), die exponentielle Glättung erster Ordnung (vgl. Brown und Meyer, 1960, S. 673 ff.) oder das Verfahren von Holt (vgl. Holt, 2004, S. 5 ff.) zur Verfügung. In Kapitel 5 wird die Kapazitätsbedarfsplanung durch Prognoseverfahren ausführlich beschrieben.
Im Rahmen der Kapazitäts- und Zeitwirtschaft wird die Betriebsmittelbelegung unter Beachtung zeitlicher Restriktionen und der verfügbaren Kapazitäten der Betriebsmittel geplant.

Kapazitäts- und Zeitwirtschaft

Für Kundenaufträge zur Erbringung von IT-Dienstleistungen werden Fertigungsaufträge auf Basis von Arbeitsplänen angelegt. Die Vergabe von Start- und Endterminen für die Fertigungsaufträge wird als Durchlaufterminierung bezeichnet. Bei der Durchlaufterminierung werden Verfahren zur Vorwärts-, Rückwärts- und zur Doppelten Terminierung verwendet (vgl. Kurbel, 2005, S. 139 ff.). Der Fertigungsauftrag umfasst sowohl den Arbeitsschritt des Betriebs der Informationsinfrastruktur als auch die Arbeitsschritte zur Bereitstellung. Start- und Endtermin des Betriebs sind durch den Kundenauftrag zur Erbringung der IT-Dienstleistung festgelegt. Im Rahmen der

Durchlaufterminierung sind die Start- und Endtermine der Arbeitsschritte zur Bereitstellung so zu definieren, dass mit dem Betrieb der Infrastruktur entsprechend dem Kundenauftrag begonnen werden kann. Diese Restriktionen lassen sich durch Einsatz der Rückwärtsterminierung erfüllen, indem ausgehend vom Endtermin der Erbringung der IT-Dienstleistung alle Arbeitsschritte rückwärtsschreitend terminiert werden. Liegt der Starttermin des ersten Arbeitsschritts der Bereitstellung in der Vergangenheit, kann der Kundenauftrag in der vorliegenden Form nicht erfüllt werden. In diesem Fall kann durch Vorwärtsterminierung der früheste Starttermin für den Betrieb der Informationsinfrastruktur ermittelt werden. Der Kundenauftrag ist dem Terminierungsergebnis anzupassen, da der vereinbarte zeitliche Rahmen zur Erbringung der IT-Dienstleistung nicht eingehalten werden kann.
Bei der Durchlaufterminierung wird das tatsächlich vorhandene Kapazitätsangebot nicht berücksichtigt. Im Rahmen der Kapazitätsplanung wird das Kapazitätsangebot mit dem Kapazitätsbedarf durch Anpassungsmaßnahmen in Einklang gebracht. Ein Defizit des Kapazitätsangebots an Betriebsmitteln ist durch Beschaffung weiter Hardwarekomponenten zu beseitigen. Lässt sich der Starttermin der Durchlaufterminierung des Arbeitsschritts zur Bereitstellung eines Betriebsmittels nicht mit der Wiederbeschaffungszeit vereinbaren, kann der Kundenauftrag in der vorliegenden Form nicht erfüllt werden und muss angepasst werden. Kapazitätsengpässe treten üblicherweise beim Personal auf, das für die Bereitstellung der Betriebsmittel zuständig ist. Eine Beseitigung dieses Kapazitätsengpasses erfolgt durch terminliche Anpassung, indem Arbeitsschritte unter Beachtung der zeitlichen Reihenfolge auf einen früheren Starttermin verschoben werden. Als Ergebnis der Durchlaufterminierung und Kapazitätsplanung liegt ein Grobplan für Produktionsmengen und -termine vor, der die Grundlage der Fertigungssteuerung darstellt.

Fertigungssteuerung

Im Rahmen der Fertigungssteuerung werden die grobgeplanten Aufträge zur Fertigung freigegeben. Die Auftragsfreigabe stellt den Übergang von der Pro-

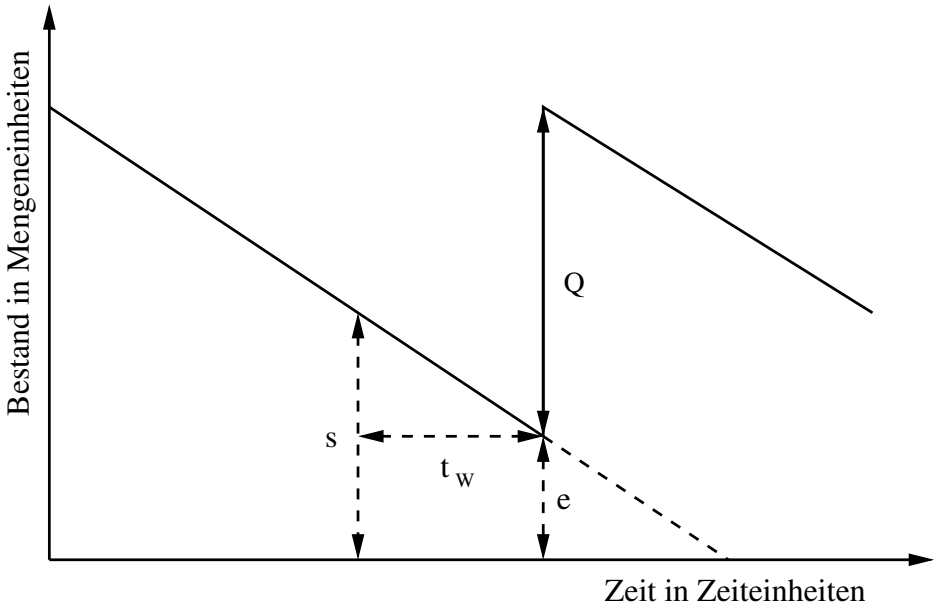

Abbildung 3.4: (s, Q)-Politik (nach Mertens, 2004, S. 81)

duktionsplanung zur Produktionssteuerung dar. Die freigegebenen Aufträge werden im Rahmen der Feinterminierung konkreten Einzelbetriebsmitteln zugeordnet.

In den zugrundeliegenden Arbeitsplänen wurden die virtuellen Betriebsmittel durch Planbetriebsmittel beschrieben. Diese Planbetriebsmittel sind vor der Auftragsfreigabe durch virtuelle Betriebsmittel zu ersetzen. Die virtuellen Betriebsmittel können entweder vor der Auftragsfreigabe entsprechend den Spezifikationen der Planbetriebsmittel angelegt werden oder auf Vorrat bereitgestellt werden. Die Vorratsbereitstellung kann auf Basis der sogenannten (s, Q)-Politik durchgeführt werden. Der Bestand an eingerichteten, noch nicht zur Produktion eingesetzten virtuellen Betriebsmitteln wird als Lagerbestand betrachtet. Nach jeder Freigabe eines Fertigungsauftrags wird überprüft, ob der Bestand an nicht zur Produktion eingesetzten Betriebsmitteln einer Art den Wert s unterschreitet. Ist dies der Fall, wird ein neues Los an virtuellen Betriebsmitteln der Größe Q eingerichtet. Die Wiederbeschaffungszeit t_w entspricht der Zeit, die zum Einrichten eines Loses Q virtueller Betriebsmittel benötigt wird. Der Bestand s ist so zu dimensionieren, dass der Bedarf an Betriebsmitteln während der Wiederbeschaffungszeit t_w nicht den Sicherheitsbestand e beansprucht (siehe Abbildung 3.4).

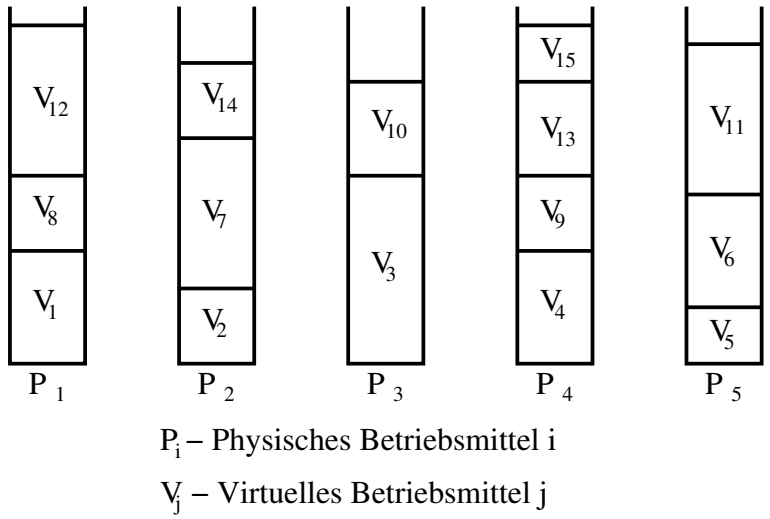

Abbildung 3.5: Zuordnung virtueller Betriebsmittel als Bin-Packing-Problem

Planbetriebsmittel beschreiben den Kapazitätsbedarf als technisches Leistungsvermögen, der an ein virtuelles Betriebsmittel gestellt wird. Ein virtuelles Betriebsmittel wird nach den Spezifikationen eines Planbetriebsmittels eingerichtet und ist einem physischen Betriebsmittel zugeordnet. Ein physisches Betriebsmittel stellt ein Kapazitätsangebot als technisches Leistungsvermögen zur Verfügung.

Die Frage der Zuordnung der virtuellen Betriebsmittel zu physischen Betriebsmitteln lässt sich im Fall der Virtualisierung durch Partitionierung als Bin-Packing-Problem beschreiben. Das Bin-Packing-Problem ist ein NP-vollständiges kombinatorisches Optimierungsproblem. Es beschreibt die Fragestellung der Verteilung einer Menge an Objekten mit einer definierten Größe auf eine Menge an Behältern mit einer vorgegebenen Größe (vgl. Galambos und Woeginger, 1995, S. 25). Im Fall der Virtualisierung werden die physischen Betriebsmittel als Behälter betrachtet. Die virtuellen Betriebsmittel werden durch die zu verteilenden Objekte repräsentiert (siehe Abbildung 3.5). Die Größe der Behälter und Objekte wird durch das technische Leistungsvermögen der Betriebsmittel bestimmt. Wird das technische

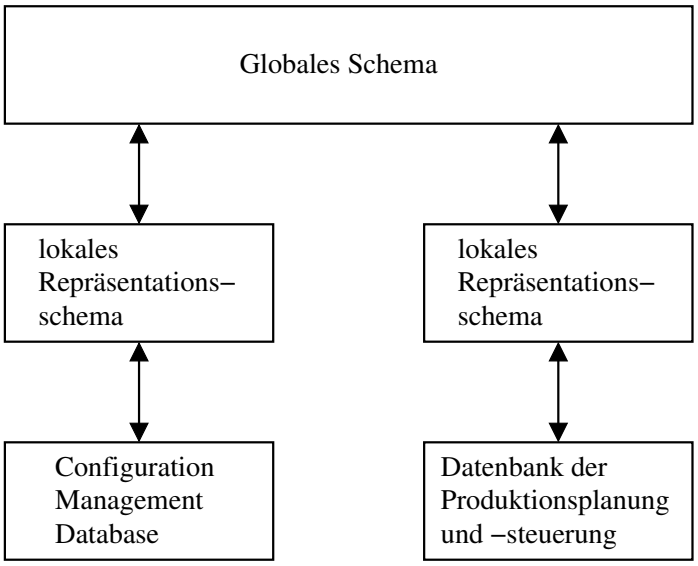

Abbildung 3.6: Realisierung der Datenintegration

Leistungsvermögen eines Betriebsmittels durch mehrere Größen spezifiziert, liegt ein mehrdimensionales Bin-Packing-Problem vor (vgl. Galambos und Woeginger, 1995, S. 40). In Kapitel 6 wird die Betriebsmittelzuordnung durch Lösung des Bin-Packing-Problems ausführlich dargestellt.
ITIL fordert den Einsatz einer Configuration Management Database zur Verwaltung sämtlicher Konfigurationselemente. Die Umsetzung der Kapazitätswirtschaft in einem System zur Produktionsplanung und -steuerung hat zur Folge, dass die Verwaltung der Konfigurationselemente teilweise von diesem System übernommen wird. Zur Vermeidung redundanter Datenhaltung ist eine Postintegration beider Systeme vorzunehmen. Postintegration bezeichnet die nachträgliche Integration von Datenbanken zu einer verteilten Datenbank (vgl. Dadam, 1996, S. 96 ff.). Hierbei wird für beide Systeme ein lokales Repräsentationsschema eingeführt, welches die lokalen Relationen, die global zur Verfügung gestellt werden sollen, in einer einheitlichen strukturellen Form repräsentiert. Die beiden lokalen Repräsentationsschemata werden in einem globalen Schema integriert. Abbildung 3.6 veranschaulicht die Integration. Im folgenden Abschnitt wird die Umsetzung der beschriebenen Methoden und Datenstrukturen für SAP ERP beschrieben.

3.2 Implementierungsansatz für SAP ERP

Im vorigen Abschnitt wurde aufgezeigt, wie die Methoden der Produktionsplanung und -steuerung auf den Betrieb eines Rechenzentrums angewendet werden können. In diesem Abschnitt wird an einem einfachen Beispiel untersucht, wie sich diese Methoden in SAP ERP implementieren lassen. Hierzu wird im ersten Schritt eine geeignete Fertigungsart gewählt.

3.2.1 Auswahl der Fertigungsart

Die Fertigungsart definiert das Verfahren der Fertigung. In SAP ERP wird zwischen diskreter Fertigung und Prozessfertigung unterschieden. Serien- und Kanbanfertigung sind spezielle Arten der diskreten Fertigung. Bei der Kanbanfertigung erfolgt die Produktionssteuerung auf Basis des Materialbestands (vgl. Dickmann, 2007, S. 181 ff.). Da das Endprodukt IT-Dienstleistung ein immaterielles Gut ist, das weder materielle Vorprodukte noch Rohstoffe benötigt, wird diese Fertigungsart nicht weiter betrachtet.
In Abschnitt 3.1.1 wurde der Produktionsprozess als Form der Serienfertigung definiert. In SAP ERP wird die Form der Serienfertigung eingesetzt, wenn (vgl. SAP, 2007):

- dasselbe Erzeugnis über einen längeren Zeitraum hinweg produziert wird,
- über einen bestimmten Zeitraum eine Gesamtmenge mit einer Produktionsrate pro Teilperiode erzeugt wird und
- die Erzeugnisse die Betriebsmittel immer in der gleichen Reihenfolge durchlaufen.

Für die Phase des Betriebs der Informationsinfrastruktur treffen diese Anforderungen zu. Für einen Kunden werden über einen längeren Zeitraum hinweg identische Zeiteinheiten des Endprodukts IT-Dienstleistung produziert. Die Gesamtmenge an Zeiteinheiten des Endprodukts IT-Dienstleistung wird mit einer definierten Produktionsrate erzeugt. Je Zeiteinheit wird genau eine Mengeneinheit des Endprodukts erzeugt. Die Phase des Betriebs wird in einem Arbeitsschritt abgebildet, der stets dieselben Betriebsmittel verwendet (siehe Abschnitt 3.1.1). Als Terminierungsverfahren wird die Taktterminierung eingesetzt. Hierbei werden im Gegensatz zur Durchlaufterminierung nicht die Bearbeitungszeiten einzelner Arbeitsschritte sondern Taktzeiten verwendet. Es werden keine Kapazitätsbedarfe erzeugt (vgl. SAP, 2007). In der Phase des Betriebs entspricht die Taktzeit der Zeiteinheit des Endprodukts IT-Dienstleistung. Ist die Mengeneinheit des Endprodukts beispielsweise in Tagen dimensioniert, beträgt die Taktzeit zur Produktion eines Dienstleistungstages 24 Stunden.
Für die Phase der Bereitstellung der Informationsinfrastruktur ist der Einsatz der Serienfertigung jedoch nicht geeignet. Zur Herstellung einer Serie von Zeiteinheiten des Endprodukts IT-Dienstleistung werden einmalig mehrere aufeinander abgestimmte Arbeitsschritte zum Rüsten der Betriebsmittel ausgeführt. In der Phase der Bereitstellung müssen Kapazitätsengpässe insbesondere beim Personal berücksichtigt werden. Diese Anforderung wird durch eine Rückwärtsterminierung der Bereitstellung (siehe Abschnitt 3.1.2) erfüllt. Da bei der Taktterminierung keine Zuführungslinien berücksichtigt werden (vgl. SAP, 2007), ist der Einsatz der diskreten Fertigung für die Bereitstellung mit anschließender Serienfertigung der Phase des Betriebs der Informationsinfrastruktur nicht sinnvoll.
Werden die Phasen Bereitstellung und Betrieb in einem Arbeitsplan abgebildet, würde sich die Fertigungsart der diskreten Fertigung anwenden lassen. Der Einsatz von Virtualisierungstechniken hat jedoch zur Folge, dass zum Zeitpunkt der Erstellung der Arbeitspläne die eingesetzten virtuellen Betriebsmittel noch nicht existieren und somit noch nicht bekannt sind. Die Verwendung von Planbetriebsmitteln bei der Erstellung von Arbeitsplänen ist in der diskreten Fertigung nicht vorgesehen. Die diskrete Fertigung ist

deshalb zur Produktionsplanung und -steuerung eines virtualisierten Rechenzentrums nicht geeignet.
Die Fertigungsart der Prozessfertigung ermöglicht den Einsatz von Planbetriebsmitteln bei der Erstellung von Arbeitsplänen. Die Funktionalität der diskreten Fertigung steht auch in der Prozessfertigung zur Verfügung. Die Produktionsplanung und -steuerung des Rechenzentrums wird deshalb in der Komponente Produktionsplanung für die Prozessindustrie (PP-PI) von SAP ERP durchgeführt.
Jede produzierte Serie des Endprodukts IT-Dienstleistung ist genau einem Kundenauftrag zugeordnet. Die Fertigungsart ist somit als kundenauftragsorientierte Prozessfertigung definiert.

3.2.2 Stammdaten der Prozessfertigung

In Abschnitt 3.1.1 wurden die Datenstrukturen zur Produktionsplanung und -steuerung eines virtualisierten Rechenzentrums spezifiziert. In diesem Abschnitt wird an einem einfachen Beispiel gezeigt, wie diese Datenstrukturen in SAP ERP bei kundenauftragsorientierter Prozessfertigung abgebildet werden können. Hierzu wird exemplarisch ein Endprodukt definiert, welches dem Kunden ein SAP® Business Information Warehouse (SAP BW) als Endprodukt einer IT-Dienstleistung zur Verfügung stellt. Das technische Leistungsvermögen der Rechnersysteme wird hierbei im proprietären Kapazitätsmaß SAP Application Benchmark Performance Standard (SAPS) angegeben. Hierbei entsprechen 100 SAPS der standardisierten Verarbeitung von 2000 Bestellpositionen pro Stunde, die 6000 Benutzer-Interaktionsschritte erfordern (vgl. Mißbach, 2005, S. 145).
Die organisatorische Einheit, in der Produkte hergestellt werden, wird in SAP ERP als Werk abgebildet (vgl. Benz und Höflinger, 2008, S. 49). Das Rechenzentrum als Produktionsstätte der IT-Dienstleistung wird als Werk 1000 angelegt.

Materialstammsatz des Endprodukts

Materialspezifische Daten eines Unternehmens werden im Materialstammsatz verwaltet. Dem Kunden soll eine Anwendungssoftware SAP BW als Endprodukt mit einem technischen Leistungsvermögen von 2000 SAPS für eine bestimmte Menge von Tagen zur Verfügung gestellt werden. Das Endprodukt wird hierzu für das Werk 1000 als Material *BW-2000-SAPS* der Materialart Fertigerzeugnis angelegt. Die Basismengeneinheit des Materials ist Tage (TAG). Die Beschaffungsart des Materials ist Eigenfertigung (E), da das Endprodukt vom Rechenzentrum selbst produziert wird.

Ressourcen

Personal und Betriebsmittel werden in der Prozessfertigung als Ressourcen abgebildet. Ressourcen sind eindeutig einem Werk zugeordnet. Für die Bereitstellung des Endprodukts werden Administratoren als Personalressource benötigt. Im Beispiel wird davon ausgegangen, dass im Rechenzentrum acht Administratoren mit identischen Qualifikationen arbeiten. Im Werk 1000 wird eine Ressource *ADMIN* vom Typ Prozesseinheit angelegt. Für die Ressource *ADMIN* wird eine Kapazität vom Typ Person definiert. Die Kapazität der Administratoren besteht aus acht Einzelkapazitäten die werktags von 8:00 Uhr bis 17:00 Uhr mit einer Stunde Pause zur Verfügung stehen.
Folgende Betriebsmittel werden benötigt, um dem Kunden ein SAP BW als IT-Dienstleistung anbieten zu können:

- Rechnerkapazität als virtueller Server,
- Sekundärspeicherkapazität als virtuelle Storagepartition und
- Kommunikationskapazität als virtuelles Local Area Network (VLAN).

Für jeden eingesetzten Typ Betriebsmittel wird ein Planbetriebsmittel als Planressource *SE-00000* (Server), *ST-00000* (Storage) und *VL-00000*

Betriebsmittel	**Ressource**	**Klassifizierung**
Server	SE-00001	Leistung<=1500 SAPS
	SE-00002	Leistung<=2000 SAPS
	SE-00003	Leistung<=2500 SAPS
Storagepartition	ST-00001	Speicherplatz<=150 GByte
	ST-00002	Speicherplatz<=200 GByte
	ST-00003	Speicherplatz<=250 GByte
VLAN	VL-00001	Übertragungsrate<=1000 MBit/s
	VL-00002	Übertragungsrate<=1000 MBit/s
	VL-00003	Übertragungsrate<=1000 MBit/s

Tabelle 3.1: Klassifizierung des technischen Leistungsvermögens

(VLAN) vom Typ Prozesseinheit angelegt. Von den Administratoren werden auf Vorrat jeweils drei virtuelle Server (*SE-00001* bis *SE-00003*), drei virtuelle Storagepartitionen (*ST-00001* bis *ST-00003*) und drei VLAN (*VL-00001* bis *VL-00003*) eingerichtet. Jedes dieser Betriebsmittel soll vom Kunden 24 Stunden pro Tag genutzt werden können. Die Kapazität vom Typ Prozesseinheit steht von 0:00 Uhr bis 24:00 Uhr, sieben Tage in der Woche, ohne Pausen zur Verfügung.
Für jedes Betriebsmittel ist ein technisches Leistungsvermögen zu spezifizieren. In SAP ERP werden Klassifizierungen verwendet, um Objekte mit Merkmalen zu beschreiben. Das technische Leistungsvermögen der Betriebsmittel wird in den Ressourcen als Klassifizierungsmerkmal spezifiziert. Die Klassifizierung des technischen Leistungsvermögens der auf Vorrat bereitgestellten Betriebsmittel ist in Tabelle 3.1 zusammengefasst.

Planungsrezept

Arbeitspläne werden in der Prozessfertigung als Planungsrezepte angelegt. Planungsrezepte dienen dazu, die Herstellung von Produkten zu planen und dienen als Vorlage für die Erstellung von Prozessaufträgen. Ein Pla-

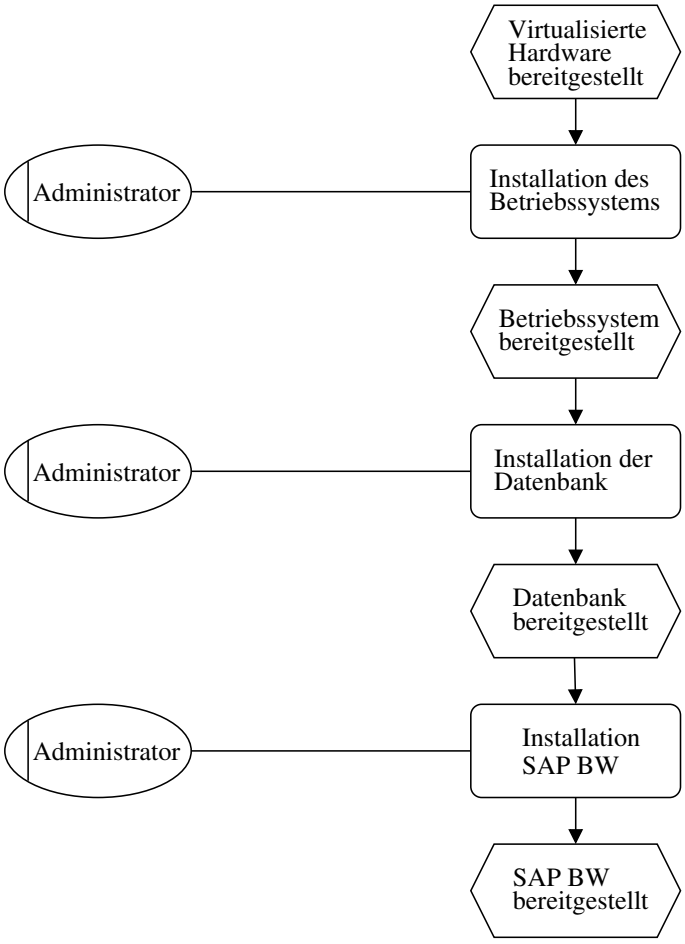

Abbildung 3.7: EPK der Bereitstellung

nungsrezept besteht aus einer Menge von Vorgängen. Jedem Vorgang ist eine Primärressource zugeordnet, auf der dieser Vorgang ausgeführt wird. Arbeitsschritte werden in der Prozessfertigung als Phasen abgebildet. Eine Phase ist eindeutig einem Vorgang zugeordnet und belegt während ihrer Ausführung die Primärressource des Vorgangs (vgl. SAP, 2007).
Die Bereitstellung und der Betrieb eines SAP BW sollen in einem Planungsrezept abgebildet werden. Die ereignisgesteuerte Prozesskette (EPK) in Abbildung 3.7 beschreibt beispielhaft den Ablauf der Bereitstellung. Die virtuellen Betriebsmittel sind auf Vorrat bereitgestellt und für die Arbeitsgänge zur Bereitstellung des SAP BW verfügbar. Im ersten Schritt wird

das Betriebssystem installiert. Auf dem Rechnersystem mit dem installierten Betriebssystem wird im nächsten Arbeitsgang das Datenbankmanagementsystem eingerichtet. Im letzten Arbeitsgang wird das SAP BW installiert. Im Beispiel werden alle Arbeitsgänge zur Bereitstellung manuell von einem Administrator ausgeführt.
Bereitstellung und Betrieb sollen in einem Planungsrezept abgebildet werden. Die im Planungsrezept eingesetzten Primärressourcen sind die Betriebsmittel Server, Storagepartition und VLAN sowie die Personalressource Administrator. Für jede Primärressource wird ein eigener Vorgang definiert. Da die tatsächlich verwendeten Betriebsmittel zum Zeitpunkt der Erstellung des Planungsrezepts nicht bekannt sind, werden die Planressourcen als Primärressourcen verwendet. In den Arbeitsgängen zur Bereitstellung werden alle vier Primärressourcen eingesetzt. Deshalb wird in jedem Vorgang für jeden Arbeitsgang eine Phase angelegt. Die Phasen eines Arbeitsgangs in den jeweiligen Vorgängen werden parallel ausgeführt. Die Arbeitsgänge zur Bereitstellung sind unabhängig von der Anzahl zu produzierender Dienstleistungstage, somit ist die Dauer der entsprechenden Phasen mengenunabhängig. Bei der Terminierung des Planungsrezepts können Kapazitätsengpässe der Personalressourcen zu einer zeitlichen Verschiebung eines Arbeitsgangs führen. Um zu verhindern, dass die Kapazitäten der Betriebsmittel zeitweise anderen Aufträgen zur Verfügung stehen, wird die Dauer der Phasen der Betriebsmittelvorgänge als dehnbar konfiguriert.
Als IT-Dienstleistung wird dem Kunden das SAP BW für eine bestimmte Menge von Tagen zur Verfügung gestellt. Für den Betrieb werden die Betriebsmittel Server, Storagepartition und VLAN benötigt. Der Betrieb des SAP BW wird in einem Arbeitsgang abgebildet. Der Arbeitsgang des Betriebs wird im Vorgang jeder Primärressource als Phase definiert. Diese Phase ist abhängig von der Anzahl zu produzierender Dienstleistungstage. Die Produktionsdauer eines Dienstleistungstages beträgt mengenabhängig 24 Stunden. Tabelle A.1 im Anhang fasst die wesentlichen Daten des Planungsrezepts zusammen.
In den Vorgängen der Betriebsmittel Server, Storagepartition und VLAN werden Planressourcen als Platzhalter für die virtualisierten Betriebsmittel

verwendet. Um die Planressourcen später gegen geeignete Ressourcen austauschen zu können, wird für jeden Vorgang eine Ressourcenauswahlbedingung spezifiziert. Im Beispiel sollen folgende Ressourcenauswahlbedingungen gelten, um ein SAP BW mit einem technischen Leistungsvermögen von 2000 SAPS betreiben zu können:

- Virtueller Server: Leistung=2000 SAPS,
- Virtuelle Storagepartition: Speicherplatz=200 GByte und
- VLAN: Übertragungsrate=1000 MBit/s.

Im folgenden Abschnitt werden die definierten Stammdaten verwendet, um eine Produktionsplanung durchzuführen.

3.2.3 Durchführung der Prozessfertigung

Im vorigen Abschnitt wurden die Stammdaten der Prozessfertigung definiert. Auf Basis dieser Stammdaten wird die Produktionsplanung und -steuerung durchgeführt. Im Beispiel bestellt ein Kunde ein SAP BW als IT-Dienstleistung für den Zeitraum von fünf Tagen. Das System soll dem Kunden ab dem 17.3.2009 zur Verfügung stehen.

Terminauftrag anlegen

Im Modul Vertrieb (SD) wird ein Terminauftrag angelegt, um die IT-Dienstleistung zu bestellen. Terminaufträge werden im Vertrieb verwendet, um die Lieferung zu einem bestimmten Termin zu veranlassen (vgl. Maassen et al., 2006, S. 431). Die Produktion des Materials *BW-2000-SAPS* entspricht der Erbringung der IT-Dienstleistung. Der Liefertermin im Terminauftrag wird verwendet, um den Zeitpunkt zu spezifizieren, an dem die Erbringung

der IT-Dienstleistung abgeschlossen ist. Die Auftragsmenge entspricht dem Zeitraum der Erbringung der IT-Dienstleistung und wird genau für einen Kunden produziert. Der Terminauftrag soll im Rechenzentrum in Kundeneinzelfertigung (vgl. Benz und Höflinger, 2008, S. 161 f.) mit folgenden Auftragsdaten ausgeführt werden:

- Auftragsart: Terminauftrag,
- Material: BW-2000-SAPS,
- Werk: 1000,
- Lieferdatum: 22.3.2009,
- Auftragsmenge: 5 TAG und
- Bedarfsart: Kundeneinzelfertigung.

Planauftrag erzeugen

Im Rahmen der Bedarfsplanung wird der Terminauftrag in einen Planauftrag überführt. Ein Planauftrag ist das Ergebnis der Bedarfsplanung und stellt eine Anforderung zur Beschaffung eines bestimmten Materials zu einem festgelegten Termin dar. Aus der Eigenfertigungszeit im Materialstammsatz werden durch Rückwärtsterminierung Eckstart- und Endtermin ermittelt (vgl. SAP, 2007). Der Planauftrag beinhaltet die Beschaffung einer Menge von fünf Tagen des Materials *BW-2000-SAPS* zum 22.3.2009.

Prozessauftrag erzeugen

Das Material *BW-2000-SAPS* soll in Eigenfertigung produziert werden. Ein Prozessauftrag wird für die Herstellung einer bestimmten Menge eines Materials zu einem bestimmten Termin verwendet. Der Prozessauftrag wird auf Basis des Planungsrezepts aus dem Planauftrag erzeugt (vgl. SAP, 2007).

Aus dem Planauftrag wird ein Prozessauftrag für die Produktion einer Menge von fünf Tagen des Materials *BW-2000-SAPS* zum 22.3.2009 erstellt.

Terminierung des Prozessauftrags

Ausgehend vom Eckendtermin wird der Prozessauftrag rückwärtsterminiert. Der späteste Starttermin der Phasen zum Betrieb des SAP BW ist der 17.3.2009 um 0:00 Uhr. Zu diesem Termin wird dem Kunden die Dienstleistung zur Verfügung gestellt. Der Betrieb des SAP BW wird am 21.3.2009 um 24:00 Uhr beendet. Die einzelnen Phasen der Bereitstellung werden entsprechend dem Planungsrezept vor dem Betrieb ausgeführt. Tabelle A.2 im Anhang zeigt die wesentlichen Daten der Terminierung des Prozessauftrags.

Ressourcenauswahl

Im Planungsrezept werden die virtualisierten Betriebsmittel durch Planressourcen beschrieben. Im Prozessauftrag werden diese Planressourcen in den entsprechenden Vorgängen durch tatsächliche Ressourcen ersetzt. Die Ressourcenauswahl wird anhand der Ressourcenauswahlbedingungen des Planungsrezepts durchgeführt. Zur Auswahl werden die Ressourcen angeboten, deren Klassifizierungsmerkmale den Ressourcenauswahlbedingungen entsprechen. Die Ressourcenauswahl wird manuell durchgeführt. Die Planressourcen werden durch die Ressourcen *SE-00002*, *ST-00002* und *VL-00001* ersetzt. Tabelle A.2 im Anhang zeigt das Ergebnis der Ressourcenauswahl im Prozessauftrag. Ansätze für eine automatisierte Entscheidungsfindung werden in Kapitel 6 aufgezeigt.

Kapazitätsabgleich

Der Kapazitätsabgleich ist Teil der Kapazitätsplanung. Ziel des Kapazitätsabgleichs ist die Verbesserung der Kapazitätsauslastung, der Termin-

treue und der Durchlaufzeiten. Wesentliche Aufgaben des Kapazitätsabgleichs sind die geeignete Auswahl und Belegung von Ressourcen sowie der Ausgleich von Über- und Unterbelastungen. Der Kapazitätsabgleich wird in grafischen oder tabellarischen Plantafeln durchgeführt (vgl. SAP, 2007).

Auftragsfreigabe

Vor der Durchführung der eigentlichen Produktion der IT-Dienstleistung muss der Prozessauftrag zur Fertigung freigegeben werden. Voraussetzung für die Auftragsfreigabe ist die Substitution aller Planressourcen durch tatsächliche Ressourcen. Die Auftragsfreigabe ist manuell durchzuführen, da die Ressourcenauswahl eine automatische Auftragsfreigabe verhindert. Mit der Auftragsfreigabe erfolgt die Übergabe des Prozessauftrags an die Produktion. Nach der Auftragsfreigabe können die Arbeitspapiere gedruckt werden. Arbeitspapiere sind Dokumente, die für die Rückmeldung verwendet werden (vgl. Maassen et al., 2006, S. 366 f.).

Rückmeldung

Der Bearbeitungsstand eines Prozessauftrags wird durch Rückmeldungen aktualisiert. Voraussetzung für die Ausführung von Rückmeldungen ist die Auftragsfreigabe. Rückmeldungen geben Auskunft darüber, welche Gut- und Ausschussmengen produziert, wieviel Zeit dafür in Anspruch genommen und durch welche Ressourcen der Vorgang ausgeführt wurde. Die Kapazitäten der betroffenen Ressourcen werden entlastet und die entstandenen Istkosten verrechnet (vgl. Benz und Höflinger, 2008, S. 250 f.). Die Rückmeldung des Prozessauftrags wird auf Phasenebene vorgenommen.
Die Fertigungsart der Prozessfertigung ermöglicht die Anbindung von Prozesssteuerungssystemen. Durch den Einsatz von Prozesssteuerungssystemen können Rückmeldungen automatisiert über Prozessmeldungen durchgeführt werden (vgl. SAP, 2007).
Bei der Ausführung des Prozessauftrags werden Bewegungsdaten erzeugt.

Diese Bewegungsdaten ermöglichen eine qualitative Bewertung der Produktionsplanung und -steuerung.

3.3 Bewertung der Kapazitätswirtschaft

Aufgabe der Kapazitätswirtschaft ist die Sicherstellung einer den Anforderungen gerecht werdenden Bereitstellung und Überwachung der IT-Kapazitäten (vgl. Zarnekow et al., 2005, S. 49). Im folgenden Abschnitt soll ein Qualitätsmodell erstellt werden, das eine Bewertung ermöglicht, inwieweit die Kapazitätswirtschaft auf Ressourcenebene im Rechenzentrum diesen Anforderungen gerecht wird.

3.3.1 Qualität

Qualität ist ein Maßstab dafür, wie gut oder schlecht eine betrachtete Einheit die Forderung an ihre Beschaffenheit erfüllt (vgl. Geiger und Kotte, 2008, S. 67 ff.). Somit wird durch Qualität eine Relation zwischen einer realisierten und einer geforderten Beschaffenheit an eine Einheit dargestellt. Dabei bezeichnet eine Einheit das, was einzeln beschrieben und betrachtet werden kann (vgl. Geiger und Kotte, 2008, S. 61 ff.). Eine Einheit kann demnach materiell wie Produkte oder Personen oder immateriell wie Tätigkeiten sein. Die Beschaffenheit einer Einheit ist dabei die Gesamtheit der inhärenten Merkmale und ihrer Werte, die zur Einheit selbst gehören.
Bei der Bestimmung der Qualität einer Einheit kommen Qualitätsmodelle zum Einsatz. Sie legen die operationalen Qualitätsmerkmale für jede betrachtete Einheit fest (vgl. Heinrich et al., 2004, S. 547). Eine Zerlegung der Qualität in einzelne Merkmale ermöglicht eine Zuordnung von Kenngrößen, die quantifizierbar sind. Qualitätsanforderungen legen fest, welche

Qualitätsmerkmale im konkreten Fall als relevant erachtet werden und welche Ausprägungen diese Merkmale erreichen sollen. Zur Bewertung der Qualitätsmerkmale sind Qualitätsstufen zu definieren. Es ist festzulegen, welche Qualitätsstufen erreicht werden sollen (vgl. Balzert, 1998, S. 269 ff.). Eine Qualitätsstufe ist ein Wertebereich auf einer Skala, mit deren Hilfe die betrachtete Einheit entsprechend den festgelegten Qualitätsanforderungen klassifiziert wird (vgl. ISO/IEC 9126-1, 2001, S. 2).
Nominal-, Ordinal-, Intervall- und Verhältnisskala können als Skalentypen zur Abbildung der Qualitätsstufen eingesetzt werden. Zwischen diesen Skalentypen existiert eine hierarchische Ordnung (vgl. Rasch et al., 2006, S. 9 ff.). Die Nominalskala verfügt über das niedrigste Skalenniveau. Für die Werte auf einer Nominalskala gilt Exklusivität und Exhaustivität. Unterschiedliche Ausprägungen der Qualitätsmerkmale werden unterschiedlichen Qualitätsstufen zugeordnet. Es existiert eine Qualitätsstufe für jede beobachtete oder potentielle Ausprägung eines Qualitätsmerkmals.
Für die Ordinalskala gelten dieselben Eigenschaften wie für die Nominalskala. Zusätzlich repräsentieren die Qualitätsstufen Unterschiede einer bestimmten Größe in Bezug auf das Qualitätsmerkmal. Aus diesen Unterschieden lässt sich eine hierarchische Ordnung der Qualitätsstufen ableiten.
Die Intervallskala schließt die Eigenschaften der Ordinalskala ein. Weiterhin werden über gleich große Abstände der Qualitätsstufen gleich große Abstände der Qualitätsmerkmale abgebildet.
Neben den aufgeführten Eigenschaften wird bei einer Verhältnisskala ein Nullpunkt für ein Qualitätsmerkmal definiert. Dieser Nullpunkt definiert den Zustand, in dem die betrachtete Einheit über keine Qualität verfügt.
Zur Bewertung eines Qualitätsmerkmals kann es notwendig sein, eine Skalentransformation durchzuführen. Grundsätzlich kann ein höheres Skalenniveau in ein niedrigeres überführt werden (vgl. Weiß, 2005, S. 26 f.).
Um ein Qualitätsmerkmal einer Qualitätsstufe zuordnen zu können, muss es ermittelbar sein. Unter Ermittlung wird die Bestimmung eines oder mehrerer Werte einer oder mehrerer betrachteter Einheiten verstanden. Versuch, Test, Untersuchung und Messung sind Methoden zur Ermittlung eines Qualitätsmerkmals (vgl. Geiger und Kotte, 2008, S. 118). Die Qualitätsmerkmale

von Dienstleistungen lassen sich folgenden Dimensionen zuordnen (vgl. Donabedian, 1980, S. 83):

- Potentialqualität,
- Prozessqualität und
- Ergebnisqualität.

Die zur Erbringung der Dienstleistung notwendigen Leistungsvoraussetzungen werden durch die Potentialqualität bewertet. Die Qualität der Aktivitäten während der Leistungserstellung wird durch die Prozessqualität abgebildet. Die Übereinstimmung von versprochener Leistung und dem Dienstleistungsergebnis wird durch die Ergebnisqualität beschrieben (vgl. Ziemeck, 2006, S. 24). Bei Dienstleistungen finden Produktion und Konsumtion simultan statt (vgl. Bruhn, 2006, S. 21). Prozess- und Ergebnisqualität beziehen sich somit beide auf die Phase der Leistungserstellung, repräsentieren jedoch verschiedene Sichten auf diesen Prozess. Die Prozessqualität beschreibt die Erfüllung der Forderungen an die Beschaffenheit der IT-Dienstleistung aus Sicht des Anbieters. Aus Sicht des Kunden ist die Ergebnisqualität die relevante Qualitätsdimension.
Die konkreten Anforderungen an die Dienstleistungsqualität aus Kunden- und Anbietersicht werden im Rahmen der Qualitätsplanung festgelegt (vgl. Bruhn, 2006, S. 253). Im folgenden Abschnitt werden die Qualitätsmerkmale von IT-Dienstleistungen identifiziert, die im Rahmen der Kapazitätswirtschaft vom Rechenzentrum beeinflussbar sind. Diese Qualitätsmerkmale sollen eine Bewertung der Qualität der Kapazitätswirtschaft im Rechenzentrum ermöglichen.

3.3.2 Qualität der Kapazitätswirtschaft

Die vom Rechenzentrum erbrachte IT-Dienstleistung ist die Einheit, deren Qualität aus Kunden- und Anbietersicht bewertet werden soll. Es

werden ausschließlich Qualitätsmerkmale betrachtet, die im Rahmen der Kapazitätswirtschaft beeinflusst werden können und messbar sind. Die Qualitätsmerkmale der IT-Dienstleistung als Ausbringung der Produktionsfunktion werden aus Kunden- und Anbietersicht durch die Verfügbarkeit der geforderten Kapazitäten bestimmt. Die Verfügbarkeit als Qualitätsmerkmal lässt sich in drei Hauptmerkmale untergliedern:

- Bereitstellung zum vereinbarten Termin,
- zeitliche Verfügbarkeit und
- technisches Leistungsvermögen.

Die Qualitätsstufen der Merkmale werden in Service Level Vereinbarungen zwischen Anbieter und Kunden definiert. Service Level legen dabei fest, welche Qualitätsstufen erreicht werden sollen. Service Level Vereinbarungen sind Bestandteil des Service Level Managements der IT Infrastructure Library (vgl. OGC, 2007a, S. 66).
Mit dem Kunden wird ein Termin vereinbart, ab dem die Erbringung der IT-Dienstleistung beginnt. Rechtzeitig zu diesem Termin müssen alle Arbeitsschritte zur Bereitstellung beendet sein. Das Qualitätsmerkmal der Bereitstellung zum vereinbarten Termin wird über die zeitliche Differenz zwischen geplanter und tatsächlicher Beendigung der Bereitstellung bestimmt. Eine Überschreitung des Bereitstellungstermins mindert die Qualität der IT-Dienstleistung aus Kunden- und Anbietersicht. Werden die benötigten Potentialfaktoren zu früh bereitgestellt, führt dies aus Anbietersicht zu Ressourcenineffizienz und mindert die Qualität. Aus Kundensicht kann eine zu frühe Bereitstellung die Qualität jedoch erhöhen, da die IT-Dienstleistung schon früher als geplant zur Verfügung steht. Benötigt der Kunde die IT-Dienstleistung erst zum vereinbarten Termin, bleibt die Qualität unverändert. Die Datenerhebung kann anhand des Fertigungsauftrags über die zeitliche Abweichung der Arbeitsschritte zur Bereitstellung von den Planwerten durchgeführt werden. Die Bereitstellung zum vereinbarten Termin

ist die zur Erbringung der Dienstleistung notwendige Leistungsvoraussetzung. Das Qualitätsmerkmal beschreibt somit die kapazitätswirtschaftlich relevante Potentialqualität der IT-Dienstleistung. Es werden die Effizienz des Ressourceneinsatzes an Potentialfaktoren und die Zuverlässigkeit der Bereitstellung bewertet.
Für den Zeitraum der Leistungserstellung werden dem Kunden Potentialfaktoren zur Verfügung gestellt. Der Ausfall von Potentialfaktoren mindert die Qualität der IT-Dienstleistung aus Kunden- und Anbietersicht. Dem Kunden steht die IT-Dienstleistung für den Zeitraum des Ausfalls nicht zur Verfügung. Für den Anbieter entsteht zusätzlicher Aufwand bei der Wiederherstellung der Verfügbarkeit. Der Ausfall eines Potentialfaktors mindert die Anzahl der vom Konsumenten beziehbaren Zeiteinheiten und damit auch die Produktionsmenge des Fertigungsprodukts IT-Dienstleistung. Die Kennzahl Mean Time Between Failures (MTBF) beschreibt den mittleren Abstand zwischen zwei Ausfällen. Die mittlere Dauer der Reparatur wird durch die Kennzahl Mean Time To Repair (MTTR) ermittelt (vgl. Márquez, 2007, S. 266 f.). Das Qualitätsmerkmal der prozentualen zeitlichen Verfügbarkeit lässt sich aus den Kennzahlen MTBF und MTTR wie folgt errechnen (vgl. Adam, 1989, S. 179):

$$\textit{Zeitliche Verfügbarkeit} = \frac{MTBF}{MTTR + MTBF} \cdot 100 \tag{3.1}$$

Die Datenerhebung kann anhand der Protokollierung der Ausfallzeiten durchgeführt werden. Aus Anbietersicht beschreibt die zeitliche Verfügbarkeit, ob die Aktivitäten während der Leistungserstellung entsprechend den Anforderungen ausgeführt werden. Die zeitliche Verfügbarkeit ist aus Anbietersicht ein Qualitätsmerkmal der Prozessqualität. Aus Kundensicht kann anhand der zeitlichen Verfügbarkeit die Übereinstimmung von versprochener Leistung und Dienstleistungsergebnis beschrieben werden. Die zeitliche Verfügbarkeit ist aus Kundensicht ein Qualitätsmerkmal der Ergebnisqualität. Es beschreibt die Zuverlässigkeit der IT-Dienstleistung aus Kunden-

und Anbietersicht.
Mit dem Kunden wird vereinbart, über welches technische Leistungsvermögen die einzelnen Potentialfaktoren verfügen sollen. Der Einsatz von Virtualisierungstechniken kann beispielsweise zu einem Überangebot oder einem Defizit an technischem Leistungsvermögen führen. Eine Unterschreitung des vereinbarten technischen Leistungsvermögens mindert die Qualität der IT-Dienstleistung aus Kunden- und Anbietersicht. Stellt ein Überangebot an technischem Leistungsvermögen für den Kunden einen Mehrwert dar, führt dies aus Kundensicht zu einer Erhöhung der Qualität. Aus Anbietersicht führt ein Überangebot zu Ressourcenineffizienz und mindert die Qualität. Das Qualitätsmerkmal des technischen Leistungsvermögens wird anhand des tatsächlichen technischen Leistungsvermögens eines Potentialfaktors ermittelt. Die Bestimmung dieses Qualitätsmerkmals erfolgt durch Messung des technischen Leistungsvermögens eines Potentialfaktors in definierten Zeitintervallen und dem Vergleich mit dem Planwert. Die Datenerhebung kann durch Monitoring (vgl. Mansouri-Samani und Sloman, 1992, S. 1 ff.) mittels geeigneter Werkzeuge durchgeführt werden. Die technische Verfügbarkeit während der Leistungserstellung beschreibt die Ausführung der Aktivitäten entsprechend den Anforderungen an die IT-Dienstleistung. Die technische Verfügbarkeit ist somit aus Anbietersicht ein weiteres Qualitätsmerkmal der Prozessqualität. Es beschreibt die Effizienz des Ressourceneinsatzes. Die Übereinstimmung von versprochener Leistung und Dienstleistungsergebnis lässt sich aus Kundensicht anhand der technischen Verfügbarkeit beschreiben. Aus Kundensicht ist die technische Verfügbarkeit ein weiteres Qualitätsmerkmal der Ergebnisqualität. Es beschreibt die Zuverlässigkeit der IT-Dienstleistung.
In Tabelle 3.2 sind die Zuordnungen zwischen Qualitätsmerkmalen, Qualitätsdimensionen und Qualitätssichten zusammengefasst. Die beschriebenen Qualitätsmerkmale stehen in einer Abhängigkeitsbeziehung. Die Überschreitung des Bereitstellungstermins führt auch zu einer Reduzierung der zeitlichen Verfügbarkeit. Steht eine IT-Dienstleistung dem Kunden zeitlich nicht zur Verfügung, ist auch keine technische Verfügbarkeit gegeben. Aus Kundensicht kann die Minderung der Qualität eines Potentialfaktors

	Qualitätsdimensionen		
Qualitätsmerkmale	**Potentialqualität**	**Prozessqualität**	**Ergebnisqualität**
Bereitstellung	Ressourceneffizienz *(Anbietersicht)*, Zuverlässigkeit *(Kundensicht)*	-	-
Zeitliche Verfügbarkeit	-	Zuverlässigkeit *(Anbietersicht)*	Zuverlässigkeit *(Kundensicht)*
Technisches Leistungsvermögen	-	Ressourceneffizienz *(Anbietersicht)*	Zuverlässigkeit *(Kundensicht)*

Tabelle 3.2: Zuordnungen zwischen Qualitätsmerkmalen, Qualitätsdimensionen und Qualitätssichten

zur Minderung der Qualität weiterer Potentialfaktoren führen. So entsteht beispielsweise aus der Nichtverfügbarkeit der Kommunikationskapazität auch eine Nichtverfügbarkeit der Prozessorkapazität, obwohl aus Anbietersicht technisch genügend Prozessorkapazität bereitgestellt wird.
Ein Überangebot an technischer Verfügbarkeit verdeutlicht, dass Qualitätsmerkmale aus Kunden- und Anbietersicht gegensätzlich bewertet werden können. Aus Kundensicht kann ein Überangebot als Mehrwert positiv bewertet werden, während aus Anbietersicht dies als Ressourcenineffizienz negativ bewertet wird.
Zur Bewertung der Qualitätsmerkmale soll eine einheitliche Skala verwendet werden. Die Verwendung einheitlicher Skalenstufen vereinfacht die Qualitätsanalyse. Qualitätsdefizite einzelner Qualitätsmerkmale können so eindeutig identifiziert werden. Verschiedene Qualitätsmerkmale sind bei Ver-

wendung einer einheitlichen Skala vergleichbar. Folgenden Qualitätsstufen lassen sich alle Qualitätsmerkmale zuordnen:

- **Qualität verletzt SLA:**
  Die Qualität erfüllt nicht die zwischen Anbieter und Kunde vereinbarten Service-Level.
- **Schlechte Qualität:**
  Die Qualität erfüllt die vereinbarten Service-Level, wird jedoch aus einem bestimmten Grund negativ bewertet.
- **Gute Qualität:**
  Die Qualität erfüllt die vereinbarten Service-Level und wird aus keinem Grund negativ bewertet.
- **Sehr gute Qualität:**
  Die vereinbarten Service-Level werden deutlich übererfüllt und aus keinem Grund negativ bewertet. Die Übererfüllung stellt einen Mehrwert dar.

Die Qualitätsstufen werden auf einer Ordinalskala abgebildet. Die gemessenen quantitativen Qualitätsmerkmale werden hierbei durch eine Skalentransformation in nichtquantitative ordinale Qualitätsstufen überführt.
Die Qualität der Bereitstellung verletzt die Service-Level, wenn der vereinbarte Bereitstellungstermin überschritten wird. Die Qualität wird als gut bewertet, wenn der Bereitstellungstermin eingehalten wird. Ist die Bereitstellung deutlich vor dem vereinbarten Termin abgeschlossen, wird das Qualitätsmerkmal aus Anbietersicht als schlecht bewertet. Der Grund hierfür ist Ressourcenineffizienz. Aus Kundensicht wird eine Bereitstellung deutlich vor dem vereinbarten Zeitpunkt als gut bewertet. Falls der Kunde das System bereits zu diesem Zeitpunkt nutzen kann, wird die Bereitstellung als sehr gut bewertet. Zur Bewertung ist ein Zeitintervall zu definieren, ab dem der Bereitstellungstermin als deutlich unterschritten gilt. Abbildung 3.8 veranschaulicht die Qualitätsstufen des Qualitätsmerkmals Bereitstellung aus Anbieter- und Kundensicht. Im Beispiel wird die Bereitstellung auf

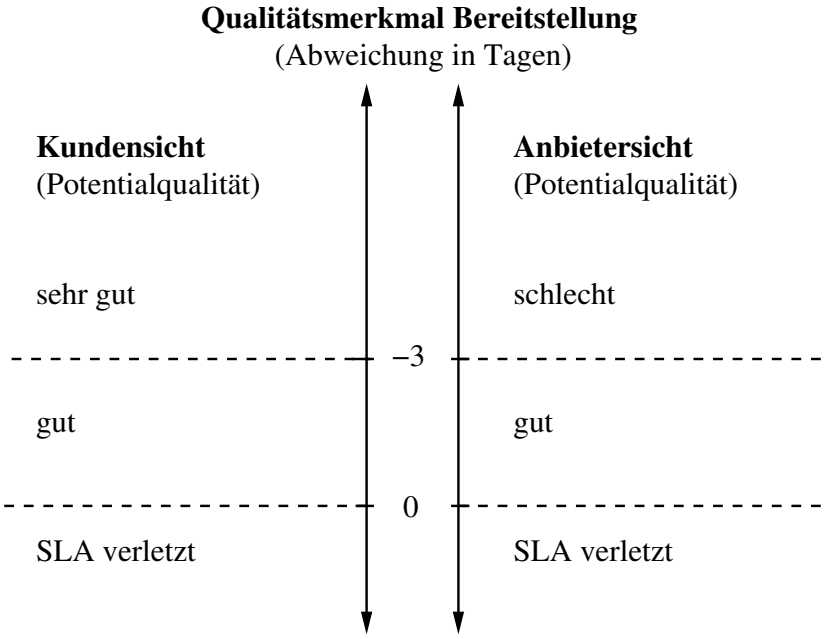

Abbildung 3.8: Qualitätsmerkmal Bereitstellung aus Anbieter- und Kundensicht

Tagesebene terminiert. Das Qualitätsmerkmal ist die Differenz zwischen geplanter und tatsächlicher Bereitstellung in Tagen. Eine Unterschreitung des Bereitstellungstermins von drei Tagen wurde als deutliche Abweichung vom vereinbarten Termin definiert.

Die Qualität der zeitlichen Verfügbarkeit verletzt die Service-Level, wenn die vereinbarte Verfügbarkeit der IT-Dienstleistung unterschritten wird. Die zeitliche Verfügbarkeit wird üblicherweise in Prozent angegeben und bezieht sich auf den Zeitraum der Leistungserstellung. Die Einhaltung der vereinbarten zeitlichen Verfügbarkeit wird aus Anbieter- und Kundensicht gut bewertet. Eine deutliche Übererfüllung der vereinbarten zeitlichen Verfügbarkeit wird aus Anbieter- und Kundensicht sehr gut bewertet. Zur Bewertung ist ein Prozentwert zu definieren, ab dem die vereinbarte zeitliche Verfügbarkeit als deutlich übererfüllt gilt.

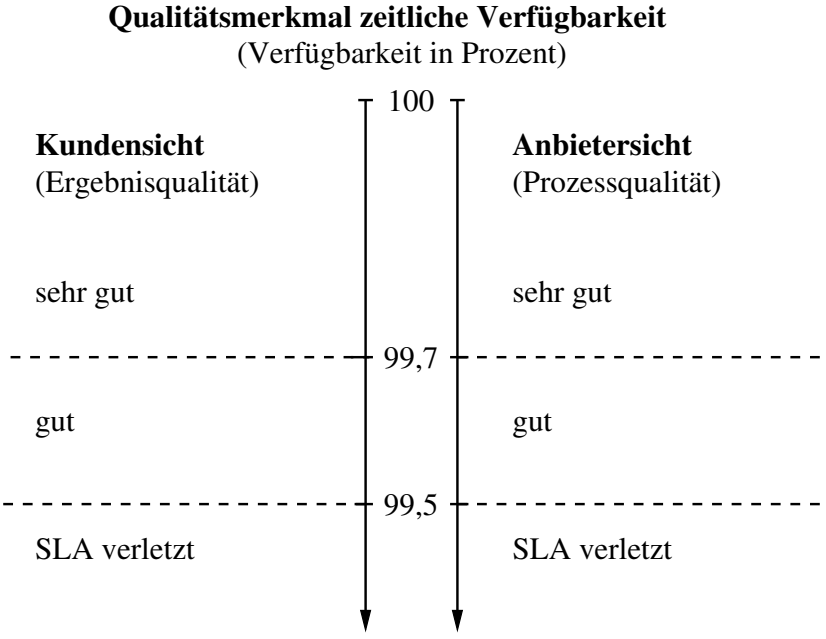

Abbildung 3.9: Qualitätsmerkmal zeitliche Verfügbarkeit aus Anbieter- und Kundensicht

Abbildung 3.9 veranschaulicht die Qualitätsstufen des Qualitätsmerkmals zeitliche Verfügbarkeit aus Anbieter- und Kundensicht. Im Beispiel wurde eine zeitliche Verfügbarkeit von 99,5 % mit dem Kunden vereinbart. Eine tatsächliche zeitliche Verfügbarkeit von 99,7 % wurde als deutliche Übererfüllung der vereinbarten zeitlichen Verfügbarkeit definiert.
Die Qualität der technischen Verfügbarkeit verletzt die Service-Level, wenn das technische Leistungsvermögen eines Potentialfaktors den mit dem Kunden vereinbarten Wert unterschreitet. Die technische Verfügbarkeit jedes Potentialfaktors stellt hierbei ein eigenständiges Qualitätsmerkmal dar. Die Einhaltung des technischen Leistungsvermögens wird aus Anbieter- und Kundensicht als gut bewertet. Steht dem Kunden deutlich mehr technisches Leistungsvermögen zur Verfügung als vereinbart und stellt dieses zusätzliche Leistungsvermögen einen Mehrwert dar, wird dies aus Kundensicht als sehr gut bewertet. Aus Anbietersicht liegt in diesem Fall eine Ressource-

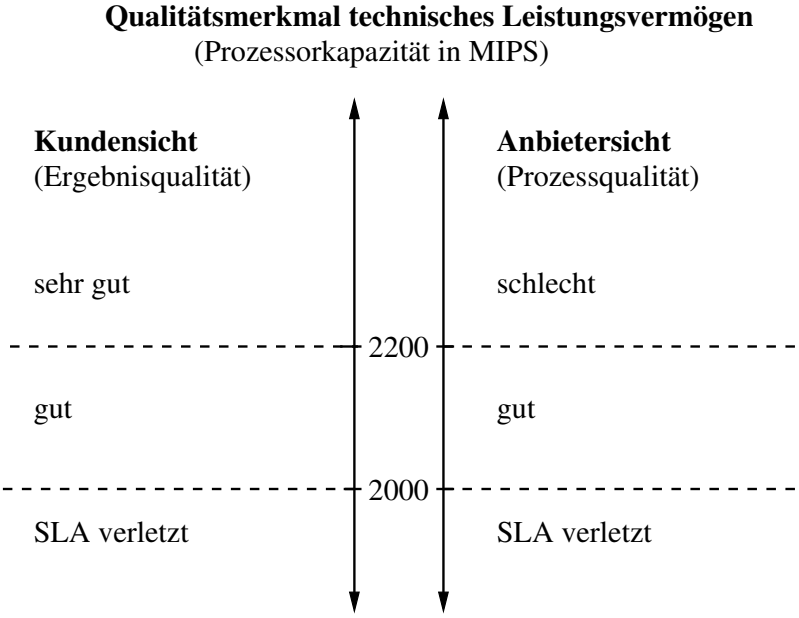

Abbildung 3.10: Qualitätsmerkmal technisches Leistungsvermögen aus Anbieter- und Kundensicht

nineffizienz vor, die zu einer schlechten Bewertung der Qualität führt. Zur Bewertung ist für jeden Potentialfaktor ein Wert festzulegen, ab dem deutlich mehr technisches Leistungsvermögen als vereinbart zur Verfügung steht. Abbildung 3.10 veranschaulicht die Qualitätsstufen des Qualitätsmerkmals technische Verfügbarkeit aus Anbieter- und Kundensicht. Im Beispiel wird die Abweichung der Prozessorkapazität vom vereinbarten technischen Leistungsvermögen bewertet. Eine Abweichung von 200 MIPS wird als deutliche Übererfüllung des technischen Leistungsvermögens definiert.

3.4 Zusammenfassung

Der Forschungsgegenstand IT-Dienstleistung lässt sich in einem System zur Produktionsplanung und -steuerung abbilden. Die Hypothese wurde durch

ein Referenzmodell bestätigt. Das Referenzmodell wurde anhand eines Prototypen erfolgreich evaluiert. Ein neues Anwendungsgebiet der Wirtschaft für den Einsatz von Systemen zur Produktionsplanung und -steuerung konnte erschlossen werden.
Bei der Bestätigung der Hypothese wurde von der Bedingung ausgegangen, dass der Einsatz von Virtualisierungstechniken aus Sicht der Kapazitätswirtschaft sinnvoll ist. Im folgenden Kapitel wird die Gültigkeit dieser Bedingung als eigene Forschungsfrage untersucht.
Es wurde festgestellt, dass die Methode der Produktionsprogrammplanung nicht ausreicht, um die Planung des Kapazitätsbedarfs an Potentialfaktoren im Rechenzentrum durchzuführen. Die Planung des Kapazitätsbedarfs wird als eigene Forschungsfrage in Kapitel 5 untersucht.
Die Evaluation des Referenzmodells durch einen Prototypen führte zu der Erkenntnis, dass die Zuordnung der Betriebsmittel durch standardisierte Methoden der Produktionsplanung und -steuerung nicht automatisiert durchgeführt werden kann. Die Automatisierung der Betriebsmittelzuordnung wird als eigene Forschungsfrage in Kapitel 6 untersucht.

Kapitel 4

Planung des Kapazitätsangebots

Im vorigen Kapitel wurde das Rechenzentrum als Produktionsbetrieb beschrieben. Es wurde davon ausgegangen, dass der Einsatz von Virtualisierungstechniken aus Sicht der Kapazitätswirtschaft sinnvoll ist. In diesem Kapitel wird diese Behauptung als eigene Forschungsfrage untersucht. Folgende Hypothese soll überprüft werden:

> Der Einsatz von Virtualisierungstechniken führt zu Einsparungen des Kapazitätsangebots.

Bei der Bestätigung der Hypothese wird davon ausgegangen, dass die Auslastung der virtuellen Server einer Verteilungsfunktion unterliegt. Basierend auf der Warteschlangentheorie wird ein mathematisch-formales Modell zur Planung des Kapazitätsangebots für physische und virtuelle Server entwickelt. Es wird untersucht, welchen Einfluss die Virtualisierung von Servern auf die Kapazitätsangebotsplanung hat.

4.1 Bestimmung der kritischen Serverauslastung

Der Betreiber eines Rechenzentrums kann das Kapazitätsangebot eines Rechnersystems nicht vollständig in Anspruch nehmen, da dies zu einem inakzeptablen Antwortzeitverhalten führt. Die Inanspruchnahme des Kapazitätsangebots wird als Serverauslastung bezeichnet. Unter Verwendung der Warteschlangentheorie lässt sich im Modell zeigen, ab welchem Grad der Serverauslastung ein Rechnersystem nicht mehr den Kundenanforderungen genügt. Eine Warteschlange kann als zeitkontinuierliche Markovkette modelliert werden.

4.1.1 Markovketten

Markovketten sind stochastische Prozesse, die folgende Eigenschaft erfüllen (vgl. Zikun und Xiangqun, 1992, S. 20):

$$P(X_{t_{n+1}} = i_{n+1} | X_{t_0} = i_0, ..., X_{t_n} = i_n) = P(X_{t_{n+1}} = i_{n+1} | X_{t_n} = i_n) \quad (4.1)$$

Der Zeitpunkt t_n kann als Gegenwart interpretiert werden. Die Zeitpunkte $t_0, ..., t_{n-1}$ repräsentieren die Vergangenheit. Für den zukünftigen Zeitpunkt t_{n+1} ist die bedingte Wahrscheinlichkeit für das Eintreten des Ereignisses $X_{t_{n+1}}$ nur von dem gegenwärtigen Ereignis X_{t_n} abhängig. Die Folge von Ereignissen $X_{t_0}, ..., X_{t_n}$ ist für das Eintreten von $X_{t_{n+1}}$ irrelevant (vgl. Kolmogoroff, 1936, S. 155).
Formel 4.1 beschreibt die Übergangswahrscheinlichkeit p_{ij} von einem Zustand i zu einem Zustand j innerhalb eines vorgegebenen Zeitintervalls u, v (vgl. Bolch, 1998, S. 49):

$$p_{ij}(u, v) = P(X_v = j | X_u = i) \quad (4.2)$$

Hängt die Übergangswahrscheinlichkeit p_{ij} nur von der Differenz der Zeitpunkte u und v und nicht von den Zeitpunkten selber ab, wird die Markovkette als zeithomogen bezeichnet. Für den zeithomogenen Fall gilt:

$$p_{ij}(t, t + \Delta t) = p_{ij}(0, \Delta t) \tag{4.3}$$

Wird die Menge aller Übergangswahrscheinlichkeiten p in einer Matrix P abgebildet, gilt im zeithomogenen Fall:

$$P(t, t + \Delta t) = P(0, \Delta t) \tag{4.4}$$

Die Wahrscheinlichkeit, dass sich der Server zum Zeitpunkt u im Zustand i befindet beträgt $\pi_i(u)$. Die Wahrscheinlichkeit $\pi_j(v)$, dass sich der Server zum Zeitpunkt v im Zustand j befindet, errechnet sich aus der Summe aller Übergangswahrscheinlichkeiten von i nach j unter der Vorbedingung $\pi_i(u)$:

$$\pi_j(v) \quad = \quad \sum_{i \in S} p_{ij}(u, v)\pi_i(u) \tag{4.5}$$

Die Menge der Zustandswahrscheinlichkeiten $\pi_i(u)$ kann in einem Vektor $\vec{\pi}(u) = (\pi_0(u), \pi_1(u), \pi_1(u), ...)$ zusammengefasst werden. In Vektorschreibweise gilt für die Zustandwahrscheinlichkeiten $\vec{\pi}(v)$:

$$\vec{\pi}(v) \quad = \quad \vec{\pi}(u)P(u, v) \tag{4.6}$$

Im zeithomogenen Fall ist die Zustandswahrscheinlichkeit $\vec{\pi}$ nur von der Zeitdifferenz Δt abhängig:

$$\vec{\pi}(t + \Delta t) \quad = \quad \vec{\pi}(t)P(0, \Delta t) \tag{4.7}$$

Der Differentialquotient aus der Zustandswahrscheinlichkeit $\vec{\pi}$ und der Zeit t lässt sich im zeithomogenen Fall wie folgt herleiten:

$$\begin{aligned}
\lim_{\Delta t \to 0} \vec{\pi}(t + \Delta t) &= \lim_{\Delta t \to 0} \vec{\pi}(t) P(0, \Delta t) \\
\lim_{\Delta t \to 0} \vec{\pi}(t + \Delta t) - \vec{\pi}(t) &= \lim_{\Delta t \to 0} \vec{\pi}(t) P(0, \Delta t) - \vec{\pi}(t) \\
\lim_{\Delta t \to 0} \frac{\vec{\pi}(t + \Delta t) - \vec{\pi}(t)}{\Delta t} &= \lim_{\Delta t \to 0} \frac{\vec{\pi}(t) P(0, \Delta t) - \vec{\pi}(t)}{\Delta t} \\
\frac{\mathrm{d}\vec{\pi}(t)}{\mathrm{d}t} &= \lim_{\Delta t \to 0} \frac{\vec{\pi}(t) P(0, \Delta t) - \vec{\pi}(t)}{\Delta t} \\
\frac{\mathrm{d}\vec{\pi}(t)}{\mathrm{d}t} &= \vec{\pi}(t) \lim_{\Delta t \to 0} \frac{P(0, \Delta t) - E}{\Delta t}
\end{aligned} \tag{4.8}$$

Die Matrix E repräsentiert die Einheitsmatrix. Im Folgenden wird eine Matrix als Intensitätsmatrix Q definiert, wenn sie im zeithomogenen Fall folgende Eigenschaften erfüllt:

$$\begin{aligned}
Q &= \lim_{\Delta t \to 0} \frac{P(0, \Delta t) - E}{\Delta t} \\
q_{ij} &= \lim_{\Delta t \to 0} \frac{p_{ij}(0, \Delta t)}{\Delta t} \qquad \text{für} \qquad i \neq j \\
q_{ii} &= \lim_{\Delta t \to 0} \frac{p_{ii}(0, \Delta t) - 1}{\Delta t}
\end{aligned} \tag{4.9}$$

Der Differentialquotient aus der Zustandswahrscheinlichkeit $\vec{\pi}$(t) und der Zeit t lautet somit im zeithomogenen Fall:

$$\begin{aligned}
\frac{\mathrm{d}\vec{\pi}(t)}{\mathrm{d}t} &= \vec{\pi}(t) Q \\
\frac{\mathrm{d}\pi_j(t)}{\mathrm{d}t} &= \sum_{i \in S} q_{ij} \pi_i(t)
\end{aligned} \tag{4.10}$$

Die Variable q_{ij} beschreibt die Übergangsrate von einem Zustand i in einen Zustand j. Für die Zustandsübergänge wird eine Poissonverteilung mit einer Ereignisrate λ unterstellt. Für diese Verteilung ist die Wahrscheinlichkeit, dass ein Ereignis innerhalb eines kleinen Zeitintervalls Δt eintritt annähernd $\lambda \Delta t$ (vgl. Fahrmeir, 2004, S. 350). Für Formel 4.9 folgt:

$$\begin{aligned} \lim_{\Delta t \to 0} p_{ij}(0, \Delta t) &= \lambda_{ij} \Delta t \\ q_{ij} &= \lambda_{ij} \end{aligned} \tag{4.11}$$

Die Wahrscheinlichkeit, dass sich die Markovkette in einem stabilen Zustand befindet, wird als Gleichgewichtswahrscheinlichkeit bezeichnet. Die Gleichgewichtswahrscheinlichkeit π_i, dass sich die Markovkette im stabilen Zustand i befindet, entspricht der Konvergenz der Zustandswahrscheinlichkeit $\pi_i(t)$ im Zeitverlauf:

$$\pi_i = \lim_{t \to \infty} \pi_i(t) \tag{4.12}$$

Aus dem Differentialquotient der Zustandswahrscheinlichkeit $\pi_j(t)$ und der Zeit t (Formel 4.10) lässt sich der Differentialquotient aus der Gleichgewichtswahrscheinlichkeit π_j und der Zeit t herleiten:

$$\begin{aligned} \lim_{t \to \infty} \frac{\mathrm{d}\pi_j(t)}{\mathrm{d}t} &= \lim_{t \to \infty} \sum_{i \in S} q_{ij} \pi_i(t) \\ \frac{\mathrm{d} \lim\limits_{t \to \infty} \pi_j(t)}{\mathrm{d}t} &= \sum_{i \in S} q_{ij} \lim_{t \to \infty} \pi_i(t) \\ \frac{\mathrm{d}\pi_j}{\mathrm{d}t} &= \sum_{i \in S} q_{ij} \pi_i \end{aligned} \tag{4.13}$$

Befindet sich die Markovkette im Gleichgewicht, wird ihr stationärer Zustand im weiteren Zeitverlauf nicht mehr geändert. Die Gleichgewichtswahrscheinlichkeit ist somit zeitunabhängig:

$$\frac{\mathrm{d}\pi_i}{\mathrm{d}t} = 0 \tag{4.14}$$

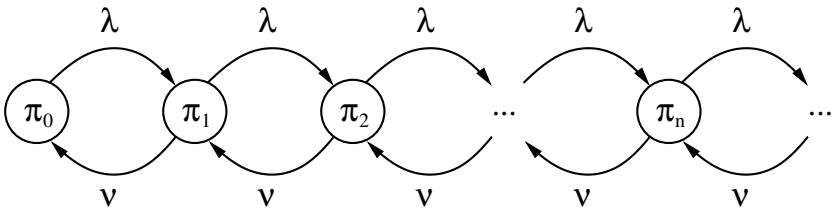

Abbildung 4.1: Warteschlange als Markov-Kette (nach Bolch, 1998, S. 105)

Aus der Zeitunabhängigkeit lässt sich folgende Beziehung zwischen der Intensitätsmatrix Q und der Gleichgewichtswahrscheinlichkeit $\vec{\pi}$ herleiten:

$$\begin{aligned} 0 &= \sum_{i \in S} q_{ij} \pi_i \\ \vec{0} &= \vec{\pi} Q \end{aligned} \tag{4.15}$$

4.1.2 Warteschlangen als Markovketten

Im Modell wird der Server als Wartschlange abgebildet. An den Server werden Anfragen gestellt, die durch Ereignisse repräsentiert werden. Diese Ereignisse unterliegen einer Poissonverteilung. Die Ankunftsrate λ definiert die Anzahl von Anfragen an den Server, die in einem bestimmten Zeitintervall eintreffen. Mit dem Eintreffen einer neuen Anfrage ändert sich der Zustand i des Servers in den Zustand $i + 1$. Die Zustandswahrscheinlichkeit $\pi_i(t)$ repräsentiert die Wahrscheinlichkeit, dass sich i Anfragen zum Zeitpunkt t im Server befinden. Der Server arbeitet diese Anfragen ab. Ist eine Anfrage abgearbeitet, wird dies durch ein Ereignis abgebildet. Dieses Ereignis unterliegt ebenfalls einer Poissonverteilung. Die Bearbeitungsrate ν definiert die Anzahl von Anfragen, die vom Server in einem bestimmten Zeitintervall abgearbeitet werden. Die Ankunfts- und Bearbeitungsraten λ und ν sind konstant und unabhängig von der Anzahl an Anfragen. In Abbildung 4.1 ist dieser Zusammenhang grafisch

dargestellt. Die Intensitätsmatrix Q für den beschriebenen Server lautet:

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ \nu & -(\lambda+\nu) & \lambda & 0 & \dots \\ 0 & \nu & -(\lambda+\nu) & \lambda & \dots \\ 0 & 0 & \nu & -(\lambda+\nu) & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{4.16}$$

Eine Warteschlange ist eine spezielle zeitkontinuierliche Markovkette, bei der Zustandsübergänge nur zwischen benachbarten Zuständen zulässig sind. Aus dem Zustand, dass sich im Server Null Anfragen befinden, ist nur ein Übergang in den Zustand einer Anfrage zulässig. Die Anzahl an Anfragen, die sich im Server befinden, kann im Modell beliebig groß werden. Für die Warteschlange wird Zeithomogenität unterstellt. Da aus dem Zustand, in dem sich Null Anfragen im Server befinden, nur eine Zustandsänderung in eine Richtung möglich ist, folgt aus der Bedingung 4.15 für die erste Spalte der gegebenen Intensitätsmatrix Q:

$$0 \quad = \quad -\pi_0\lambda + \pi_1\nu \tag{4.17}$$

Befinden sich eine oder mehr Anfragen im Server, ist eine Zustandsänderung in beide Richtungen möglich. Aus der Bedingung 4.15 folgt für alle weiteren Spalten der Intensitätsmatrix Q:

$$0 \quad = \quad -\pi_k(\lambda+\nu) + \pi_{k-1}\lambda + \pi_{k+1}\nu \qquad \text{für} \qquad k > 0 \tag{4.18}$$

Durch vollständige Induktion lässt sich folgender Zusammenhang zwischen der Gleichgewichtswahrscheinlichkeit π und den Ankunfts- und Bearbeitungsraten λ und ν zeigen:

$$\pi_k \quad = \quad \pi_0 \left(\frac{\lambda}{\nu}\right)^k \qquad \text{für} \qquad k > 0 \tag{4.19}$$

Die Behauptung 4.19 gilt für den Fall $k = 1$. Die Gleichgewichtswahrscheinlichkeit für eine Anfrage im Server lässt sich aus Formel 4.17 herleiten:

$$\begin{aligned} 0 &= -\pi_0\lambda + \pi_1\nu \\ \pi_1 &= \pi_0 \left(\frac{\lambda}{\nu}\right)^1 \end{aligned} \tag{4.20}$$

Im Induktionsschritt lässt sich zeigen, dass für $k+1$ gilt:

$$\pi_{k+1} = \pi_0 \left(\frac{\lambda}{\nu}\right)^{k+1} \tag{4.21}$$

Durch Einsetzen der Behauptung 4.19 in Formel 4.17 folgt:

$$\begin{aligned} 0 &= -\pi_0 \left(\frac{\lambda}{\nu}\right)^k (\lambda+\nu) + \pi_0 \left(\frac{\lambda}{\nu}\right)^{k-1} \lambda + \pi_{k+1}\nu \\ 0 &= -\pi_0 \left(\frac{\lambda}{\nu}\right)^k (\lambda+\nu) + \pi_0 \left(\frac{\lambda}{\nu}\right)^{k-1} \frac{\nu\lambda}{\nu} + \pi_{k+1}\nu \\ 0 &= \pi_0 \left(\frac{\lambda}{\nu}\right)^k (\lambda+\nu) - \pi_0 \left(\frac{\lambda}{\nu}\right)^k \nu - \pi_{k+1}\nu \\ \pi_{k+1} &= \frac{\pi_0 \left(\frac{\lambda}{\nu}\right)^k (\lambda+\nu) - \pi_0 \left(\frac{\lambda}{\nu}\right)^k \nu}{\nu} \\ \pi_{k+1} &= \pi_0 \left(\frac{\lambda}{\nu}\right)^k \left(\frac{\lambda+\nu-\nu}{\nu}\right) \\ \pi_{k+1} &= \pi_0 \left(\frac{\lambda}{\nu}\right)^{k+1} \qquad \text{q. e. d.} \end{aligned} \tag{4.22}$$

Anhand der Wahrscheinlichkeit, dass sich der Server im Gleichgewicht mit k unbearbeiteten Anfragen befindet, kann die Auslastung des Servers ermittelt werden.

4.1.3 Auslastung des Servers

Das Verhältnis zwischen Ankunfts- und Bearbeitungsrate beschreibt die Auslastung ρ des Servers. Damit die Anzahl nicht bearbeiteter Anfragen nicht unendlich groß wird, muss die Bearbeitungsrate größer sein als die Ankunftsrate:

$$\rho = \frac{\lambda}{\nu} < 1 \tag{4.23}$$

Die Wahrscheinlichkeit, dass sich der Server im Gleichgewicht mit k unbearbeiteten Anfragen befindet, lautet in Abhängigkeit von der Serverauslastung ρ und gegebener Wahrscheinlichkeit, dass sich der Server im Gleichgewicht mit Null Anfragen befindet:

$$\begin{aligned} \pi_k &= \pi_0 \left(\frac{\lambda}{\nu}\right)^k \qquad \text{für} \qquad k > 0 \\ &= \pi_0 \rho^k \end{aligned} \tag{4.24}$$

Nach dem Gesetz der totalen Wahrscheinlichkeit beträgt die Summe aller möglichen Gleichgewichtswahrscheinlichkeiten, in denen sich der Server befinden kann, Eins:

$$1 = \pi_0 + \sum_{k=1}^{\infty} \pi_0 \rho^k \tag{4.25}$$

Aus Formel 4.25 lässt sich die Wahrscheinlichkeit für den Zustand herleiten, in dem sich der Server im Gleichgewicht mit Null Anfragen befindet:

$$\begin{aligned} \pi_0 &= \frac{1}{1 + \sum_{k=1}^{\infty} \rho^k} \\ &= \frac{1}{1 - \rho^0 + \sum_{k=0}^{\infty} \rho^k} \\ &= \frac{1}{\sum_{k=0}^{\infty} \rho^k} \end{aligned} \tag{4.26}$$

Für die Konvergenz der unendlichen Reihe im Divisor gilt (vgl. Ghorpade und Limaye, 2006, S. 47 f.):

$$\sum_{k=0}^{\infty} \rho^k = \frac{1}{1-\rho} \tag{4.27}$$

Die Wahrscheinlichkeit, dass sich der Server im Gleichgewicht mit Null Anfragen befindet, lautet somit:

$$\pi_0 = 1 - \rho \tag{4.28}$$

In Formel 4.24 lässt sich durch Einsetzen der Formel 4.28 die Wahrscheinlichkeit eliminieren, dass sich der Server im Gleichgewicht mit Null Anfragen befindet. Die Wahrscheinlichkeit, dass sich der Server im Gleichgewicht mit k unbearbeiteten Anfragen befindet, lautet in Abhängigkeit der Serverauslastung ρ:

$$\begin{aligned} \pi_k &= \pi_0 \rho^k \\ &= (1-\rho)\rho^k \end{aligned} \tag{4.29}$$

Aus den Gleichgewichtswahrscheinlichkeiten, dass sich der Server im Gleichgewicht mit k unbearbeiteten Anfragen befindet, lässt sich die mittlere Anzahl an Anfragen bestimmen, die sich im Server befinden:

$$\begin{aligned} \overline{K} &= \sum_{k=1}^{\infty} k\pi_k \\ &= \sum_{k=1}^{\infty} k(1-\rho)\rho^k \\ &= (1-\rho)\sum_{k=1}^{\infty} k\rho^k \\ &= (1-\rho)(-0\rho^0 + \sum_{k=0}^{\infty} k\rho^k) \end{aligned} \tag{4.30}$$

$$
\begin{aligned}
&= (1-\rho)(-0\rho^0 + \sum_{k=0}^{\infty} k\rho^k) \\
&= (1-\rho)\sum_{k=0}^{\infty} k\rho^k \\
&= \rho(1-\rho)\sum_{k=0}^{\infty} k\frac{\rho^k}{\rho} \\
&= \rho(1-\rho)\sum_{k=0}^{\infty} k\rho^{k-1} \\
&= \rho(1-\rho)\frac{\mathrm{d}\sum_{k=0}^{\infty}\rho^k}{\mathrm{d}\rho}
\end{aligned}
\tag{4.31}
$$

Durch Substitution der unendlichen Reihe nach Formel 4.27 beträgt die mittlere Anzahl an Anfragen, die sich im Server befinden, in Abhängigkeit von der Serverauslastung ρ (vgl. Bolch, 1998, S. 214):

$$
\begin{aligned}
\overline{K} &= \rho(1-\rho)\frac{\mathrm{d}\frac{1}{1-\rho}}{\mathrm{d}\rho} \\
&= \frac{\rho}{1-\rho}
\end{aligned}
\tag{4.32}
$$

Der Satz von Little (vgl. Little, 1961, S. 383 ff.) beschreibt die mittlere Anzahl von Anfragen im Server in Abhängigkeit von der Ankunftsrate λ und der mittleren Antwortzeit $\overline{K}$:

$$
\overline{K} = \lambda\overline{T} \tag{4.33}
$$

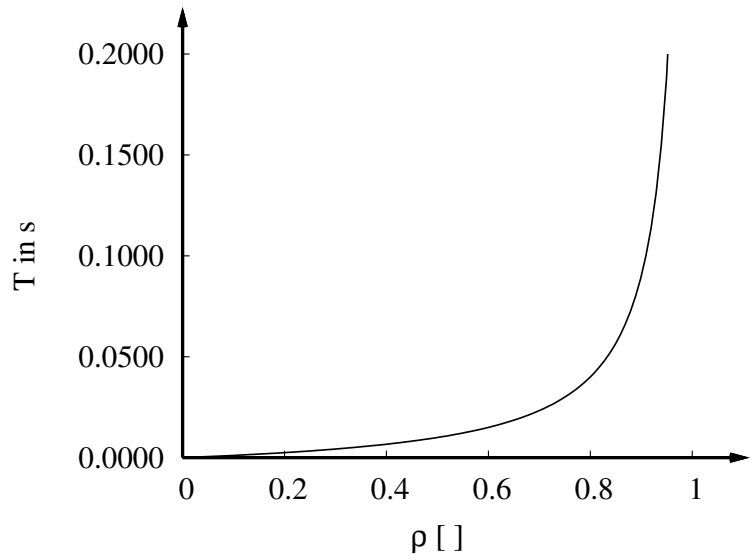

Abbildung 4.2: Antwortzeit in Abhängigkeit von der Serverauslastung

Durch Einsetzten des Satzes von Little in Formel 4.32 lässt sich die mittlere Antwortzeit $\overline{T}$ in Abhängigkeit von der Serverauslastung ρ und der Ankunftsrate λ der Anfragen bestimmen:

$$\overline{T} = \frac{\rho}{\lambda(1-\rho)} \tag{4.34}$$

Abbildung 4.2 zeigt die Entwicklung der mittleren Antwortzeit $\overline{T}$ in Abhängigkeit von der Serverauslastung ρ für eine Ankunftsrate von 100 Anfragen pro Sekunde an einen Server ($\lambda = 100$). Aus der Abbildung wird ersichtlich, dass ein Server nicht zu 100% ausgelastet werden kann, da die Antwortzeit in diesem Fall gegen unendlich strebt. Aus Formel 4.34 folgt für die Auslastung ρ des Servers:

$$\rho = \frac{\overline{T}\lambda}{1 + \overline{T}\lambda} \tag{4.35}$$

Sind für einen Server die mittlere Antwortzeit und die Ankunftsrate der Anfragen bekannt, lässt sich durch Formel 4.35 die kritische Auslastung des Servers ρ^* bestimmen, die nicht überschritten werden darf.

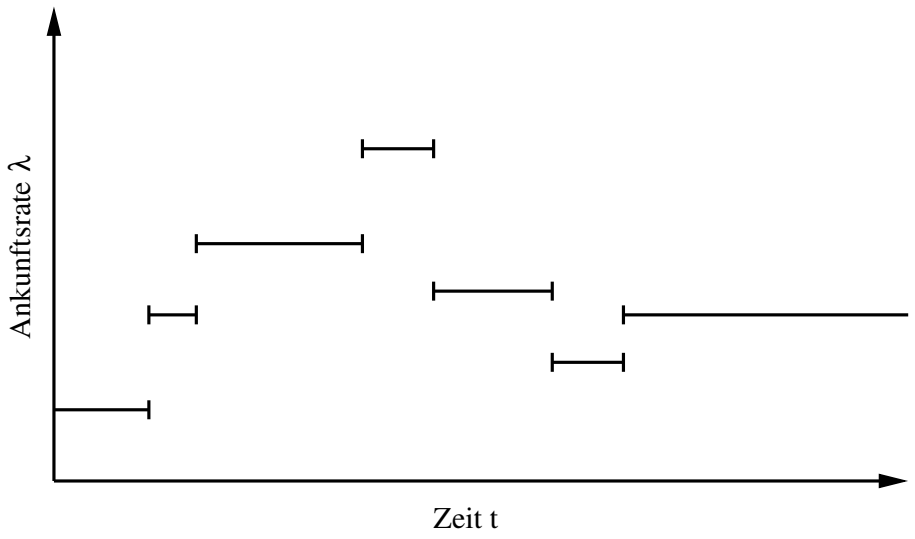

Abbildung 4.3: Ankunftsrate λ im Zeitverlauf

4.2 Bestimmung des Kapazitätsangebots

Bei der Bestimmung der kritischen Serverauslastung wurde davon ausgegangen, dass die Ankunftsrate λ der Anfragen im Zeitverlauf konstant ist. In der Realität sind zu verschiedenen Zeiten verschiedene Ankunftsraten für Anfragen an einen Server zu beobachten. Die Ankunftsrate ist somit nicht konstant sondern unterliegt Schwankungen im Zeitverlauf. Im Modell wird angenommen, dass die Ankunftsrate in verschiedenen Zeitintervallen auf unterschiedlichen Niveaus konstant ist (siehe Abbildung 4.3). Innerhalb eines Zeitintervalls stellt sich für die Anfragen im Modell ein Gleichgewichtszustand ein. Die Ankunftsrate λ, die zur Bestimmung der kritischen Serverauslastung verwendet wird, entspricht der maximalen Ankunftsrate, die der Server bearbeiten können muss, um unkritisches Antwortzeitverhalten zu gewährleisten. Die Niveaus der Ankunftsrate λ unterliegen selbst einer Verteilungsfunktion. Für die Niveaus der Ankunftsrate wird im Modell eine Normalverteilung unterstellt. Aus Formel 4.23 folgt, dass auch die Serverauslastung ρ einer Normalverteilung unterliegt, wenn die Niveaus der Ankunftsrate durch eine Normalverteilung bestimmt sind.

4.2.1 Verteilungsfunktion des Kapazitätsbedarfs

Ein Server verfügt über ein Kapazitätsangebot c_A, das sein technisches Leistungsvermögen beschreibt. Der Bedarf an technischem Leistungsvermögen, der durch Anfragen an den Server in Anspruch genommen wird, stellt den Kapazitätsbedarf c_B dar. Die Auslastung des Servers ρ lässt sich als Verhältnis aus Kapazitätsangebot c_A und Kapazitätsbedarf c_B beschreiben:

$$\rho = \frac{c_B}{c_A} \tag{4.36}$$

Da die Serverauslastung ρ einer Normalverteilung unterliegt, wird auch der Kapazitätsbedarf c_B durch eine Normalverteilung bestimmt. Der Anteil an Serveranfragen, die einen bestimmten Kapazitätsbedarf nicht überschreiten, lässt sich durch die Verteilungsfunktion F der Normalverteilung bestimmen:

$$F(c_B) = P(C \leq c_B) \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{c_B} \exp\left(-\frac{(c_B - \mu)^2}{2\sigma^2}\right) \tag{4.37}$$

Das Integral der Verteilungsfunktion des Kapazitätsbedarfs c_B lässt sich nicht analytisch berechnen und durch bekannte Funktionen in geschlossener Form beschreiben (vgl. Fahrmeir, 2004, S. 294). Die Normalverteilung des Kapazitätsbedarfs wird deshalb durch die Weibullverteilung approximiert. Die Verteilungsfunktion $F(x;\Theta)$ der 3-Parameter-Weibullverteilung beträgt (vgl. Murthy et al., 2003, S. 9):

$$F(x;\Theta) = 1 - \exp\left[-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right] \qquad \text{für} \qquad x \geq \gamma \tag{4.38}$$

Die Parameter der Verteilungsfunktion sind durch $\Theta = \{\alpha, \beta, \gamma\}$ als Skalierungs-, Form- und Lageparameter definiert. Die Dichtefunktion der Weibullverteilung ist gegeben als (vgl. Murthy et al., 2003, S. 54):

$$f(x) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} - \exp\left[-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right] \qquad \text{für} \qquad x \geq \gamma \tag{4.39}$$

Der Erwartungswert der 3-Parameter-Weibullverteilung wird über die Skalierungs-, Form- und Lageparameter bestimmt (vgl. Murthy et al., 2003, S. 54):

$$\mu = E(X) = \gamma + \alpha\Gamma\left(1 + \frac{1}{\beta}\right) \tag{4.40}$$

Die Skalierungs- und Formparameter definieren die Varianz der 3-Parameter-Weibullverteilung (vgl. Murthy et al., 2003, S. 52):

$$\sigma^2 = \alpha^2 \left[\Gamma\left(1 + \frac{2}{\beta}\right) - \left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2\right] \tag{4.41}$$

Über den Formparameter β wird die Schiefe der Weibullverteilung festgelegt. Der Formparameter lässt sich so definieren, dass die Weibullverteilung die Normalverteilung approximiert (vgl. Dubey, 1967, S. 78):

$$\beta = 3,60232 \tag{4.42}$$

Für den Formgebungsparameter β, der die Normalverteilung approximiert, ergeben sich folgende Werte der Gammafunktion (vgl. Bronshtein et al., 2007, S. 1063):

$$\begin{aligned} \Gamma\left(1 + \frac{1}{\beta}\right) &= \Gamma\left(1 + \frac{1}{3,60232}\right) \approx 0,90072 \\ \Gamma\left(1 + \frac{2}{\beta}\right) &= \Gamma\left(1 + \frac{2}{3,60232}\right) \approx 0,88964 \end{aligned} \tag{4.43}$$

Der Skalierungsparameter α beträgt in Abhängigkeit von der Varianz σ^2 für den Formparameter β, der die Normalverteilung approximiert:

$$\alpha = \sqrt{\frac{\sigma^2}{0,07834}} \tag{4.44}$$

Der Lageparameter γ beträgt in Abhängigkeit von der Varianz σ^2 und dem Erwartungswert μ für den Formparameter β, der die Normalverteilung approximiert:

$$\gamma = \mu - 0,90072 \cdot \sqrt{\frac{\sigma^2}{0,07834}} \tag{4.45}$$

Die durch die 3-Parameter-Weibullverteilung approximierte Verteilungsfunktion des Kapazitätsbedarfs c_B beträgt somit:

$$\begin{aligned} F(c_B) &= 1 - \exp\left[-\left(\frac{c_B - \gamma}{\alpha}\right)^{\beta}\right] \\ &= 1 - \exp\left[-\left(\frac{c_B - \mu - 0,90072 \cdot \sqrt{\frac{\sigma^2}{0,07834}}}{\sqrt{\frac{\sigma^2}{0,07834}}}\right)^{3,60232}\right] \end{aligned} \tag{4.46}$$

4.2.2 Berechnung des Kapazitätsangebots

Der Anteil an Serveranfragen, durch den ein bestimmter Kapazitätsbedarf c_B überschritten wird, ergibt sich aus der Differenz der totalen Wahrscheinlichkeit und der Verteilungsfunktion des Kapazitätsbedarfs $F(c_B)$. Für einen gegebenen Kapazitätsbedarf, der einer Normalverteilung unterliegt, die durch die 3-Parameter-Weibullverteilung approximiert wird, beträgt der Anteil an kritischen Serveranfragen κ, die einen Kapazitätsbedarf c_B überschreiten:

$$\begin{aligned} \kappa &= P(C > c_B) = 1 - F(c_B) \\ \kappa &= \exp\left[-\left(\frac{c_B - \gamma}{\alpha}\right)^{\beta}\right] \end{aligned} \tag{4.47}$$

Ist die kritische Serverauslastung ρ^* bekannt, lässt sich der Kapazitätsbedarf c_B nach Formel 4.36 ersetzen:

$$\kappa = \exp\left[-\left(\frac{\rho^* c_A - \gamma}{\alpha}\right)^{\beta}\right] \tag{4.48}$$

Aus Formel 4.48 lässt sich das Kapazitätsangebot c_A in Abhängigkeit von dem Erwartungswert μ und der Standardabweichung σ^2 des Kapazitätsbedarfs c_B sowie der kritischen Serverauslastung κ ermitteln:

$$\begin{aligned} \exp\left[-\left(\frac{\rho^* c_A - \gamma}{\alpha}\right)^{\beta}\right] &= \kappa \\ -\left(\frac{\rho^* c_A - \gamma}{\alpha}\right)^{\beta} &= \ln(\kappa) \\ \frac{\rho^* c_A - \gamma}{\alpha} &= [-\ln(\kappa)]^{\frac{1}{\beta}} \\ c_A &= \frac{\alpha\,[-\ln(\kappa)]^{\frac{1}{\beta}} + \gamma}{\rho^*} \end{aligned} \tag{4.49}$$

Für Form-, Skalierungs- und Lageparameter bei approximierter Normalverteilung beträgt das benötigte Kapazitätsangebot c_A eines Servers mit einem Anteil von κ Serveranfragen, die im kritischen Bereich liegen dürfen:

$$\begin{aligned} c_A &= \frac{\sqrt{\frac{\sigma^2}{0,07834}}\,[-\ln(\kappa)]^{\frac{1}{3,60232}} + \mu - 0,90072 \cdot \sqrt{\frac{\sigma^2}{0,07834}}}{\rho^*} \\ &= \frac{\sqrt{\frac{\sigma^2}{0,07834}}\left([-\ln(\kappa)]^{\frac{1}{3,60232}} - 0,90072\right) + \mu}{\rho^*} \end{aligned} \tag{4.50}$$

Das Kapazitätsangebot wird somit durch den Erwartungswert und die Standardabweichung des Kapazitätsbedarfs, die Anzahl an Anfragen die in einem definierten Zeitabschnitt bearbeitet werden sollen, sowie die mittlere Antwortzeit einer Anfrage bestimmt. Im nächsten Schritt wird der Einfluss der Virtualisierung auf das Kapazitätsangebot untersucht.

4.2.3 Berechnung des Kapazitätsangebots bei Virtualisierung

Ein physischer Server kann in mehrere virtuelle Server partitioniert werden. Im Modell wird eine Virtualisierungstechnik unterstellt, die es ermöglicht, das Kapazitätsangebot des physischen Servers dynamisch den virtuellen Servern zuzuteilen. Das Kapazitätsangebot des physischen Servers kann im vollen Umfang den virtuellen Servern zur Verfügung gestellt werden. Im Modell existiert somit kein Virtualisierungsoverhead. Die Kapazitätsbedarfe der virtuellen Server stellen statistische Zufallsgrößen dar. Diese Zufallsgrößen sind korreliert, wenn zwischen ihnen eine statistische Abhängigkeit besteht (vgl. Wunsch und Schreiber, 2006, S. 41). Die Kapazitätsbedarfe der virtuellen Server sollen unkorreliert sein. Für die Kovarianzmatrix (vgl. Fahrmeir, 1996, S. 23) gilt:

$$\text{cov}(x) = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \dots & \sigma_{nn} \end{pmatrix} \quad \text{mit} \quad \sigma_{ij} = \begin{cases} 0, & i \neq j \\ \sigma_i, & i = j \end{cases} \tag{4.51}$$

Der Erwartungswert des Kapazitätsbedarfs des physischen Servers $E(C_{PB})$ ist die Summe der Erwartungswerte der Kapazitätsbedarfe der virtuellen Server $E(C_{VB})$ (vgl. Mosler und Schmid, 2006, S. 140):

$$\begin{aligned} E(C_{PB}) &= E(C_{VB1} + C_{VB2} + ... + C_{VBn}) \\ &= \sum_{i=1}^{n} E(C_{VBi}) = \sum_{i=1}^{n} \mu_{VBi} \end{aligned} \tag{4.52}$$

Die Varianz des Kapazitätsbedarfs des physischen Servers $\text{Var}(C_{PB})$ beträgt für zwei virtuelle Server im allgemeinen Fall (vgl. Mosler und Schmid, 2006, S. 140):

$$\begin{aligned} \text{Var}(C_{PB}) &= \text{Var}(C_{VB1} + C_{VB2}) \\ &= \text{Var}(C_{VB1}) + \text{Var}(C_{VB2}) + 2\ \text{cov}(C_{VB1}, C_{VB2}) \end{aligned} \tag{4.53}$$

Die Verteilungen der Kapazitätsbedarfe der virtuellen Server sollen im Modell unkorreliert sein. Die Varianz des Kapazitätsbedarfs des physischen Servers für n virtuelle Server ist in diesem Fall:

$$\begin{aligned} \mathrm{Var}(C_{PB}) &= \sigma_{PB}^2 &= \mathrm{Var}(C_{VB1} + C_{VB2} + ... + C_{VBn}) \\ &= \sum_{i=1}^{n} \mathrm{Var}(C_{VBi}) &= \sum_{i=1}^{n} \sigma_{VBi}^2 \end{aligned} \tag{4.54}$$

Bei vernachlässigtem Virtualisierungsoverhead ist das Kapazitätsangebot c_{PA} des physischen Servers die Summe der Kapazitätsangebote c_{VAi} der virtuellen Server:

$$c_{PA} = \sum_{i=1}^{n} c_{VAi} \tag{4.55}$$

Im Modell wird vereinfachend angenommen, dass auf einem physischen Server n identische virtuelle Server betrieben werden. Der Erwartungswert und die Standardabweichung der Kapazitätsbedarfe sind für jeden der virtuellen Server gleich. Erwartungswert $E(C_{PB})$ und Varianz $\mathrm{Var}(C_{PB})$ des Kapazitätsbedarfs des physischen Server betragen:

$$\begin{aligned} E(C_{PB}) &= \mu_{PB} &= n\mu_{VB} \\ \mathrm{Var}(C_{PB}) &= \sigma_{PB}^2 &= n\sigma_{VB}^2 \end{aligned} \tag{4.56}$$

Für n identische virtuelle Server beträgt das Kapazitätsangebot c_{PA} des physischen Servers:

$$c_{PA} = nc_{VA} \tag{4.57}$$

Formel 4.50 beschreibt das Kapazitätsangebot eines physischen Servers ohne Virtualisierung:

$$c_{PA} = \frac{\sqrt{\frac{\sigma_{PB}^2}{0,07834}} \left([-\ln(\kappa)]^{\frac{1}{3,60232}} - 0,90072 \right) + \mu_{PB}}{\rho^*} \tag{4.58}$$

Erwartungswert und Varianz des physischen Servers lassen sich nach Formel 4.56 durch Erwartungswert und Varianz der virtuellen Server ersetzen. Das notwendige Kapazitätsangebot c_{PA} eines physischen Servers mit n identischen virtuellen Servern beträgt in Abhängigkeit von der kritischen Serverauslastung ρ^*, dem Anteil an Anfragen κ, die im kritischen Bereich liegen dürfen, dem Erwartungswert des Kapazitätsbedarfs eines virtuellen Servers μ_{VB} und der Varianz des Kapazitätsbedarfs eines virtuellen Servers σ^2_{VB}:

$$c_{PA} = \frac{\sqrt{\frac{n\sigma^2_{VB}}{0,07834}}\left([-\ln(\kappa)]^{\frac{1}{3,60232}} - 0,90072\right) + n\mu_{VB}}{\rho^*} \tag{4.59}$$

Da der physische Server aus n identischen virtuellen Servern besteht, lässt sich das benötigte Kapazitätsangebot des physischen Servers anteilig auf die virtuellen Server umlegen:

$$\begin{aligned} \frac{c_{PA}}{n} = c_{VA} &= \frac{\sqrt{\frac{n\sigma^2_{VB}}{0,07834}}\left([-\ln(\kappa)]^{\frac{1}{3,60232}} - 0,90072\right) + n\mu_{VB}}{\rho^* n} \\ &= \frac{\sqrt{\frac{\sigma^2_{VB}}{0,07834n}}\left([-\ln(\kappa)]^{\frac{1}{3,60232}} - 0,90072\right) + \mu_{VB}}{\rho^*} \end{aligned} \tag{4.60}$$

Behauptung: Sollen weniger als die Hälfte aller Serveranfragen zu einem kritischen Kapazitätsbedarf führen ($\kappa < 0,5$), sinkt das notwendige Kapazitätsangebot je virtuellem Server mit steigender Zahl an virtuellen Servern.
Beweis: Wenn die Aussage wahr ist, muss für $\kappa < 0,5$ die Formel 4.60 eine stetig fallende Funktion sein:

$$\frac{\mathrm{d}c_{VA}}{\mathrm{d}n} = \underbrace{-\frac{\sqrt{\frac{\sigma^2_{VB}}{0,07834}}}{2n^{\frac{3}{2}}\rho^*}}_{Faktor1}\underbrace{\left([-\ln(\kappa)]^{\frac{1}{3,60232}} - 0,90072\right)}_{Faktor2} < 0 \tag{4.61}$$

Für eine positive Anzahl virtueller Server n und eine positive Serverauslastung ρ ist Faktor 1 stets negativ. Somit muss für Faktor 2 gelten:

$$\begin{aligned}
[-\ln(\kappa)]^{\frac{1}{3,60232}} - 0,90072 &> 0 \\
[-\ln(\kappa)]^{\frac{1}{3,60232}} &> 0,90072 \\
-\ln\kappa &> 0,90072^{3,60232} \\
\ln\kappa &< -0,68615 \\
\kappa &< e^{-0,68615} \\
\kappa &< 0,50351 \approx 0,5 \qquad \text{q. e. d.}
\end{aligned} \tag{4.62}$$

Aus Sicht der Kapazitätsplanung ist es sinnvoll, so viele virtuelle Server wie möglich auf einem physischen Server zu betreiben. Auf einem unendlich großen physischen Server mit unendlich vielen virtuellen Servern ist das benötigte Kapazitätsangebot eines virtuellen Servers das Verhältnis aus dem Erwartungswert des Kapazitätsbedarfs und der kritischen Serverauslastung:

$$\lim_{n\to\infty} \frac{c_{PA}}{n} = \frac{\mu_{VB}}{\rho^*} \tag{4.63}$$

Ein solcher Server lässt sich in der Praxis nicht realisieren, da das technische Leistungsvermögen eines physischen Servers begrenzt ist. Werden auf einem physischen Server zwei identische virtuelle Server betrieben, lässt sich der Einfluss der Korrelation auf das Kapazitätsangebot zeigen. Die Formel zur Berechnung des Kapazitätsangebots zweier identischer virtueller Server, deren Kapazitätsbedarfe zueinander in statistischer Abhängigkeit stehen, lautet:

$$c_{PA} = \frac{\sqrt{\frac{2\sigma_{VB}^2 + 2\mathrm{cov}(C_{VB1}, C_{VB2})}{0,07834}} \left([-\ln(\kappa)]^{\frac{1}{3,60232}} - 0,90072 \right) + 2\mu_{VB}}{\rho^*} \tag{4.64}$$

Durch den Einfluss der Korrelation soll das benötigte Kapazitätsangebot niedriger sein als im unkorrelierten Fall. Zwei identische physische Server mit jeweils zwei identischen virtuellen Servern mit identischen Erwartungswerten und Varianzen sollen miteinander verglichen werden. Zwischen zwei virtuellen Servern eines physischen Servers soll eine statistische Abhängigkeit bestehen, zwischen den anderen beiden virtuellen Servern nicht. Dieser Zusammenhang wird durch folgende Ungleichung beschrieben:

$$\frac{\sqrt{\frac{2\sigma_{VB}^2}{0,07834}}\left([-\ln(\kappa)]^{\frac{1}{3,60232}}-0,90072\right)+2\mu_{VB}}{\rho^*} > \frac{\sqrt{\frac{2\sigma_{VB}^2+2\mathrm{cov}(C_{VB1},C_{VB2})}{0,07834}}\left([-\ln(\kappa)]^{\frac{1}{3,60232}}-0,90072\right)+2\mu_{VB}}{\rho^*} \tag{4.65}$$

Weniger als die Hälfte der Serveranfragen sollen einen kritischen Kapazitätsbedarf verursachen. Nach Formel 4.62 gilt für $\kappa < 0,5$:

$$[-\ln(\kappa)]^{\frac{1}{3,60232}}-0,90072 \;>\; 0 \tag{4.66}$$

Die Serverauslastung ρ ist ebenfalls immer positiv. Somit lässt sich die Ungleichung der Kapazitätsangebote wie folgt darstellen:

$$\begin{aligned}\sqrt{\frac{2\sigma_{VB}^2}{0,07834}} &> \sqrt{\frac{2\sigma_{VB}^2+2\mathrm{cov}(C_{VB1},C_{VB2})}{0,07834}}\\ \sqrt{\sigma_{VB}^2} &> \sqrt{\sigma_{VB}^2+\mathrm{cov}(C_{VB1},C_{VB2})}\end{aligned} \tag{4.67}$$

Diese Ungleichung gilt für negative Kovarianz zwischen den Kapazitätsbedarfen der virtuellen Server:

$$\mathrm{cov}(C_{VB1},C_{VB2}) < 0 \tag{4.68}$$

Für $\kappa < 0,5$ ist im unkorrelierten Fall ein höheres Kapazitätsangebot notwendig als bei negativer Korrelation. Die Zufallsvariablen C_{VB1} und C_{VB2}

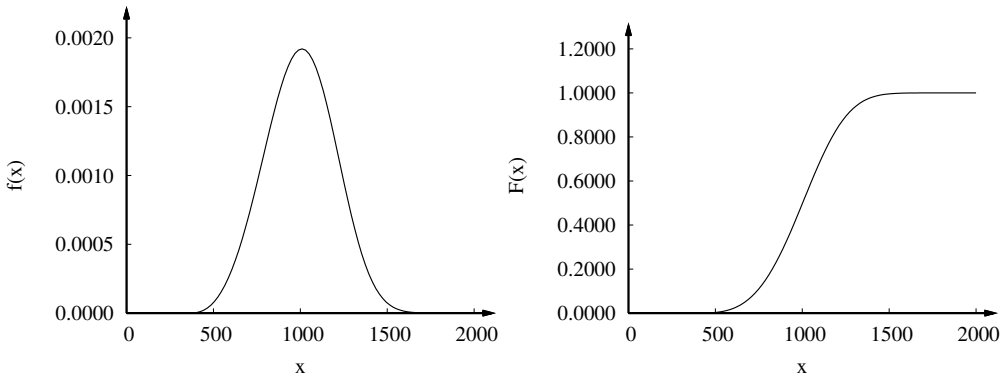

Abbildung 4.4: Verteilungs- und Dichtefunktion

repräsentieren die Kapazitätsbedarfe der virtuellen Server. Negative Korrelation liegt vor, wenn die Zufallsvariablen C_{VB1} und C_{VB2} einen gegensinnigen linearen Zusammenhang besitzen (vgl. Fahrmeir, 2004, S. 350). Positive Korrelation führt zu einer Erhöhung des Kapazitätsangebots gegenüber dem unkorrelierten Fall. Soll das benötigte Kapazitätsangebot minimiert werden, ist es sinnvoll, virtuelle Server mit negativer Korrelation auf einem physischen Server zusammenzufassen.

4.3 Beispiel einer Kapazitätsangebotsplanung

In einem Rechenzentrum soll das Kapazitätsangebot für vier identische Server geplant werden. Ein Server soll in der Lage sein, im Mittel 100 Anfragen pro Sekunde zu verarbeiten. Ankunftsrate und Bearbeitungsrate unterliegen einer Poissonverteilung. Die mittlere Antwortzeit soll 0,03 Sekunden betragen. Diese Anforderung soll für 95% aller Anfragen erfüllt werden. Die Auslastung des Kapazitätsangebots ist normalverteilt. Die Normalverteilung wird durch die Weibullverteilung approximiert. Der Erwartungswert des Kapazitätsbedarfs beträgt 1000 MIPS bei einer Standardabweichung von 200 MIPS. Abbildung 4.4 zeigt die Verteilungs- und Dichtefunktion des Kapazitätsbedarfs für einen Server. Der Kapazitätsbedarf der Server untereinander soll unkorreliert sein.

Es soll das Kapazitätsangebot ermittelt werden, wenn die vier Server als physische Server realisiert werden. Weiterhin soll gezeigt werden, wieviel Kapazität eingespart werden kann, wenn vier virtuelle Server auf einem physischen Server mit dynamischer Kapazitätszuteilung installiert werden.

4.3.1 Berechnung der kritischen Serverauslastung

Die Ankunftsrate λ der poissonverteilten Serveranfragen beträgt 100 Anfragen pro Sekunde:

$$\lambda = 100 \tag{4.69}$$

Die mittlere Antwortzeit $\overline{T}$ zur Bearbeitung einer Anfrage beträgt 0,03 Sekunden:

$$\overline{T} = 0,03 \tag{4.70}$$

Nach Formel 4.35 lässt sich aus der Ankunftsrate λ und der mittleren Antwortzeit $\overline{T}$ die kritische Serverauslastung ρ^* ermitteln. Die kritische Serverauslastung darf nicht überschritten werden, wenn die mittlere Antwortzeit eingehalten werden soll:

$$\begin{aligned} \rho^* &= \frac{\overline{T}\lambda}{1+\overline{T}\lambda} \\ &= \frac{0,03 \cdot 100}{1+0,03 \cdot 100} \\ &= 0,75 \end{aligned} \tag{4.71}$$

Soll das mittlere Antwortzeitverhalten von 0,03 Sekunden bei 100 Anfragen pro Sekunde eingehalten werden, darf eine Serverauslastung von 75% nicht überschritten werden.

4.3.2 Berechnung des Kapazitätsangebots

Der Erwartungswert μ des Kapazitätsbedarfs c_B eines Servers ist 1000 MIPS:

$$\mu = E(C) = 1000 \tag{4.72}$$

Die Standardabweichung σ des Kapazitätsbedarfs c_B eines Servers beträgt 200 MIPS:

$$\sigma = 200 \tag{4.73}$$

Die Anfragen sollen in 95% aller Fälle einen Kapazitätsbedarf beanspruchen, der eine Serverauslastung von 75% nicht überschreitet. Der Anteil kritischer Serveranfragen κ beträgt:

$$\kappa = 0,05 \tag{4.74}$$

Nach Formel 4.50 lässt sich das Kapazitätsangebot c_A für einen Server ermitteln, der diese Anforderungen erfüllt:

$$\begin{aligned} c_A &= \frac{\sqrt{\frac{200^2}{0,07834}}\left([-\ln(0,05)]^{\frac{1}{3,60232}} - 0,90072\right) + 1000}{0,75} \\ &= 1767,14995 \end{aligned} \tag{4.75}$$

Das Kapazitätsangebot eines Servers beträgt 1768 MIPS.

4.3.3 Berechnung des Kapazitätsangebots bei Virtualisierung

Im nächsten Schritt wird das Kapazitätsangebot eines physischen Servers ermittelt, auf dem vier identische virtuelle Server entsprechend den Anforderungen installiert werden. Das Kapazitätsangebot c_{PA} des physischen Servers

lässt sich nach Formel 4.59 berechnen:

$$c_{PA} = \frac{\sqrt{\frac{4 \cdot 200^2}{0,07834}} \left([-\ln(0,05)]^{\frac{1}{3,60232}} - 0,90072 \right) + 4 \cdot 1000}{0,75} \quad (4.76)$$

$$= 6200,96656$$

Das Kapazitätsangebot beträgt 6201 MIPS. Die Kapazitätsersparnis durch Virtualisierung errechnet sich aus der Differenz der Kapazitätsangebote von vier physischen Servern ohne Virtualisierung und einem physischen Server, auf dem vier virtuelle Server installiert werden:

$$4 \cdot 1768 - 6201 = 871 \quad (4.77)$$

Werden auf einem physischen Server vier virtuelle Server mit den gegebenen Anforderungen betrieben, beträgt die Kapazitätsersparnis 871 MIPS.

4.4 Zusammenfassung

Der Einsatz von Virtualisierungstechniken führt zu Einsparungen des Kapazitätsangebots. Die Hypothese konnte durch ein mathematisch-formales Modell bestätigt werden. Diese Erkenntnis rechtfertigt die in Kapitel 3 getroffene Annahme, dass der Einsatz von Virtualisierungstechniken aus Sicht der Kapazitätswirtschaft sinnvoll ist.
Die unterschiedlichen Auslastungen der virtuellen Server liefern die Erklärung für die Kapazitätseinsparung. Das Maß der Einsparung wird durch die Korrelation der Verteilungsfunktionen der Auslastung der virtuellen Server beeinflusst. Diese Erkenntnis bildet eine der Grundlagen zur Beantwortung der Forschungsfrage der Betriebsmittelzuordnung in Kapitel 6.
Bei der Planung des Kapazitätsangebots wird die zu erwartende Gesamtnachfrage nicht berücksichtigt. Durch Nachfrage von IT-Dienstleistungen entsteht ein Kapazitätsbedarf. Die Planung des Kapazitätsbedarfs wird im folgenden Kapitel erläutert.

Kapitel 5

Planung des Kapazitätsbedarfs

In Kapitel 3 wurde festgestellt, dass die Planung des Kapazitätsangebots nicht ausschließlich auf Grundlage der Produktionsprogrammplanung durchgeführt werden kann. Diese Behauptung soll als eigene Forschungsfrage untersucht werden. In diesem Kapitel wird folgende Hypothese überprüft:

> Der Kapazitätsbedarf an Potentialfaktoren wächst auch bei konstanter Nachfrage nach IT-Dienstleistungen exponentiell.

Bei der Bestätigung der Hypothese wird von Gültigkeit des Mooreschen Gesetzes (vgl. Tanenbaum, 1999, S. 25) als Bedingung ausgegangen. Als Ergebnis liegt ein mathematisch-formales Modell zur Kapazitätsentwicklung im Rechenzentrum vor.
Im folgenden Abschnitt werden die Potentialfaktoren eines Rechenzentrums aus Sicht der Kapazitätswirtschaft beschrieben. Anschließend werden die Prognoseverfahren für die Produktionsprogrammplanung erläutert. Anhand dieser Darstellungen werden die Prognoseverfahren der Produktionsprogrammplanung zur Ermittlung des Bedarfs an Potentialfaktoren in Rechenzentren angewendet.
Das beschriebene Kapazitätsmodell wurde in Auszügen bereits im Tagungsband der Multikonferenz Wirtschaftsinformatik (vgl. Osterburg et al., 2008, S. 405 ff.) veröffentlicht. Die Planung des Kapazitätsbedarfs wurde bereits im Journal Informatica Economica (vgl. Hanisch et al., 2009, S. 9 ff.) vorgestellt.

5.1 Potentialfaktoren

Zur Erbringung einer IT-Dienstleistung werden vom Rechenzentrum Potentialfaktoren zur Verfügung gestellt. Um den Kapazitätsbedarf an Potentialfaktoren für zukünftige Perioden prognostizieren zu können, muss der trendförmige Verlauf des Kapazitätsbedarfs bekannt sein. Hierzu wird in diesem Abschnitt ein Kapazitätsmodell für Potentialfaktoren im Rechenzentrum entwickelt.

5.1.1 Potentialfaktoren in Rechenzentren

Die Entwicklung der Kapazitäten der Hardwarekomponenten lässt sich anhand des Mooreschen Gesetzes beschreiben. Das Moorsche Gesetz besagt, dass sich die Anzahl an Transistoren auf einem Schaltkreis etwa alle 18 Monate verdoppelt (vgl. Tanenbaum, 1999, S. 25). Formal wächst die Kapazität c der Hardwarekomponenten in Abhängigkeit von der Zeit t exponentiell:

$$c(t) = c_0 \cdot e^{\lambda \cdot t} \tag{5.1}$$

Der Faktor c_0 beschreibt die initiale Kapazität. Der Exponent λ beschreibt die Wachstumsrate. Steigt die Kapazität innerhalb der Zeit T_n um den Faktor n, gilt:

$$\begin{aligned} n \cdot c_0 \cdot e^{\lambda \cdot t} &= c_0 \cdot e^{\lambda (t + T_n)} \\ n \cdot c_0 \cdot e^{\lambda \cdot t} &= c_0 \cdot e^{\lambda \cdot t} \cdot e^{\lambda \cdot T_n} \\ n &= e^{\lambda \cdot T_n} \\ \ln n &= \lambda \cdot T_n \\ \lambda &= \frac{\ln n}{T_n} \end{aligned} \tag{5.2}$$

Die Wachstumsrate für eine Verdopplung der Kapazität alle 1,5 Perioden beträgt beispielsweise:

$$\lambda = \frac{\ln 2}{1,5} \approx 0,462 \tag{5.3}$$

5.1.2 Modell eines Rechenzentrums

Werden im Modell nach n Perioden alle q Hardwareeinheiten der ältesten Generation gegen q aktuelle Hardwareeinheiten ausgetauscht, lässt sich zeigen, dass auch die Gesamtkapazität der Hardwarekomponenten im Rechenzentrum exponentiell wächst.
Die Gesamtkapazität des Rechenzentrums $c_{Hardware}$ zum Zeitpunkt t ist die Summe der Gesamtkapazitäten jeder Hardwaregeneration im Rechenzentrum. Aus Formel 5.1 folgt bei angenommener Substituierbarkeit aller Hardwareeinheiten:

$$c_{Hardware}(t) = \sum_{t^*=0}^{n-1} q \cdot c_0^* \cdot e^{\lambda \cdot t^*} \tag{5.4}$$

Das Rechenzentrum besteht aus n Generationen von je q Hardwareeinheiten. Die Kapazität der der ältesten Hardwaregeneration c_0^* lässt sich anhand der Formel 5.1 beschreiben:

$$c_0^* = c_0 \cdot e^{\lambda(t-n+1)} \tag{5.5}$$

Für die Kapazität des Rechenzentrums (siehe Formel 5.4) ergibt sich somit:

$$\begin{aligned} c_{Hardware}(t) &= \sum_{t^*=0}^{n-1} q \cdot c_0 \cdot e^{\lambda(t-n+1)} \cdot e^{\lambda \cdot t^*} \\ &= q \cdot c_0 \cdot e^{\lambda(t-n+1)} \cdot \sum_{t^*=0}^{n-1} \left(e^{\lambda}\right)^{t^*} \end{aligned} \tag{5.6}$$

Die Summe beschreibt eine geometrische Reihe und lässt sich deshalb weiter umformen:

$$\begin{aligned} c_{Hardware}(t) &= q \cdot c_0 \cdot e^{\lambda(t-n+1)} \cdot \frac{e^{\lambda \cdot n} - 1}{e^{\lambda} - 1} \\ &= q \cdot c_0 \cdot \frac{e^{\lambda \cdot t}}{e^{\lambda(n-1)}} \cdot \frac{e^{\lambda \cdot n} - 1}{e^{\lambda} - 1} \\ &= q \cdot c_0 \cdot e^{\lambda \cdot t} \cdot \frac{e^{\lambda}}{e^{\lambda \cdot n}} \cdot \frac{e^{\lambda \cdot n} - 1}{e^{\lambda} - 1} \\ &= q \cdot \frac{e^{\lambda(n+1)} - e^{\lambda}}{e^{\lambda(n+1)} - e^{\lambda \cdot n}} \cdot c_0 \cdot e^{\lambda \cdot t} \\ &= q \cdot \frac{e^{\lambda \cdot n} - 1}{e^{\lambda \cdot n} - e^{\lambda(n-1)}} \cdot c_0 \cdot e^{\lambda \cdot t} \end{aligned} \tag{5.7}$$

Die Gesamtkapazität der Hardwarekomponenten eines Rechenzentrums wächst somit auch exponentiell, im Durchschnitt jedoch langsamer als die eigentliche Kapazität der Hardwarekomponenten. Dieser Effekt lässt sich dadurch erklären, dass im Rechenzentrum auch Hardwareeinheiten älterer Generationen eingesetzt werden. Für die durchschnittliche Kapazität $c^{\varnothing}_{Hardware}$ in einem Rechenzentrum mit n Hardwaregenerationen mit je q Hardwareeinheiten gilt:

$$\begin{aligned} c^{\varnothing}_{Hardware}(t) &= \frac{q}{q \cdot n} \cdot \frac{e^{\lambda \cdot n} - 1}{e^{\lambda \cdot n} - e^{\lambda(n-1)}} \cdot c_0 \cdot e^{\lambda \cdot t} \\ &= \frac{e^{\lambda \cdot n} - 1}{n\left(e^{\lambda \cdot n} - e^{\lambda(n-1)}\right)} \cdot c_0 \cdot e^{\lambda \cdot t} \end{aligned} \tag{5.8}$$

Für den Spezialfall, dass das Rechenzentrum nur aus einer aktuellen Hardwaregeneration besteht, ist die durchschnittliche Kapazität gleich der allge-

meinen Kapazität. Es lässt sich zeigen, dass gilt:

$$\begin{aligned} \frac{\mathrm{d}c^{\varnothing}_{Hardware}(t)}{\mathrm{d}t} \leq \frac{\mathrm{d}c(t)}{\mathrm{d}t} \qquad & \text{für} \qquad n \in \mathbb{N}^{+} \\ & \qquad\quad \lambda \in \mathbb{R}, \quad \lambda > 0 \\ & \qquad\quad c_0 \in \mathbb{R}, \quad c_0 > 0 \end{aligned} \tag{5.9}$$

Für die erste Ableitung der Ungleichung ergibt sich:

$$\begin{aligned} \frac{e^{\lambda \cdot n} - 1}{n(e^{\lambda \cdot n} - e^{\lambda(n-1)})} \cdot c_0 \cdot \lambda \cdot e^{\lambda \cdot t} &\leq c_0 \cdot \lambda \cdot e^{\lambda \cdot t} \\ \frac{e^{\lambda \cdot n} - 1}{n(e^{\lambda \cdot n} - e^{\lambda(n-1)})} &\leq 1 \end{aligned} \tag{5.10}$$

Durch vollständige Induktion lässt sich nachweisen, dass gilt:

$$e^{\lambda \cdot n} - 1 \leq n(e^{\lambda \cdot n} - e^{\lambda(n-1)}) \tag{5.11}$$

Die Ungleichung 5.11 gilt für den Fall $n = 1$:

$$\begin{aligned} e^{\lambda \cdot 1} - 1 &\leq 1(e^{\lambda \cdot 1} - e^{\lambda(1-1)}) \\ e^{\lambda} - 1 &\leq e^{\lambda} - 1 \end{aligned} \tag{5.12}$$

Im Induktionsschritt ergibt sich für die Ungleichung 5.11:

$$e^{\lambda(n+1)} - 1 \leq (n+1)(e^{\lambda(n+1)} - e^{\lambda \cdot n}) \tag{5.13}$$

Da für $\lambda > 0$ die Exponentialfunktion $e^{\lambda} > 1$ ist, muss gelten:

$$\begin{aligned}
e^{\lambda \cdot n} - 1 &\leq n(e^{\lambda \cdot n} - e^{\lambda(n-1)}) \leq e^{\lambda} \cdot n(e^{\lambda \cdot n} - e^{\lambda(n-1)}) \\
e^{\lambda \cdot n} - 1 &\leq n \cdot e^{\lambda(n+1)} - n \cdot e^{\lambda \cdot n} \\
e^{\lambda(n+1)} + e^{\lambda \cdot n} - 1 &\leq n \cdot e^{\lambda(n+1)} + e^{\lambda(n+1)} - n \cdot e^{\lambda \cdot n} \\
e^{\lambda(n+1)} - 1 &\leq n \cdot e^{\lambda(n+1)} + e^{\lambda(n+1)} - n \cdot e^{\lambda \cdot n} - e^{\lambda \cdot n} \\
e^{\lambda(n+1)} - 1 &\leq (n+1)e^{\lambda(n+1)} - (n+1)e^{\lambda \cdot n} \\
e^{\lambda(n+1)} - 1 &\leq (n+1)(e^{\lambda(n+1)} - e^{\lambda \cdot n}) \qquad \text{q. e. d.}
\end{aligned} \tag{5.14}$$

Abbildung 5.1 zeigt die durchschnittliche Entwicklung der Prozessorkapazität im Rechenzentrum unter der Annahme, dass sich die Kapazität der Prozessoren alle zwei Perioden verdoppelt, eine Hardwaregeneration nach fünf Perioden ausgetauscht wird und zum Zeitpunkt T_0 die initiale Kapazität 8000 MIPS beträgt. Nach Formel 5.8 beträgt die durchschnittliche Prozessorkapazität in diesem Fall zum Zeitpunkt T_0 4477 MIPS.

Die exponentiell steigende Kapazität der Hardwarekomponenten wird von den Konsumenten auch in Anspruch genommen. Die Kompensation des technologischen Fortschritts durch einen steigenden Ressourcenverbrauch ist als Rebound-Effekt bekannt (vgl. Radermacher, 2000, S. 232). Die Steigerung des Ressourcenverbrauchs wird als primärer Rebound-Effekt bezeichnet (vgl. Kuhlen, 2004, S. 108).

Der Kapazitätsbedarf des Personals folgt nicht dem Trend des exponentiellen Wachstums der Kapazität der Hardwarekomponenten, sondern steht in Beziehung zur Menge der eingesetzten Potentialfaktoren. Ansonsten wäre der Betrieb eines Rechenzentrums aus finanziellen Gründen nicht durchführbar, da auch die Personalkosten exponentiell steigen würden.

Für den Kapazitätsbedarf an Personal ergeben sich Skaleneffekte. Als Skalen-

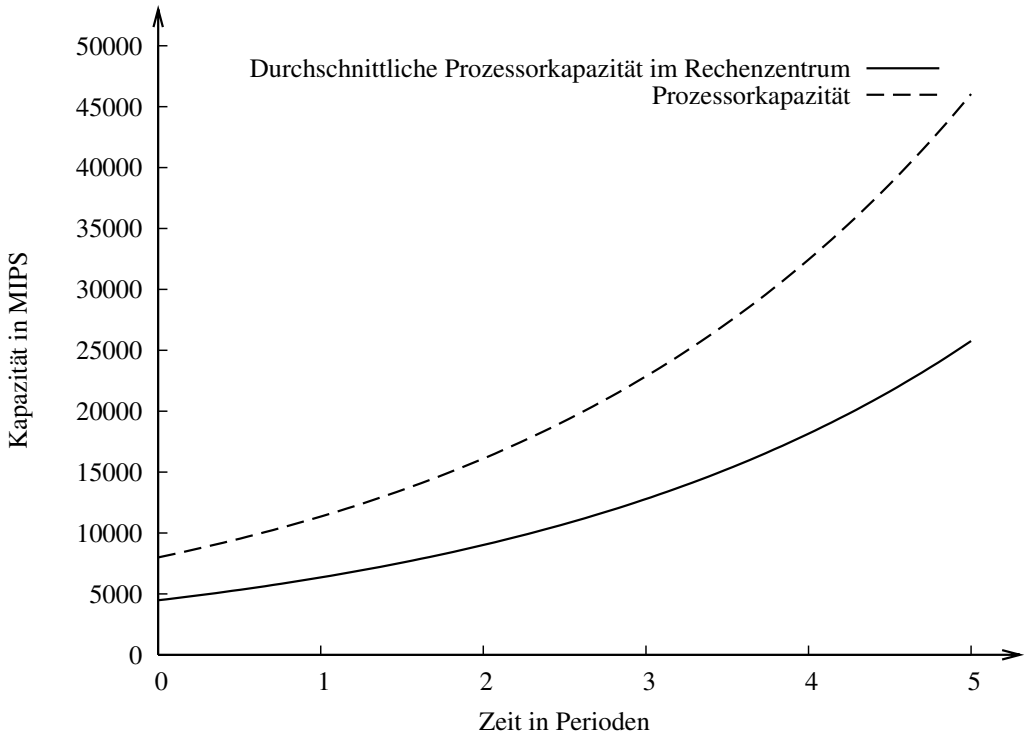

Abbildung 5.1: Entwicklung der Prozessorkapazität

effekt wird eine unterproportionale Erhöhung der Menge an Potentialfaktoren bei einer Erhöhung der Ausbringungsmenge bezeichnet. Dieser Effekt wird durch einen erhöhten Einsatz von Automatisierungstechnologien und höhere Spezialisierung des Personals bei steigender Ausbringungsmenge erzielt (vgl. Haasis, 2008, S. 93). Formal lässt sich der Kapazitätsbedarf des Personals wie folgt beschreiben:

$$c_{Personal} = c_{Fix} + \kappa \ln(n+1) \tag{5.15}$$

Die Kapazität an Personal setzt sich aus einem fixen und einem variablen Anteil zusammen. Der fixe Anteil c_{Fix} beschreibt die Kapazität an Personal, die benötigt wird, um das Rechenzentrum zu betreiben, selbst wenn keine Hardwareeinheit installiert ist. Der variable Anteil der Personalkapazität berechnet sich aus der Menge eingesetzter Hardwareeinheiten und einem

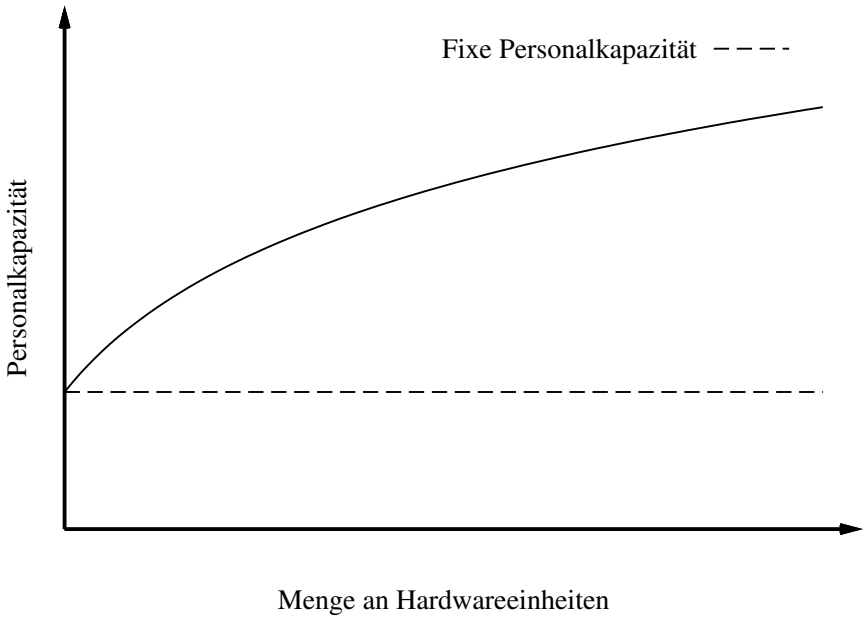

Abbildung 5.2: Entwicklung der Personalkapazität

Skalierungsfaktor κ. Abbildung 5.2 veranschaulicht die Entwicklung der Personalkapazität in Abhängigkeit der eingesetzten Betriebsmittel.
Die zur Verfügung gestellte Kapazität an Potentialfaktoren entspricht nicht dem tatsächlichen Bedarf. Um den Bedarf an Potentialfaktoren zu ermitteln, ist eine Bedarfsprognose durchzuführen.

5.2 Bedarfsprognose

Im Rahmen der Produktionsprogrammplanung werden Prognoseverfahren eingesetzt, um aus dem Verlauf des Vergangenheitsbedarfs eines Verbrauchsfaktors auf den zu erwartenden zukünftigen Bedarf zu schließen (vgl. Tempelmeier, 2006, S. 36). Bei regelmäßigem Bedarf eines Verbrauchsfaktors wird zwischen Prognoseverfahren für konstanten, trendförmigen und saisonal schwankenden Bedarf unterschieden.
Anstelle von Prognoseverfahren für konstanten Bedarf können immer Ver-

fahren für trendförmigen Bedarf angewendet werden. Für den Spezialfall des konstanten Bedarfs beträgt der Trend Null (vgl. Brown, 1984, S. 81). Der Einsatz von Prognoseverfahren für saisonal schwankenden Bedarf ist zur Bedarfsprognose innerhalb eines saisonalen Zyklus sinnvoll.
Die bekanntesten Verfahren für die Bedarfsprognose bei trendförmigem Bedarf sind die lineare Regressionsrechnung, die exponentielle Glättung zweiter Ordnung und das Verfahren von Holt. Die aufgeführten Prognosemodelle beruhen auf der Annahme, dass das Niveau der Bedarfszeitreihe einem linearen Trend folgt. Es gilt (vgl. Tempelmeier, 2006, S. 50):

$$y_k = b_0 + b_1 \cdot k + \epsilon_k \tag{5.16}$$

Der Achsenabschnitt der Trendgeraden beträgt b_0. Der Faktor b_1 beschreibt die Steigung der Trendgeraden. Die unabhängige Variable k repräsentiert den zeitlichen Verlauf. Zufällige Schwankungen werden durch ϵ_k abgebildet. Im Modell unterliegt ϵ_k einer Standardnormalverteilung mit dem Erwartungswert $E_\epsilon = 0$ (vgl. Makridakis et al., 1983, S. 220).
Für trendförmige Verläufe höherer Ordnung existieren weitere Prognoseverfahren. Dem Einsatz dieser Methoden ist jedoch eine Linearisierung der Bedarfsfunktion vorzuziehen (vgl. Brown, 1984, S. 83). Für einen exponentiellen Verlauf der Bedarfsfunktion ergibt sich folgende lineare Transformation (vgl. Weber, 1990, S. 70):

$$\begin{aligned} y_k &= b_0 \cdot e^{b_1 \cdot k} \\ \ln y_k &= \ln b_0 + b_1 \cdot k \cdot \ln e \\ \ln y_k &= b_0^* + b_1 \cdot k \qquad \text{mit} \qquad b_0^* = \ln b_0 \end{aligned} \tag{5.17}$$

5.2.1 Lineare Regressionsrechnung

Im Rahmen der linearen Regressionsrechnung wird versucht, mit Hilfe der Methode der kleinsten Quadrate eine lineare Trendfunktion zu finden, für die der Fehler ϵ_k minimal wird (vgl. Tempelmeier, 2006, S. 51 f.):

$$\begin{aligned} Q(b_0, b_1) &= \sum_{k=t-n+1}^{t} (y_k - b_0 - b_1 \cdot k)^2 \\ \frac{\delta Q(b_0, b_1)}{\delta b_0} &= -2 \cdot \sum_{k=t-n+1}^{t} y_k - b_0 - b_1 \cdot k \\ \frac{\delta Q(b_0, b_1)}{\delta b_1} &= -2 \cdot \sum_{k=t-n+1}^{t} k(y_k - b_0 - b_1 \cdot k) \end{aligned} \tag{5.18}$$

Aus den partiellen Ableitungen nach b_0 und b_1 ergeben sich folgende Normalgleichungen:

$$\begin{aligned} \sum_{k=t-n+1}^{t} y_k &= n \cdot b_0 + b_1 \cdot \sum_{k=t-n+1}^{t} k \\ \sum_{k=t-n+1}^{t} y_k \cdot k &= b_0 \cdot \sum_{k=t-n+1}^{t} k + b_1 \cdot \sum_{k=t-n+1}^{t} k^2 \end{aligned} \tag{5.19}$$

Die Normalgleichungen lassen sich wie folgt in Matrizenform darstellen (vgl. Neter et al., 1996, S. 200):

$$\begin{bmatrix} \sum_{k=t-n+1}^{t} y_k \\ \sum_{k=t-n+1}^{t} y_k \cdot k \end{bmatrix} = \begin{bmatrix} n & \cdot \sum_{k=t-n+1}^{t} k \\ \sum_{k=t-n+1}^{t} k & \sum_{k=t-n+1}^{t} k^2 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \tag{5.20}$$

$$(K^T \cdot Y) = (K^T \cdot K) \cdot b$$

Durch Ausmultiplizieren mit der inversen Matrix von $(K^T \cdot K)$ lässt sich das Gleichungssystem umformen:

$$b = (K^T \cdot K)^{-1} \cdot (K^T \cdot Y) \tag{5.21}$$

Sind die Parameter b_0 und b_1 der Trendgeraden bekannt, lässt sich der Bedarfswert $\hat{y}$ für die zukünftige Periode k ermitteln (vgl. Tempelmeier, 2006, S. 55):

$$\hat{y}_k = b_0 + b_1 \cdot k \tag{5.22}$$

Um den Grad der Anpassung der Trendgeraden zu bewerten, kann das korrigierte Bestimmtheitsmaß $\bar{R}^2$ verwendet werden. Das korrigierte Bestimmtheitsmaß $\bar{R}^2$ errechnet sich aus der Gesamtvariation und der nicht erklärten Restvariation und beträgt für zwei Freiheitsgrade der linearen Trendfunktion (vgl. DeLurgio, 1998, S. 104 f.):

$$\bar{R}^2 = 1 - \frac{\dfrac{\sum_{k=t-n+1}^{t} (y_k - \hat{y}_k)^2}{n-2}}{\dfrac{\sum_{k=t-n+1}^{t} (y_k - \bar{y})^2}{n-1}} \tag{5.23}$$

Die Variable $\bar{y}$ beschreibt den Mittelwert der Zeitreihe bezogen auf die letzten n Perioden. Eine perfekte Anpassung der Trendgeraden wird durch $\bar{R}^2 = 1$ beschrieben. Strebt der Anstieg der Trendgeraden gegen Null, sind negative Werte für das korrigierte Bestimmtheitsmaß möglich.

5.2.2 Exponentielle Glättung zweiter Ordnung

Bei einem angenommenen konstanten Bedarf lässt sich für die Periode t der gleitende Durchschnitt des Bedarfs S_t wie folgt beschreiben (vgl. Brown und Meyer, 1960, S. 677):

$$S'_t = \alpha \cdot y_t + (1 - \alpha) \cdot S'_{t-1} \qquad \text{mit} \qquad 0 < \alpha < 1 \tag{5.24}$$

Der gleitende Durchschnitt wird als Mittelwert erster Ordnung bezeichnet. Das Ausmaß, mit dem sich ein aufgetretener Prognosefehler auf die nächste Periode auswirkt, wird durch den Parameter α bestimmt. Je größer der Parameter α ist, um so stärker ist der Einfluss jüngerer Beobachtungen (vgl. Tempelmeier, 2006, S. 47 f.).
Der Mittelwert erster Ordnung lässt sich ebenfalls wieder exponentiell glätten und liefert den Mittelwert zweiter Ordnung:

$$S''_t = \alpha \cdot S'_t + (1 - \alpha) \cdot S''_{t-1} \qquad \text{mit} \qquad 0 < \alpha < 1 \tag{5.25}$$

Wird der Mittelwert erster Ordnung S_t als Prognosewert für den Bedarf bei einem vorliegenden linearen Trend betrachtet, tritt bei der Prognose ein systematischer Fehler auf. Der Mittelwert zweiter Ordnung wird zur Korrektur des auftretenden systematischen Fehlers eingesetzt. Für die Prognose des Bedarfs $\hat{y}$ in der Periode $t + j$ gilt (vgl. Gardner, 1984, S. 48):

$$\hat{y}_{t+j} = (2 \cdot S'_t - S''_t) + j \cdot (S'_t - S''_t) \cdot \frac{\alpha}{1 - \alpha} \tag{5.26}$$

Die Berechnung der Mittelwerte ist ein iterativer Prozess. Der Achsenabschnitt und der Anstieg der Trendgeraden der Startwerte der gleitenden

Durchschnitte lassen sich durch lineare Regressionsrechnung abschätzen. Für die Startwerte gilt (vgl. Brown und Meyer, 1960, S. 679):

$$\begin{aligned} S'_0 &= b_{0,0} - \frac{1-\alpha}{\alpha} \cdot b_{1,0} \\ S''_0 &= b_{0,0} - 2 \cdot \frac{1-\alpha}{\alpha} \cdot b_{1,0} \end{aligned} \tag{5.27}$$

Das Verfahren der exponentiellen Glättung zweiter Ordnung reagiert dynamischer auf aktuelle Änderungen im Trendverlauf als die lineare Regressionsrechnung. Zur Erreichung zufriedenstellender Prognoseergebnisse sind jedoch geeignete Werte für den Glättungsparameter α zu schätzen. Zur Schätzung des Glättungsparameters erster Ordnung bei n vorliegenden Beobachtungen wird folgende Regel vorgeschlagen (vgl. DeLurgio, 1998, S. 156):

$$\alpha_1 = \frac{2}{n+1} \tag{5.28}$$

Der Glättungsparameter m-ter Ordnung steht zum Glättungsparameter erster Ordnung in folgender Beziehung (vgl. Brown und Meyer, 1960, S. 680):

$$1 - \alpha_1 = (1 - \alpha_m)^m \tag{5.29}$$

Für den Glättungsparameter zweiter Ordnung gilt somit:

$$\begin{aligned} \alpha_2 &= 1 - \sqrt{1 - \alpha_1} \\ &= 1 - \sqrt{1 - \frac{2}{n+1}} \end{aligned} \tag{5.30}$$

5.2.3 Verfahren von Holt

Das Verfahren von Holt verwendet einen zusätzlichen Glättungsparameter zur Bestimmung des Anstiegs der Trendgeraden (vgl. Holt, 2004, S. 8). Achsenabschnitt und Anstieg der Trendgeraden werden getrennt voneinander einer exponentiellen Glättung erster Ordnung unterzogen. Der Achsenabschnitt der Trendgeraden lässt sich wie folgt prognostizieren (vgl. Tempelmeier, 2006, S. 68):

$$b_{0,t} = \alpha \cdot y_t + (1 - \alpha)(b_{0,t-1} + b_{1,t-1}) \tag{5.31}$$

Für den Anstieg der Trendgeraden gilt:

$$b_{1,t} = \beta(b_{0,t} - b_{0,t-1}) + (1 - \beta)b_{1,t-1} \tag{5.32}$$

Die Prognose des Bedarfs $\hat{y}$ in der Periode $t+j$ lässt sich wie folgt berechnen:

$$\hat{y}_{t+j} = b_{0,t} + b_{1,t} \cdot j \tag{5.33}$$

Bei geeigneter Wahl der Glättungsparameter liefert das Verfahren von Holt bessere Ergebnisse als die exponentielle Glättung zweiter Ordnung. Die Schätzung zweier unabhängiger Glättungparameter ist jedoch komplexer als die Schätzung eines Parameters.

5.2.4 Bewertung der Prognoseergebnisse

Zur Bewertung der Prognoseergebnisse werden die prognostizierten Werte mit den Beobachtungswerten verglichen. Das Residuum ϵ beschreibt die Dif-

ferenz zwischen prognostiziertem Wert $\hat{y}$ und Beobachtungswert y der Periode t (vgl. Neter et al., 1996, S. 97):

$$\epsilon_t = y_t - \hat{y}_t \tag{5.34}$$

Anhand der Streuung der Residualwerte ist eine Bewertung der Prognoseergebnisse möglich. Zur Beurteilung der Streuung der Residualwerte kann die Standardabweichung des Residuums herangezogen werden. Die Varianz σ^2 des Residuums ϵ beträgt (vgl. Tempelmeier, 2006, S. 35):

$$\sigma_{\epsilon t}^2 = \frac{1}{n-1} \cdot \sum_{k=t-n+1}^{t} (\epsilon_k - \bar{\epsilon})^2 \tag{5.35}$$

Der Mittelwert des Residuums $\bar{\epsilon}$ berechnet sich wie folgt:

$$\bar{\epsilon} = \frac{1}{n} \cdot \sum_{k=t-n+1}^{t} (\epsilon_k) \tag{5.36}$$

Die Standardabweichung des Residuums ϵ ist somit:

$$\sigma_{\epsilon t} = \sqrt{\sigma_{\epsilon t}^2} \tag{5.37}$$

Je geringer die Werte der Standardabweichung des Residuums für ein Prognoseverfahren sind, um so besser ist das Prognoseergebnis.

5.3 Bedarfsprognose für Potentialfaktoren

Im Rahmen der Produktionsplanung werden Prognoseverfahren eingesetzt, um aus dem zurückliegenden Bedarf an Verbrauchsfaktoren den zukünftigen Verbrauch zu extrapolieren. Diese Methode der Produktionsprogrammplanung soll im Rechenzentrum für den Bedarf an physischen Potentialfaktoren adaptiert werden. Die Anwendung dieser Methode wird zur Schätzung des durchschnittlichen Kapazitätsbedarfs der Hardwareeinheiten eingesetzt.
Die Anwendung von Prognoseverfahren erscheint sinnvoll, da der zurückliegende Bedarf an Kapazitäten durch entsprechende Werkzeuge erhoben werden kann. In Abschnitt 4.1.3 wurde aufgezeigt, dass es nicht sinnvoll ist, die Betriebsmittel voll auszulasten. Wird ein Potentialfaktor nicht voll ausgelastet, dann entspricht die Auslastung auch dem Bedarf. Die Entwicklung der durchschnittlichen Auslastung kann damit als Datenbasis für die Prognoseverfahren eingesetzt werden. Die Prognosen werden für jede Hardwarekomponente der Hardwareeinheiten getrennt durchgeführt.
Eine Prognose für den Bedarf an Personal erscheint nicht sinnvoll, da diese Kapazitäten üblicherweise voll ausgelastet sind. Die Prognosen des Bedarfs an Hardwareeinheiten können jedoch als Grundlage für die Personalplanung dienen.
Durch entsprechende Messwerkzeuge wird für jede Hardwarekomponente je Hardwareeinheit die durchschnittliche Auslastung ermittelt. Aus der durchschnittlichen Auslastung ρ der Hardwarekomponente einer Hardwareeinheit i und deren Kapazitätsangebot c_A kann nach Formel 4.36 der durchschnittliche Kapazitätsbedarf c_B ermittelt werden:

$$c_{Bi} = \rho_i \cdot c_{Ai} \tag{5.38}$$

Der durchschnittliche Kapazitätsbedarf c_B einer Hardwarekomponente ergibt sich aus der Summe aller Einzelbedarfe:

$$c_B = \sum_{i=1}^{n} c_{Bi} \tag{5.39}$$

Periode t	**Kapazität c_t in MIPS**	$ln\ c_t$
1	95200	11,4637
2	90800	11,4164
3	151600	11,9290
4	249600	12,4276
5	257800	12,4599
6	502200	13,1268
7	713500	13,4779
8	836900	13,6375

Tabelle 5.1: Messung des durchschnittlichen Bedarfs an Prozessorleistung

Als Beispieldaten dienen die in Tabelle 5.1 aufgeführten durchschnittlichen Kapazitätsbedarfe an Prozessorleistung für ein Rechenzentrum in den vergangenen acht Perioden. Der Kapazitätsbedarf physischer Potentialfaktoren unterliegt auch im Rechenzentrum saisonalen Schwankungen. Da die Hardwareeinheiten jedoch nicht saisonabhängig auf- und abgebaut werden, wird eine Prognoseperiode so dimensioniert, dass sie einen kompletten saisonalen Zyklus umfasst. Da das Wachstum der Kapazitätsbedarfe einem exponentiellen Trend folgt und saisonale Schwankungen nicht berücksichtigt werden müssen, können die in Abschnitt 5.2 beschriebenen Verfahren angewendet werden.
In den folgenden Unterabschnitten werden eine lineare Regressionsrechnung, eine exponentielle Glättung zweiter Ordnung sowie das Verfahren von Holt zur Prognose des durchschnittlichen Kapazitätsbedarfs an Prozessorleistung durchgeführt. Die Entwicklung des Bedarfs folgt einem exponentiellen Trend. Um die beschriebenen Prognoseverfahren anwenden zu können, ist eine Linearisierung der exponentiellen Werte nach Formel 5.17 vorzunehmen. Die Ergebnisse der Linearisierung sind ebenfalls in Tabelle 5.1 aufgeführt. Abbildung 5.3 stellt die Entwicklung des durchschnittlichen Kapazitätsbedarfs an Prozessorleistung für das Rechenzentrum und das Ergebnis der Linearisierung grafisch dar.

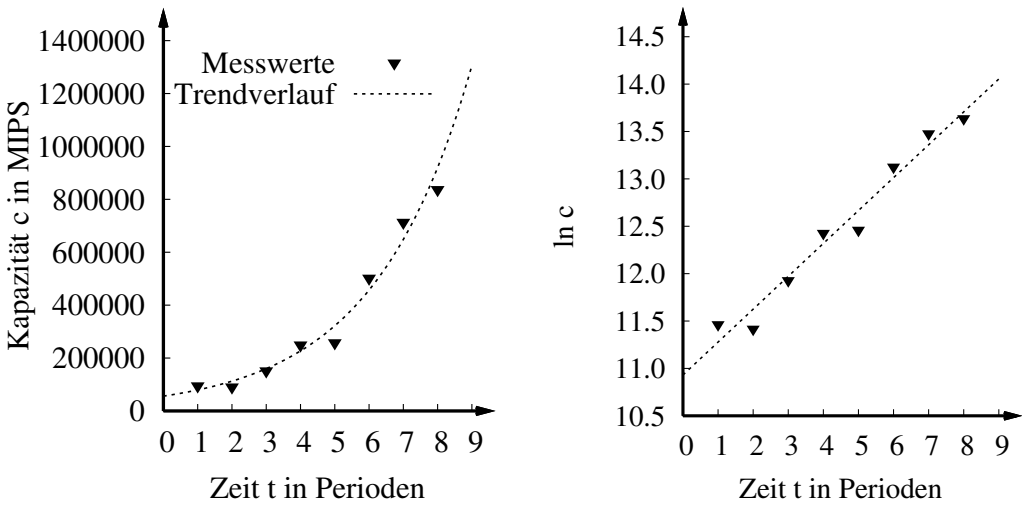

Abbildung 5.3: Messwerte der Prozessorkapazität

5.3.1 Lineare Regressionsrechnung

Zur Bestimmung des Achsenabschnitts b_0 und des Anstiegs der Trendgeraden b_1 wird Formel 5.21 angewendet. Die Matrizen K und Y sind wie folgt definiert:

$$K = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \end{pmatrix} \qquad Y = \begin{pmatrix} 11,4637 \\ 11,4164 \\ 11,9290 \\ 12,4276 \\ 12,4599 \\ 13,1268 \\ 13,4779 \\ 13,6375 \end{pmatrix} \tag{5.40}$$

Für die Matrizen $K^T \cdot K$ und $K^T \cdot Y$ ergeben sich folgende Werte:

$$K^T \cdot K = \begin{pmatrix} 8 & 36 \\ 36 & 204 \end{pmatrix} \qquad K^T \cdot Y = \begin{pmatrix} 99,9389 \\ 464,2995 \end{pmatrix} \tag{5.41}$$

Die inverse Matrix von $K^T \cdot K$ ist:

$$(K^T \cdot K)^{-1} = \begin{pmatrix} 0,6071 & -0,1071 \\ -0,1071 & 0,0238 \end{pmatrix} \tag{5.42}$$

Der Achsenabschnitt und der Anstieg der Trendgeraden sind:

$$b = \begin{pmatrix} 10,9308 \\ 0,3470 \end{pmatrix} \tag{5.43}$$

Die Trendgleichung 5.22 ist somit wie folgt definiert:

$$\begin{aligned} \hat{y}_k &= b_0 + b_1 \cdot k \\ \hat{y}_k &= 10,9308 + 0,347 \cdot k \end{aligned} \tag{5.44}$$

Der Prognosewert für den linearisierten durchschnittlichen Bedarf an Prozessorkapazität für das Rechenzentrum $\hat{y}$ in der Periode 9 beträgt:

$$\begin{aligned} \hat{y}_9 &= 10,9308 + 0,347 \cdot 9 \\ &= 14,0539 \end{aligned} \tag{5.45}$$

Der durchschnittliche Bedarf an Prozessorkapazität $\hat{c}$ in Periode 9 ist:

$$\begin{aligned} \hat{c}_9 &= e^{14,0539} \\ &\approx 1269000 \text{ MIPS} \end{aligned} \tag{5.46}$$

Das korrigierte Bestimmtheitsmaß $\bar{R}^2$ liefert für die Trendgerade den Wert:

$$\bar{R}^2 = 0,963 \tag{5.47}$$

Es kann davon ausgegangen werden, dass die Trendgerade an die vorliegenden Beobachtungswerte angepasst ist.

5.3.2 Exponentielle Glättung zweiter Ordnung

Für die Durchführung einer exponentiellen Glättung zweiter Ordnung ist eine Schätzung des Glättungsparameters α vorzunehmen. Für die vorliegenden Beobachtungswerte ergibt sich nach Formel 5.30:

$$\begin{aligned} \alpha &= 1 - \sqrt{1 - \frac{2}{n+1}} \\ &= 1 - \sqrt{1 - \frac{2}{8+1}} \\ &= 0,1181 \end{aligned} \tag{5.48}$$

Zur Initialisierung der gleitenden Durchschnitte werden die Ergebnisse der linearen Regressionsrechnung zur Bestimmung des Achsenabschnitts und der Trendgeraden aus Gleichung 5.43 verwendet. Die Startwerte betragen nach Formel 5.27:

$$\begin{aligned} S'_0 &= b_{0,0} - \frac{1-\alpha}{\alpha} \cdot b_{1,0} \\ &= 10,9308 - \frac{1-0,1181}{0,1181} \cdot 0,347 \\ &= 8,3391 \\ S''_0 &= b_{0,0} - 2 \cdot \frac{1-\alpha}{\alpha} \cdot b_{1,0} \\ &= 10,9308 - 2 \cdot \frac{1-0,1181}{0,1181} \cdot 0,347 \\ &= 5,7473 \end{aligned} \tag{5.49}$$

Periode t	S'_t	S''_t
1	8,7080	6,0969
2	9,0278	6,4430
3	9,3704	6,7887
4	9,7314	7,1362
5	10,0536	7,4807
6	10,4165	7,8274
7	10,7780	8,1758
8	11,1157	8,5229

Tabelle 5.2: Gleitende Durchschnitte

Der gleitende Durchschnitt erster Ordnung ist für Periode 1 nach Formel 5.24:

$$\begin{aligned} S'_1 &= \alpha \cdot y_1 + (1-\alpha) \cdot S'_0 \\ &= 0,1181 \cdot 11,4637 + (1-0,1181) \cdot 8,3391 \\ &= 8,7080 \end{aligned} \tag{5.50}$$

Der gleitende Durchschnitt zweiter Ordnung in Periode 1 beträgt nach Formel 5.25:

$$\begin{aligned} S''_1 &= \alpha \cdot S'_1 + (1-\alpha) \cdot S''_0 \\ &= 0,1181 \cdot 8,7080 + (1-0,1181) \cdot 5,7473 \\ &= 6,0969 \end{aligned} \tag{5.51}$$

Tabelle 5.2 zeigt die Ergebnisse der Bildung der gleitenden Durchschnitte für die Folgeperioden. Der Prognosewert $\hat{y}$ für den linearisierten durchschnittlichen Bedarf an Prozessorkapazität für das Rechenzentrum in der Periode 9 beträgt nach Formel 5.26:

$$\begin{aligned} \hat{y}_9 &= (2 \cdot S'_8 - S''_8) + 1 \cdot (S'_8 - S''_8) \cdot \frac{\alpha}{1-\alpha} \\ &= (2 \cdot 11,1157 - 8,5229) \\ &\quad +1 \cdot (11,1157 - 8,5229) \cdot \frac{0,1181}{1-0,1181} \\ &= 14,0555 \end{aligned} \tag{5.52}$$

Der durchschnittliche Bedarf an Prozessorkapazität $\hat{c}$ in Periode 9 ist:

$$\begin{aligned} \hat{c}_9 &= e^{\hat{y}_9} \\ &= e^{14,0555} \\ &\approx 1270000 \text{ MIPS} \end{aligned} \tag{5.53}$$

5.3.3 Verfahren von Holt

Für die Umsetzung des Verfahrens von Holt ist eine Schätzung der Glättungsparameter vorzunehmen. Der Glättungsparameter α wird von der exponentiellen Glättung übernommen:

$$\alpha = 0,1181 \tag{5.54}$$

Der Glättungsparameter β wird vereinfachend als Glättungsparameter der nächsthöheren Ordnung betrachtet und nach Formel 5.29 geschätzt:

$$\begin{aligned} \beta &= 1 - \sqrt[3]{1 - \frac{2}{8+1}} \\ &= 0,0804 \end{aligned} \tag{5.55}$$

Als Startparameter für den Achsenabschnitt b_0 und den Anstieg der Trendgeraden b_1 werden wieder die Ergebnisse der linearen Regressionsrechnung verwendet. Nach Formel 5.31 beträgt der Achsenabschnitt in Periode 1:

$$\begin{aligned} b_{0,1} &= \alpha \cdot y_1 + (1-\alpha)(b_{0,0} + b_{1,0}) \\ &= 0,1181 \cdot 11,4637 + (1 - 0,1181)(10,9308 + 0,347) \\ &= 11,2998 \end{aligned} \tag{5.56}$$

Periode t	$b_{0,t}$	$b_{1,t}$
1	11,2998	0,3488
2	11,6211	0,3466
3	11,9631	0,3462
4	12,3233	0,3473
5	12,6458	0,3453
6	13,0071	0,3466
7	13,3684	0,3478
8	13,7069	0,3471

Tabelle 5.3: Prognoseergebnisse für das Verfahren von Holt

Der Anstieg der Trendgeraden in Periode 1 ist nach Formel 5.32:

$$\begin{aligned} b_{1,1} &= \beta(b_{0,1} - b_{0,0}) + (1-\beta)b_{1,0} \\ &= 0,0804 \cdot (11,2998 - 10,9308) + (1 - 0,0804) \cdot 0,347 \\ &= 0,3488 \end{aligned} \tag{5.57}$$

Die Ergebnisse der Berechnung des Achsenabschnitts und des Anstiegs der Trendgeraden für die Folgeperioden sind in Tabelle 5.3 aufgeführt. Der Prognosewert $\hat{y}$ für den linearisierten durchschnittlichen Bedarf an Prozessorkapazität für das Rechenzentrum in der Periode 9 beträgt nach Formel 5.33:

$$\begin{aligned} \hat{y}_9 &= b_{0,8} + b_{1,8} \cdot 1 \\ &= 13,7069 + 0,3471 \cdot 1 \\ &= 14,054 \end{aligned} \tag{5.58}$$

Der durchschnittliche Bedarf an Prozessorkapazität $\hat{c}$ in Periode 9 ist:

$$\begin{aligned} \hat{c}_9 &= e^{\hat{y}_9} \\ &= e^{14,054} \\ &\approx 1269000 \text{ MIPS} \end{aligned} \tag{5.59}$$

			Lineare Regressions-rechnung		**Exponentielle Glättung zweiter Ordnung**		**Verfahren von Holt**	
t	α	β	$\hat{y}_t$	ϵ_t	$\hat{y}_t$	ϵ_t	$\hat{y}_t$	ϵ_t
5	0,2254	0,1566	12,6602	-0,2003	12,6706	-0,2107	12,6603	-0,2003
6	0,1835	0,1264	12,8404	0,2863	12,8408	0,2860	12,8405	0,2862
7	0,1548	0,1061	13,3317	0,1463	13,3375	0,1404	13,3316	0,1463
8	0,1340	0,0914	13,7565	-0,1191	13,7625	-0,1250	13,7564	-0,1190
$\bar{\epsilon}$				0,0283		0,0227		0,0283
$\sigma_{\epsilon t}$				0,2269		0,2306		0,2268

Tabelle 5.4: Residualwerte der Prognoseverfahren

5.3.4 Bewertung der Prognoseergebnisse

Zur Bewertung der Prognoseergebnisse werden die prognostizierten Werte der letzten vier Perioden des jeweiligen Prognoseverfahrens mit den entsprechenden Beobachtungswerten verglichen. Die Beobachtungswerte der ersten vier Perioden dienen ausschließlich als Datenbasis. Die Prognose einer Periode wird nur anhand der Beobachtungswerte der Vorperioden durchgeführt. In Tabelle 5.4 sind die Residualwerte der Prognoseverfahren nach Formel 5.34 der letzten vier Perioden aufgeführt. Die Ex-Post-Analyse liefert für die Standardabweichung des Residuums $\sigma_{\epsilon t}$ nach Formel 5.37 die ebenfalls in Tabelle 5.4 aufgeführten Werte. Das Verfahren von Holt besitzt den niedrigsten Wert für die Standardabweichung des Residuums und liefert für die vorliegenden Beobachtungswerte mit den gegebenen Glättungsparametern die besten Prognoseergebnisse.

5.4 Zusammenfassung

Der Kapazitätsbedarf an Potentialfaktoren wächst auch bei konstanter Nachfrage nach IT-Dienstleistungen exponentiell. Die Hypothese konnte für die Kapazität der Hardwarekomponenten durch ein mathematisch-formales

Modell bestätigt werden. Der Kapazitätsbedarf des Personals steht in Beziehung zur Menge der eingesetzten Potentialfaktoren und folgt nicht dem exponentiellen Trend.
Die Kapazitätsbedarfsplanung prognostiziert auf taktischer Planungsebene den Kapazitätsbedarf an Betriebsmitteln für die kommende Planungsperiode. Ist der Kapazitätsbedarf bekannt, kann auf Basis der Kapazitätsangebotsplanung ermittelt werden, welche Menge an Betriebsmitteln mit welchem technischen Leistungsvermögen benötigt wird. Auf operativer Planungsebene ist die Frage zu klären, welche virtuellen Betriebsmittel auf welchen physischen Betriebsmitteln eingerichtet und welche Fertigungsaufträge auf welchen Betriebsmitteln ausgeführt werden. Im folgenden Kapitel werden Lösungsansätze für eine automatisierte Betriebsmittelzuordnung aufgezeigt.

Kapitel 6

Zuordnung der Betriebsmittel

In Kapitel 3 führte die die Evaluation des Referenzmodells zur Produktion von IT-Dienstleistungen anhand eines Prototypen zu der Erkenntnis, dass die Zuordnung der Betriebsmittel nicht durch standardisierte Methoden der Produktionsplanung und -steuerung automatisiert durchgeführt werden kann. Die Automatisierung der Betriebsmittelzuordnung wird in diesem Kapitel als eigene Forschungsfrage untersucht. Im ersten Schritt soll folgende Hypothese überprüft werden:

> Aus Sicht der Kapazitätswirtschaft ist es stets sinnvoll, die virtualisierten Betriebsmittel auf Vorrat bereitzustellen.

Bei der Bestätigung der Hypothese wird von der Bedingung ausgegangen, dass der Einsatz von Virtualisierungstechniken aus Sicht der Kapazitätswirtschaft sinnvoll ist. Die Gültigkeit dieser Bedingung wurde bereits in Abschnitt 4.2 nachgewiesen. Zur Darstellung der Betriebsmittelzuordnung wird ein mathematisch-formales Modell entwickelt.
In Kapitel 4 wurde die Planung des Kapazitätsangebots der Betriebsmittel zur Erbringung einer IT-Dienstleistung beschrieben. Die Prognose des Kapazitätsbedarfs an Betriebsmitteln wurde in Kapitel 5 dargestellt. Anhand der Ergebnisse werden die physischen Betriebsmittel beschafft. Es wird untersucht, auf welchen physischen Betriebsmitteln welche virtuellen Betriebsmittel auf Grundlage der Kapazitätsangebotsplanung eingerichtet werden.
Im Rahmen der Fertigungssteuerung sind vor der Auftragsfreigabe die in

den Arbeitsplänen verwendeten Planbetriebsmittel durch tatsächliche virtualisierte Betriebsmittel zu substituieren (siehe Abschnitt 3.1.2). Es wird aufgezeigt, wie die Zuordnung der Betriebsmittel automatisiert erfolgen kann.

6.1 Zuordnung virtualisierter Betriebsmittel

Zur Umsetzung der Fertigungsaufträge werden virtualisierte Betriebsmittel benötigt. Diese virtuellen Betriebsmittel werden auf physischen Betriebsmitteln eingerichtet. Die Zuordnung der Betriebsmittel lässt sich als Bin-Packing-Problem beschreiben (vgl. Bichler et al., 2006, S. 2). Heuristiken zur Lösung des Bin-Packing-Problems lassen sich auf die Betriebsmittelzuordung anwenden.

6.1.1 Bin-Packing

Gegeben sind n Objekte verschiedener fester Größe sowie eine Menge von Behältern gleicher Größe. Ziel ist es, die Objekte den Behältern so zuzuordnen, dass möglichst wenig Behälter verwendet werden. Diese Problemstellung wird als Bin-Packing-Problem bezeichnet (vgl. Korte und Vygen, 2008, S. 485).
Erfüllt eine Lösung x aus der Menge der zulässigen Lösungen M eine gegebene Bedingung α, lässt sich die Suche nach dieser Lösung als Entscheidungsproblem formulieren:

$$\text{Existiert zu einem Wert } \alpha \text{ ein } x \in M \text{ mit } f(x) \leq \alpha \ ? \tag{6.1}$$

Ein solches Entscheidungsproblem wird Zielfunktionsseparierungsproblem genannt (vgl. Dempe und Schreier, 2006, S. 350). Die Klasse der Entscheidungsprobleme, für deren Lösung ein nichtdeterministisch polynomialer Algorithmus existiert, wird als NP bezeichnet (vgl. Dempe und Schreier, 2006,

S. 362). Existiert für die Lösung eines Entscheidungsproblems E ein nichtdeterministisch polynomialer Algorithmus ($E \in NP$) und lässt sich jedes andere Entscheidungsproblem der Klasse NP polynomial auf das Problem E transformieren, ist dieses Problem ein NP-vollständiges Entscheidungsproblem (vgl. Dempe und Schreier, 2006, S. 366). Ein Optimierungsproblem ist NP-schwer, wenn dessen Zielfunktionsseparierungsproblem NP-vollständig ist (vgl. Dempe und Schreier, 2006, S. 371). Das Bin-Packing-Problem ist ein NP-schweres kombinatorisches Optimierungsproblem (vgl. Coffman et al., 1997, S. 46). Zur Lösung des Bin-Packing-Problems werden Heuristiken eingesetzt.
Um das Bin-Packing-Problem zu untersuchen, wird es formalisiert (vgl. Coffman et al., 1997, S. 46 f.). Gegeben ist eine Liste L mit n Gegenständen a_i:

$$L = (a_1, a_2, ..., a_n) \tag{6.2}$$

Jeder Gegenstand a_i hat eine bestimmte Größe s. Vereinfachend wird angenommen, dass sich die Größe s der Gegenstände im linksoffenen Intervall von Null bis Eins befindet:

$$s(a_i) \in (0, 1] \tag{6.3}$$

Zum Verpacken der Gegenstände a_i stehen Behälter B_j zur Verfügung. Jeder Behälter B_j hat die Größe Eins:

$$s(B_j) = 1 \tag{6.4}$$

Der Füllstand v eines Behälters B_j berechnet sich aus der Summe der Größen der Gegenstände, die sich in dem Behälter befinden:

$$v(B_j) = \sum_{a_i \in B_j} s(a_i) \quad \text{mit} \quad 1 \leq j \tag{6.5}$$

Die Gegenstände a_i sollen in möglichst wenige Behälter B gepackt werden. Dabei darf die Summe der Größen s der Gegenstände a_i in einem Behälter nicht Größe des Behälters B_j übersteigen:

$$\sum_{a_i \in B_j} s(a_i) \leq s(B_j) \quad \text{mit} \quad 1 \leq j \tag{6.6}$$

Unter Berücksichtigung der Restriktion 6.4 gilt:

$$\sum_{a_i \in B_j} s(a_i) \leq 1 \quad \text{mit} \quad 1 \leq j \tag{6.7}$$

Zum Verpacken der Gegenstände a_i der Liste L wird ein Algorithmus A angewendet. Die vom Algorithmus A benötigte Menge an Behältern B sei $A(L)$. Die optimale Menge an Behältern B zum Verpacken der Gegenstände a_i der Liste L beträgt $OPT(L)$. Es gilt folgende Bedingung:

$$OPT(L) \leq A(L) \tag{6.8}$$

Die Bewertung des Algorithmus A erfolgt durch einen kompetitiven Faktor. Der Algorithmus A sei n-kompetitiv, wenn es einen Wert m gibt, so dass für jede Liste L gilt (vgl. Klein, 2005, S. 334 f.):

$$A(L) \leq n \cdot OPT(L) + m \quad \text{mit} \quad n, m \in \mathbb{R} \tag{6.9}$$

Die Berechnungskomplexität des Bin-Packing-Problems wird anhand der Zeitkomplexität T bestimmt. Die Zeitkomplexität beschreibt die Anzahl von Rechenoperationen, die zur Lösung des Problems notwendig sind (vgl. Hromkovič, 2007, S. 207). Die Darstellung der Zeitkomplexität erfolgt in der Ω- und der O-Notation (vgl. Hromkovič, 2007, S. 211):

$$O(f(n)) = \{r : \mathbb{N} \to \mathbb{R}^+ \mid \exists n_0 \in \mathbb{N}, \exists c \in \mathbb{N}, \forall n \geq n_0 : r(n) \leq c \cdot f(n)\} \tag{6.10}$$

Jede Funktion $r \in O(f(n))$ wächst asymptotisch nicht schneller als die Funktion f.

$$\Omega(g(n)) = \{s : \mathbb{N} \to \mathbb{R}^+ \mid \exists n_0 \in \mathbb{N}, \exists d \in \mathbb{N}, \forall n \geq n_0 : s(n) \geq \frac{1}{d} \cdot g(n)\} \tag{6.11}$$

Jede Funktion $s \in \Omega(g(n))$ wächst asymptotisch mindestens so schnell wie die Funktion g.
Bin-Packing-Probleme lassen sich in Online- und Offline-Bin-Packing-Probleme unterteilen. Beim Offline-Bin-Packing sind die Mengen an Objekten und Behältern vor dem Beginn der Lösung des Problems bekannt.

In der Realität kann nicht immer von dieser Voraussetzung ausgegangen werden. Treffen die Objekte in einer geordneten Reihenfolge ein und sind unmittelbar einem Behälter zuzuordnen, wird diese Problemstellung als Online-Bin-Packing bezeichnet (vgl. Coffman et al., 1997, S. 47).
Bin-Packing-Probleme lassen sich weiterhin nach der Anzahl der Kapazitätsrestriktionen unterscheiden. Verfügen die Behälter nur über eine Kapazitätsrestriktion, liegt ein eindimensionales Bin-Packing-Problem vor. Werden die Behälter durch mehrere Kapazitätsrestriktionen limitiert, handelt es sich um ein mehrdimensionales Bin-Packing-Problem.

Eindimensionales Online-Bin-Packing

Beim Online-Bin-Packing werden die Gegenstände $a_1, ..., a_k$ der Liste L nacheinander verpackt. Der Algorithmus hat keine Kenntnis über den Inhalt der Liste sondern berücksichtigt immer nur den gerade zu verpackenden Gegenstand a_j.
Der einfachste Algorithmus zur Lösung des Online-Bin-Packing-Problems ist der Next-Fit-Algorithmus (NF). Hierbei wird immer nur ein aktueller Behälter B_m betrachtet. Der aktuelle Behälter B_m ist entweder leer oder bereits mit den Objekten $a_i \in B_m$ gefüllt. Der Gegenstand a_j wird in den Behälter gepackt wenn gilt:

$$s(a_j) + \sum_{a_i \in B_m} s(a_i) \leq 1 \tag{6.12}$$

Ist diese Bedingung nicht erfüllt, wird der Behälter B_m geschlossen und Gegenstand a_j wird in den Behälter B_{m+1} gepackt. Der Behälter B_m wird nicht weiter betrachtet. Dieser Vorgang wird solange wiederholt bis alle k Gegenstände verpackt sind.
Der Next-Fit-Algorithmus ist 2-kompetitiv. Es lässt sich zeigen, dass gilt (vgl. Korte und Vygen, 2008, S. 488):

$$NF(L) \leq 2 \cdot OPT(L) - 1 \tag{6.13}$$

Betrachtet man zwei benachbarte, durch den Next-Fit-Algorithmus befüllte Behälter B_m und B_{m+1} so gilt:

$$v(B_m) + v(B_{m+1}) > 1 \tag{6.14}$$

Die Summe der Füllstände zweier benachbarter Behälter muss immer größer als Eins sein, da der Inhalt des zweiten Behälters sonst noch in den ersten Behälter gepasst hätte. Für Summe der Füllstände aller vom Next-Fit-Algorithmus verwendeten Behälter gilt somit:

$$\left\lfloor \frac{NF(L)}{2} \right\rfloor < \sum_{m=1}^{NF(L)} v(B_m) \tag{6.15}$$

Die linke Seite der Ungleichung wird für eine ungerade Anzahl Behälter abgerundet. Die Summe der Füllstände ist sowohl für einen optimalen Algorithmus als auch für den Next-Fit-Algorithmus identisch, wenn beide auf dieselbe Liste mit Gegenständen angewendet werden:

$$\sum_{m=1}^{NF(L)} v(B_m) = \sum_{m=1}^{OPT(L)} v(B_m) \tag{6.16}$$

Aus Formel 6.15 folgt deshalb:

$$\left\lfloor \frac{NF(L)}{2} \right\rfloor < \sum_{m=1}^{OPT(L)} v(B_m) \tag{6.17}$$

Die Ungleichung lässt sich erweitern und umformen:

$$\begin{aligned} \frac{NF(L)-1}{2} \leq \left\lfloor \frac{NF(L)}{2} \right\rfloor &\leq \left\lceil \sum_{m=1}^{OPT(L)} v(B_m) \right\rceil - 1 \\ NF(L) - 1 &\leq 2 \cdot \left\lceil \sum_{m=1}^{OPT(L)} v(B_m) \right\rceil - 2 \\ NF(L) &\leq 2 \cdot \left\lceil \sum_{m=1}^{OPT(L)} v(B_m) \right\rceil - 1 \end{aligned} \tag{6.18}$$

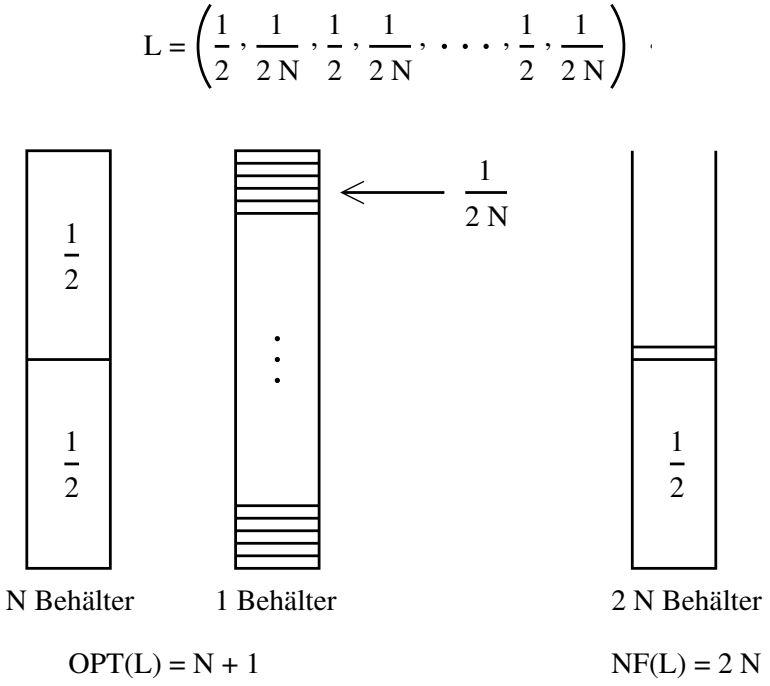

Abbildung 6.1: Beispiel für den Next-Fit-Algorithmus (nach Coffman et al., 1997, S. 49)

Der Füllstand eines Behälters darf die Größe des Behälters niemals überschreiten. Deshalb gilt:

$$v(B_m) \leq 1$$

$$NF(L) \leq 2 \cdot \left\lceil \sum_{m=1}^{OPT(L)} v(B_m) \right\rceil - 1 \leq 2 \cdot \left\lceil \sum_{m=1}^{OPT(L)} 1 \right\rceil - 1 \qquad (6.19)$$

$$NF(L) \leq 2 \cdot OPT(L) - 1 \qquad \text{q. e. d.}$$

Abbildung 6.1 zeigt ein Beispiel für eine Liste von Gegenständen, die durch den Next-Fit-Algorithmus verpackt wurden. Die Zeitkomplexität T_{NF} des Next-Fit-Algorithmus ist linear (vgl. Korte und Vygen, 2008, S. 488):

$$T_{NF} = O(n) \qquad (6.20)$$

Ein weiterer Algorithmus zur Lösung des Online-Bin-Packing-Problems ist der First-Fit-Algorithmus (FF) (vgl. Garey et al., 1972, S. 143). Ein Behälter B_m ist entweder leer oder ist bereits mit den Objekten $a_i \in B_m$ gefüllt. Beginnend mit dem ersten Behälter wird versucht, den Gegenstand a_j zu verpacken. Der Gegenstand a_j wird in den Behälter m gepackt wenn gilt:

$$\begin{aligned} s(a_j) + v(B_m) &\leq 1 \\ s(a_j) + \textstyle\sum_{a_i \in B_m} s(a_i) &\leq 1 \end{aligned} \tag{6.21}$$

Ist diese Bedingung nicht erfüllt, wird versucht den Gegenstand a_j in den Behälter B_{m+1} zu packen. Im Gegensatz zum Next-Fit-Algorithmus wird der Behälter B_m nicht geschlossen. Dieser Vorgang wird solange wiederholt, bis alle k Gegenstände verpackt sind.
Der First-Fit-Algorithmus ist 1,7-kompetitiv. Es lässt sich zeigen, dass gilt (vgl. Garey et al., 1976, S. 262):

$$FF(L) \leq \left\lceil \frac{17}{10} \cdot OPT(L) \right\rceil \tag{6.22}$$

Die Zeitkomplexität T_{FF} des First-Fit-Algorithmus beträgt (vgl. Coffman et al., 1997, S. 50):

$$T_{FF} = O(n \log n) \tag{6.23}$$

Ein weiterer 1,7-kompetitiver Algorithmus ist der Best-Fit-Algorithmus (BF) (vgl. Johnson et al., 1974, S. 300). Beim Best-Fit-Algorithmus wird ein Gegenstand a_j in den Behälter gepackt, in dem der höchste Füllstand $v'(B)$ erreicht wird:

$$v'(B_m) = s(a_j) + v(B_m) \quad \text{mit} \quad v'(B_m) = \max_{i=1}^{n} \left(s(a_j) + v(B_i) \leq 1\right) \tag{6.24}$$

Die Zeitkomplexität T_{BF} des Best-Fit-Algorithmus beträgt ebenfalls (vgl. Coffman et al., 1997, S. 53):

$$T_{BF} = O(n \log n) \tag{6.25}$$

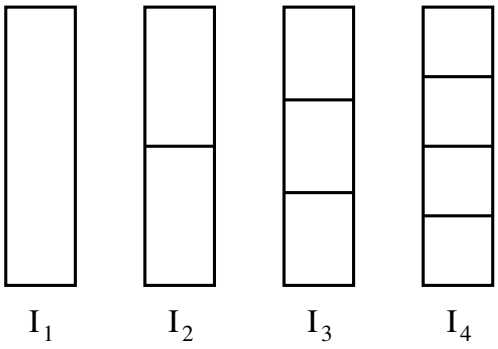

Abbildung 6.2: Partitionierung des HARMONIC-k-Algorithmus für M=4

Bessere Ergebnisse liefert der HARMONIC-k-Algorithmus (vgl. Lee und Lee, 1985, S. 562 ff.). Hierbei wird das Intervall $(0, 1]$ wie folgt in Partitionierungen I unterteilt:

$$(0,1] = \bigcup_{k=1}^{M} I_k \quad \text{mit} \qquad I_k = \left(\frac{1}{k+1}, \frac{1}{k}\right] \quad \text{für} \quad 1 \leq k < M$$
$$I_M = \left(0, \frac{1}{M}\right] \qquad M \in \mathbb{N}^+ \tag{6.26}$$

Abbildung 6.2 veranschaulicht die Partitionierung für den Fall $M = 4$. Für jede Partitionierung I_k wird ein eigener I_k-Behälter verwendet. Ein Gegenstand a wird entsprechend seiner Größe $s(a) \in I_k$ in den I_k-Behälter gepackt. Ist in dem I_k-Behälter nicht mehr genügend Platz, wird der Behälter geschlossen, und ein neuer Behälter verwendet. Die Zeitkomplexität $T_{HARMONIC}$ des HARMONIC-k-Algorithmus ist linear (vgl. Lee und Lee, 1985, S. 563):

$$T_{HARMONIC} = O(n) \tag{6.27}$$

Der kompetitive Faktor des Algorithmus hängt von der Anzahl der Partitionierungen M ab. Für sechs Partitionierungen ($M = 6$) ist der HARMONIC-k-Algorithmus 1,7-kompetitiv und damit genau so gut wie der First-Fit- und

der Best-Fit-Algorithmus. Für mehr als sechs Partitionierungen liefert der HARMONIC-k-Algorithmus bessere Ergebnisse (vgl. Lee und Lee, 1985, S. 567).
Auf Grundlage des HARMONIC-k-Algorithmus wurden verschiedene Modifikationen entwickelt, die den kompetitiven Faktor weiter verbessern. Eine Variante ist der HARMONIC++-Algorithmus mit einem kompetitiven Faktor von 1,58889 (vgl. Seiden, 2002, S. 640 ff.). Zur Lösung des Online-Bin-Packing-Problems existieren noch weitere Algorithmen (vgl. Coffman et al., 1997, S. 46 ff.).

Eindimensionales Offline-Bin-Packing

Im Unterschied zum Online-Bin-Packing sind beim Offline-Bin-Packing alle Gegenstände a der Liste L vor dem Beginn der Zuordnung zu den Behältern B bekannt. Der Online-Algorithmus First-Fit lässt sich zu einem Offline-Algorithmus First-Fit-Decreasing (FFD) modifizieren, in dem die Gegenstände a vor dem Verpacken nach ihrer Größe $s(a)$ absteigend sortiert werden. Für die neue Liste L' gilt:

$$a_i \in L' \quad \text{mit} \quad s(a_i) \geq s(a_{i+1}) \tag{6.28}$$

Anschließend wird die neue Liste L' nach dem First-Fit-Algorithmus verpackt. Es lässt sich zeigen, dass der First-Fit-Decreasing-Algorithmus $\frac{11}{9}$-kompetitiv ist (vgl. Yue, 1991, S. 321 ff.):

$$FFD(L) \leq \frac{11}{9} \cdot OPT(L) + 1 \tag{6.29}$$

Da die Liste L vor dem Verpacken noch zusätzlich sortiert werden muss, gilt für die Zeitkomplexität des First-Fit-Decreasing-Algorithmus (vgl. Coffman et al., 1984, S. 56):

$$T_{FFD} = \Omega(n \log n) \tag{6.30}$$

Der Online-Algorithmus Best-Fit lässt sich ebenfalls zu einem Offline-Algorithmus Best-Fit-Decreasing (BFD) modifizieren. Auch dieser Algorithmus ist $\frac{11}{9}$-kompetitiv und hat dieselbe Zeitkomplexität wie der First-Fit-Decreasing-Algorithmus (vgl. Coffman et al., 1997, S. 60 ff.). Enthält die Liste L keine Gegenstände a deren Größe $s(a)$ kleiner als $\frac{1}{6}$ ist, liefert der Best-Fit-Algorithmus kompetitiv mindestens so gute Ergebnisse wie der First-Fit-Algorithmus (vgl. Johnson et al., 1974, S. 310 ff.):

$$BFD(L) \leq FFD(L) \quad \text{mit} \quad \forall a_i \in L : \frac{1}{6} \leq s(a_i) \leq 1 \tag{6.31}$$

Weiterhin lässt sich zeigen, dass es keinen Online-Algorithmus geben kann, der kompetitiv besser als die Offline-Algorithmen First-Fit-Decreasing und Best-Fit-Decreasing ist. Es lässt sich zeigen, dass für jeden Online-Algorithmus ONLINE gilt (vgl. Gehweiler und auf der Heide, 2008, S. 401 ff.):

$$ONLINE(L) \geq \frac{4}{3} \cdot OPT(L) \tag{6.32}$$

Die Liste L' soll $2 \cdot x$ Gegenstände a' enthalten, für deren Größe $s(a')$ gilt:

$$s(a') = \frac{1}{2} - \varepsilon \quad \text{mit} \quad 0 < \varepsilon < \frac{1}{6} \tag{6.33}$$

Ein optimaler Algorithmus OPT benötigt x Behälter B zum Verpacken der Gegenstände a' der Liste L', da immer zwei Gegenstände in einen Behälter passen:

$$OPT(L') = x \tag{6.34}$$

Ein Online-Algorithmus ONLINE verpackt die Gegenstände a' so in die Behälter B, dass sich entweder ein oder zwei Gegenstände in einem Behälter befinden. Die verwendete Anzahl der Behälter $ONLINE(L')$ ist die Summe der Behälter n_1 in denen sich ein Gegenstand a' befindet und der Behälter n_2 in denen sich zwei Gegenstände a' befinden:

$$ONLINE(L') = n_1 + n_2 \tag{6.35}$$

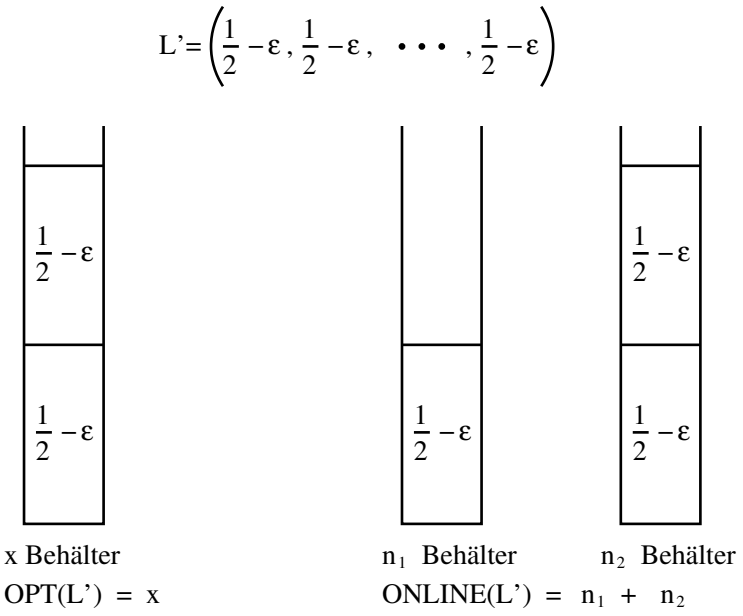

Abbildung 6.3: Zuordnungen für L'

Für die Gesamtzahl aller Gegenstände gilt somit:

$$\begin{aligned} 2 \cdot x &= n_1 + 2 \cdot n_2 \\ n_1 &= 2 \cdot x - 2 \cdot n_2 \end{aligned} \tag{6.36}$$

Durch Einsetzen von Formel 6.36 in Formel 6.35 erhält man:

$$\begin{aligned} ONLINE(L') &= 2 \cdot x - 2 \cdot n_2 + n_2 \\ &= 2 \cdot x - n_2 \end{aligned} \tag{6.37}$$

Abbildung 6.3 veranschaulicht die Zuordnungen. Eine Liste L'' soll neben den $2 \cdot x$ Gegenständen a', welche die Bedingung 6.33 erfüllen, $2 \cdot x$ zusätzliche Gegenstände a'' enthalten, für deren Größe $s(a'')$ gilt:

$$s(a'') = \frac{1}{2} + \varepsilon \quad \text{mit} \quad 0 < \varepsilon < \frac{1}{6} \tag{6.38}$$

Ein optimaler Algorithmus OPT benötigt $2 \cdot x$ Behälter B zum Verpacken der Gegenstände a und a'. Ein kleiner Gegenstand a' (mit $s(a') = \frac{1}{2} - \epsilon$) und

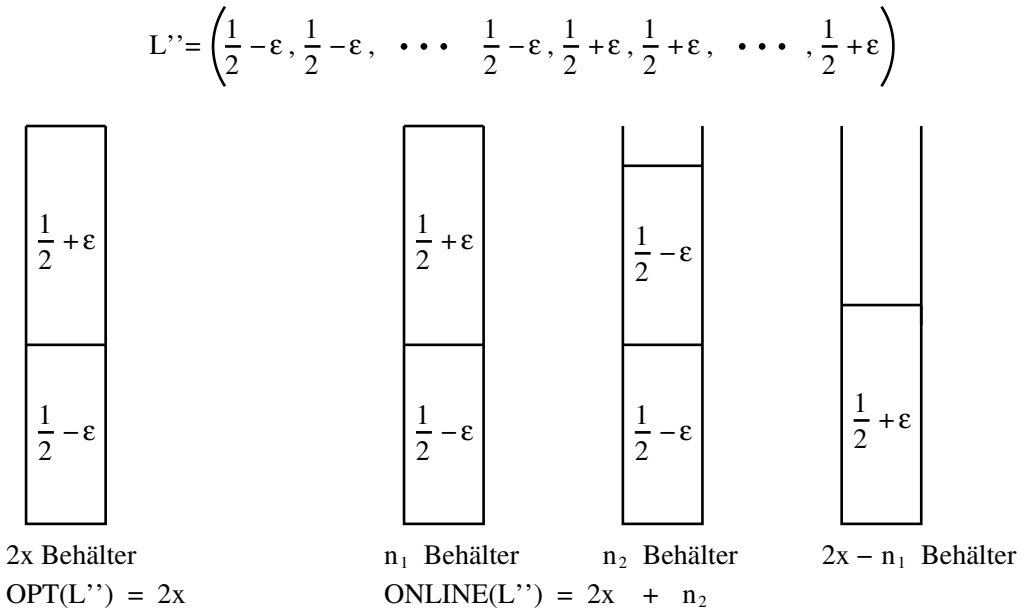

Abbildung 6.4: Zuordnungen für L"

ein großer Gegenstand a'' (mit $s(a'') = \frac{1}{2} + \epsilon$) werden immer zusammen in einen Behälter gepackt:

$$OPT(L'') = 2 \cdot x \quad \text{mit} \quad s(a') + s(a'') = \frac{1}{2} - \epsilon + \frac{1}{2} + \epsilon = 1 \tag{6.39}$$

Ein Online-Algorithmus ONLINE wird die ersten $2 \cdot x$ Gegenstände a der Liste L'' mit der Größe $s(a') = \frac{1}{2} - \epsilon$ genauso verpacken wie die Liste L'. Mit einem Teil der weiteren Gegenstände a'' der Liste L'' mit der Größe $s(a'') = \frac{1}{2} + \epsilon$ lassen sich die Behälter n_1 auffüllen, in denen sich nur ein Gegenstand befindet. Für den restlichen Teil der Gegenstände a'' muss jeweils ein neuer Behälter verwendet werden. Die Anzahl der vom Online-Algorithmus benötigten Behälter $ONLINE(L'')$ beträgt somit:

$$\begin{aligned} ONLINE(L'') &= n + 2 \cdot x - n_1 \\ &= n_1 + n_2 + 2 \cdot x - n_1 \\ &= 2 \cdot x + n_2 \end{aligned} \tag{6.40}$$

Abbildung 6.4 veranschaulicht die Zuordnungen. Um die Behauptung 6.32 zu

bestätigen, wird ein Negativbeweis geführt. Es ist zu widerlegen, dass gilt:

$$ONLINE(L) < \frac{4}{3} \cdot OPT(L) \tag{6.41}$$

Setzt man die Ergebnisse für die Liste L' in die Behauptung 6.41 ein, erhält man:

$$\begin{aligned} 2 \cdot x - n_2 &< \frac{4}{3} \cdot x \\ n_2 &> \frac{2}{3} \cdot x \end{aligned} \tag{6.42}$$

Setzt man die Ergebnisse für die Liste L'' in die Behauptung 6.41 ein, erhält man:

$$\begin{aligned} 2 \cdot x + n_2 &< \frac{4}{3} \cdot 2 \cdot x \\ n_2 &< \frac{2}{3} \cdot x \end{aligned} \tag{6.43}$$

Aussage 6.42 und Aussage 6.43 stehen in Widerspruch zueinander. Die Behauptung 6.41 ist somit falsch. Es kann keinen Online-Algorithmus geben, der besser als $\frac{4}{3}$-kompetitiv ist. Die Offline-Algorithmen First-Fit-Decreasing und Best-Fit-Decreasing sind kompetitiv besser als jeder Online-Algorithmus.
Es lässt sich nachweisen, dass die untere kompetitive Schranke von Online-Algorithmen noch höher liegt, ohne die getroffenen Aussagen zu beeinflussen. So kann man in einem aufwändigeren Beweis zeigen, dass kein Online-Algorithmus besser als $\frac{3}{2}$-kompetitiv ist (vgl. Yao, 1980, S. 212 ff.).

Mehrdimensionales Bin-Packing

Bei eindimensionalen Bin-Packing-Problemen wurde davon ausgegangen, dass die Behälter und die Gegenstände nur eine Größenrestriktion haben. Bestehen für die Behälter mehrere Größenrestriktionen, liegt ein mehrdimensionales Bin-Packing-Problem vor. Können die Größenrestriktionen unabhängig voneinander betrachtet werden, handelt es sich um ein Vector-Packing-Problem (vgl. Caprara et al., 2002, S. 58). Jeder Behälter B hat beim

Vector-Packing-Problem d voneinander unabhängige Größenrestriktionen s. Die Größenrestriktionen werden einheitlich skaliert, so dass gilt:

$$s_i(B) = 1 \quad \text{mit} \quad 1 \leq i \leq d \tag{6.44}$$

Die Größe eines zu verpackenden Gegenstands a wird durch d Größenmerkmale s beschrieben. Für die Größe der zu verpackenden Gegenstände a gilt:

$$s_i(a) \leq 1 \quad \text{mit} \quad 1 \leq i \leq d \tag{6.45}$$

Der Füllstand v des Behälters darf für jede Dimension die Größe des Behälters in dieser Dimension nicht überschreiten:

$$\begin{aligned} \forall i \in d: \qquad v_i(B) &\leq 1 \\ \sum_{a_j \in B} s_i(a_j) &\leq 1 \end{aligned} \tag{6.46}$$

Algorithmen zur Lösung von Vector-Packing-Problemen lassen sich wieder in Online- und Offline-Algorithmen unterteilen. Ein einfacher Online-Algorithmus zur Lösung des Vector-Packing-Problems ist der bereits beschriebene First-Fit-Algorithmus. Im Gegensatz zum eindimensionalen Fall werden jedoch d Größenrestriktionen beim Verpacken berücksichtigt. Ein Gegenstand a_j wird in den Behälter m gepackt wenn gilt:

$$\begin{aligned} \forall i \in d: \qquad s_i(a_j) + v_i(B_m) &\leq 1 \\ s_i(a_j) + \sum_{a_k \in B_m} s_i(a_k) &\leq 1 \end{aligned} \tag{6.47}$$

Die Kompetitivität des First-Fit-Algorithmus ($\widehat{FF}$) ist beim Vector-Packing-Problem von der Anzahl der Dimensionen d abhängig (vgl. Garey et al., 1976, S. 263):

$$\widehat{FF(L)} \leq \left(d + \frac{7}{10}\right) \cdot OPT(L) + \frac{5}{2} \tag{6.48}$$

Ist die Liste L der Gegenstände vor dem Verpacken vollständig bekannt, kann der Offline-Algorithmus First-Fit-Decreasing eingesetzt werden. Die Gegenstände werden absteigend nach der Größe sortiert. Für die neue Liste L' gilt:

$$a_j \in L' \quad \text{mit} \quad \max_{i \in d} s_i(a_j) \geq \max_{i \in d} s_i(a_{j+1}) \tag{6.49}$$

Anschließend wird die Liste L' nach dem First-Fit-Algorithmus verpackt. Für die Kompetitivität des First-Fit-Decreasing-Algorithmus ($\widehat{FFD}$) gilt beim Vector-Packing-Problem in Abhängigkeit von der Anzahl an Dimensionen d (vgl. Garey et al., 1976, S. 277):

$$\widehat{FFD(L)} \leq \left(d + \frac{1}{3}\right) \cdot OPT(L) \tag{6.50}$$

Das Sortieren der Liste liefert somit auch im mehrdimensionalen Fall kompetitiv bessere Ergebnisse. Im folgenden Abschnitt werden auf Grundlage der beschriebenen Algorithmen Methoden zur Zuordnung virtualisierter Betriebsmittel entwickelt.

6.1.2 Zuordnung virtualisierter Betriebsmittel als Bin-Packing-Problem

Zur Erbringung einer IT-Dienstleistung werden Betriebsmittel benötigt. Betriebsmittel stellen ihr Nutzungspotential der IT-Dienstleistung über einen längeren Zeitraum zur Verfügung. Jedes Betriebsmittel kann durch sein Kapazitätsangebot spezifiziert werden. Das Kapazitätsangebot beschreibt das technische Leistungsvermögen eines Betriebsmittels. Virtualisierungstechniken ermöglichen eine effiziente Ausnutzung des technischen Leistungsvermögens dieser Betriebsmittel. Das technische Leistungsvermögen physischer Betriebsmittel kann durch Partitionierung auf mehrere virtuelle Betriebsmittel aufgeteilt werden.
Die Zuordnung der virtuellen Betriebsmittel zu physischen Betriebsmitteln lässt sich als Bin-Packing-Problem formulieren. Es sollen mehrere virtu-

elle Betriebsmittel mit einer spezifischen Kapazität auf möglichst wenige physische Betriebsmittel verteilt werden. Die Summe der Kapazitäten der virtuellen Betriebsmittel darf dabei die Nettokapazität des physischen Betriebsmittels nicht überschreiten. Unter der Nettokapazität wird hierbei die tatsächlich für die virtuellen Betriebsmittel zur Verfügung stehende Kapazität des physischen Betriebsmittels verstanden. Die Bruttokapazität berechnet sich aus der Nettokapazität und der für den Virtualisierungsoverhead benötigten Kapazität:

$$c^{Brutto} = c^{Netto} + c^{Overhead} \tag{6.51}$$

Für die Kapazität $c^{Virtuell}$ der virtuellen Betriebsmittel, die einem physischen Betriebsmittel P zugeordnet sind, gilt:

$$\sum_{i \in P} c_i^{Virtuell} \leq c_P^{Netto} \tag{6.52}$$

Diese Bedingung entspricht der Bedingung 6.6 des Bin-Packing Problems. Die physischen Betriebsmittel werden in Losen beschafft. Ein Los physischer Betriebsmittel steht für die Zuordnung der virtuellen Betriebsmittel zur Verfügung. Geht man davon aus, dass in einem Los Betriebsmittel mit gleichem technischen Leistungsvermögen beschafft werden, verfügen alle physischen Betriebsmittel über die gleiche Kapazität. Die Kapazitäten lassen sich auf einen Einheitswert von Eins skalieren und entsprechen damit den Anforderungen an die Größe der Behälter des Bin-Packing-Problems.
IT-Dienstleistungen werden kundenauftragsorientiert produziert. Die hierfür benötigten virtuellen Betriebsmittel können für jeden Fertigungsauftrag einzeln angelegt werden. Die vollständige Liste der einzurichtenden virtuellen Betriebsmittel ist in diesem Fall bei der Zuordnung der virtuellen Betriebsmittel nicht bekannt. Die auftragsorientierte Einrichtung der virtuellen Betriebsmittel entspricht somit dem Online-Bin-Packing-Problem. Die virtuellen Betriebsmittel können auch auf Vorrat bereitgestellt werden. Bei der Vorratsbereitstellung virtueller Betriebsmittel ist die Liste der einzurichtenden virtuellen Betriebsmittel bekannt. Diese Variante stellt ein

Offline-Bin-Packing-Problem dar. In Abschnitt 6.1.1 wurde für den Fall des eindimensionalen Bin-Packings bewiesen, dass es Algorithmen zur Lösung des Offline-Bin-Packing-Problems gibt, die kompetitiv besser sind als jeder Online-Algorithmus. Aus diesem Beweis lässt sich schlussfolgern, dass es aus Sicht der Kapazitätswirtschaft sinnvoll ist, die virtuellen Betriebsmittel auf Vorrat bereitzustellen, wenn die virtuellen Betriebsmittel nur eine Kapazitätsrestriktion haben. Für eine Verallgemeinerung dieser Behauptung auf den Fall, dass ein virtuelles Betriebsmittel durch mehr als eine Kapazitätsrestriktion beschrieben wird, besteht weiterer Forschungsbedarf. Am Beispiel des First-Fit-Algorithmus lässt sich jedoch auch im mehrdimensionalen Fall zeigen, dass kompetitiv bessere Ergebnisse erzielt werden, wenn die Liste der zu verpackenden Gegenstände vorher bekannt ist. Es ist also anzunehmen, dass die Vorratsbereitstellung virtueller Betriebsmittel immer sinnvoll ist.

Die Ermittlung des Bedarfs an Komponenten und Kapazitäten wird als Betriebsmittelbedarfsplanung bezeichnet (engl. resource requirements planning [RRP]) (vgl. Schönsleben, 2007, S. 228). Als virtuelle Betriebsmittelbedarfsplanung (engl. virtual resource requirements planning [VRRP]) wird im Folgenden die Ermittlung des Bedarfs an virtuellen Betriebsmitteln, deren Einrichtung und die Zuordnung zu physischen Betriebsmitteln bezeichnet.

Ein Rechenzentrum bietet eine bestimmte Menge verschiedener IT-Dienstleistungen an. Zur Erbringung einer IT-Dienstleistung werden virtuelle Betriebsmittel verwendet. Diese virtuellen Betriebsmittel sollen entsprechend der (s,Q)-Politik auf Vorrat in Losen Q bereit gestellt werden. Unterschreitet der Bestand eines virtuellen Betriebsmittels V einen definierten Meldebestand s, soll eine virtuelle Betriebsmittelbedarfsplanung (VRRP-Lauf) durchgeführt werden (siehe Abbildung 6.5). Der Bestand ist die Menge an eingerichteten virtuellen Betriebsmitteln mit gleichem technischen Leistungsvermögen, die noch nicht in Fertigungsaufträgen verwendet werden. Vor der Freigabe eines Fertigungsauftrags werden die Planbetriebsmittel durch tatsächliche virtuelle Betriebsmittel substituiert (siehe Abschnitt 3.1.2). Nach jeder Substitution wird überprüft, ob der Bestand eines virtuellen Betriebsmittels den Meldebestand s erreicht oder unterschritten hat. Ist dies

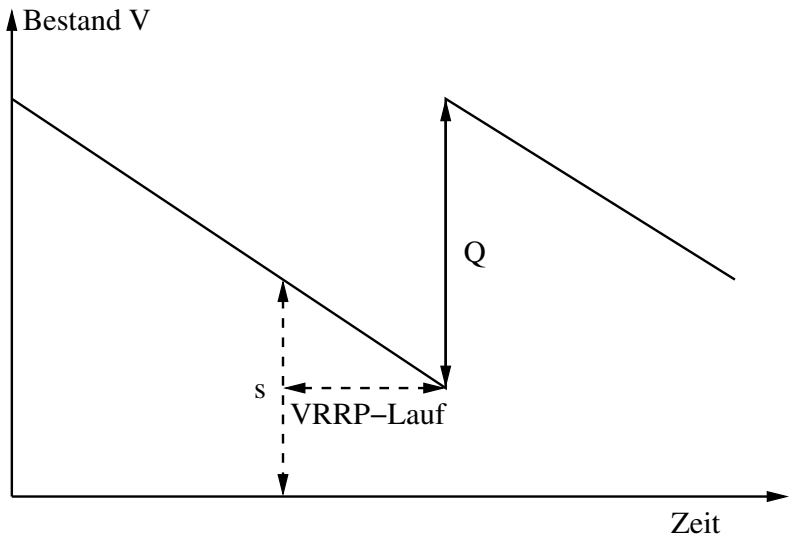

Abbildung 6.5: VRRP-Lauf

der Fall, wird ein VRRP-Lauf angestoßen. Ein VRRP-Lauf besteht aus den zwei Arbeitschritten:

- **Schritt 1:** Reservierung der physischen Kapazitäten und
- **Schritt 2:** Einrichtung der virtuellen Kapazitäten.

Im ersten Schritt werden die physischen Kapazitäten reserviert, die zur Einrichtung der virtuellen Kapazitäten benötigt werden. Voraussetzung hierfür ist das Vorhandensein einer ausreichenden Menge physischer Betriebsmittel, die im Vorfeld beschafft wurden. Die Nettokapazität eines physischen Betriebsmittels P ist wie folgt aufgeteilt:

- **Fertigungsaufträge (F):** Virtuelle Betriebsmittel, die bereits in Fertigungsaufträgen verwendet werden,
- **Installationen (I):** Virtuelle Betriebsmittel, die bereits installiert sind aber noch nicht in Fertigungsaufträgen verwendet werden und
- **Reservierungen (R):** Platzhalter für einzurichtende virtuelle Betriebsmittel.

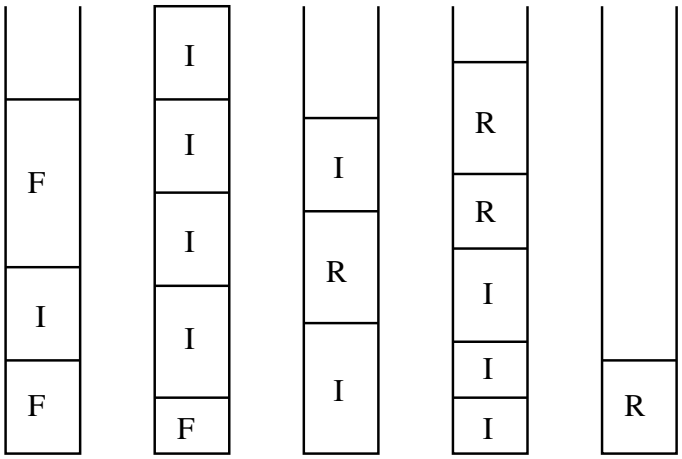

Abbildung 6.6: Mögliche Aufteilung physischer Betriebsmittel

Die nicht zugeordnete Nettokapazität steht als freie Kapazität zur Verfügung. Abbildung 6.6 zeigt eine mögliche Aufteilung physischer Betriebsmittel. Zur Reservierung der physischen Betriebsmittel soll ein Offline-Bin-Packing-Algorithmus angewendet werden. In verschiedenen IT-Dienstleistungen werden gleiche virtuelle Betriebsmittel mit unterschiedlichem technischen Leistungsvermögen eingesetzt. Für jedes technische Leistungsvermögen i eines Betriebsmittels wird ein eigener Bestand V^i geführt und eine Losgröße Q^i definiert. Bei der Durchführung des Bin-Packing-Algorithmus sollen für ein Betriebsmittel mit einem technischen Leistungsvermögen i, dessen Meldebestand erreicht oder unterschritten wurde, physische Kapazitäten reserviert werden. Um bei der Durchführung des Bin-Packing-Algorithmus bessere Resultate zu erzielen, werden zusätzlich Reservierungen für die technischen Leistungsvermögen j mit $j \neq i$ des virtuellen Betriebsmittels vorgenommen. Die Menge der neu einzurichtenden Installationen I^i_{neu} des virtuellen Betriebsmittels V^i entspricht der Losgröße Q^i:

$$I^i_{neu} = Q^i \tag{6.53}$$

Für die einzurichtenden Installationen I müssen Reservierungen R vorgenommen werden. Die Menge der Reservierungen R^i_{neu} für das Betriebsmittel

V^i berechnet sich aus der Menge der einzurichtenden Installationen abzüglich der bereits in früheren VRRP-Läufen vorgenommen Reservierungen R^i_{alt}:

$$\begin{aligned} R^i_{neu} &= I^i_{neu} - R^i_{alt} \\ &= Q^i - R^i_{alt} \end{aligned} \tag{6.54}$$

Die Menge der Reservierungen R^j_{neu} für die Betriebsmittel V^j mit $j \neq i$ ergibt sich aus der Losgröße Q^j abzüglich der bereits in früheren VRRP-Läufen vorgenommen Reservierungen R^j_{alt} sowie der Installationen I^j_{alt}, die noch nicht in Fertigungsaufträgen verwendet werden:

$$R^j_{neu} = Q^j - I^j_{alt} - R^j_{alt} \tag{6.55}$$

Nachdem die Mengen der vorzunehmenden Reservierungen für die verschiedenen technischen Leistungsvermögen eines virtuellen Betriebsmittels bekannt sind, kann ein geeigneter Offline-Bin-Packing-Algorithmus ausgeführt werden. Um eine bessere Auslastung zu erzielen, wird der Bin-Packing-Algorithmus neben den neu beschafften physischen Betriebsmitteln, auch auf die bereits im Einsatz befindlichen Betriebsmittel angewendet. Als Ergebnis liegen die Reservierungen der physischen Betriebsmittel für die virtuellen Betriebsmittel vor.

Im nächsten Schritt werden die virtuellen Betriebsmittel eingerichtet. Ausfälle physischer Betriebsmittel werden als Betriebsstörungen bezeichnet (vgl. Kollerer, 1978, S. 20). Um flexibel auf Störungen reagieren zu können, werden ausschließlich die virtuellen Betriebsmittel V^i eingerichtet. Die Kapazitäten der virtuellen Betriebsmittel V^j mit $j \neq i$ werden zwar reserviert jedoch nicht eingerichtet. Die Reservierungen R^j können bei auftretenden Betriebsstörungen wieder rückgängig gemacht werden, um die Kapazitäten für die Beseitigung von Betriebsstörungen zu verwenden. In diesem Fall werden die Reservierungen R^j beim nächsten VRRP-Lauf neu vorgenommen.

Bei der Zuordnung der virtuellen Betriebsmittel wird von einer vorgegebenen Größe des Kapazitätsbedarfs ausgegangen. In Abschnitt 4.2.3 wurde jedoch nachgewiesen, dass der Kapazitätsbedarf von der Anzahl der virtuellen Be-

triebsmittel abhängt, die sich auf einem physischen Betriebsmittel befinden. Wird die Größe des Kapazitätsbedarfs eines virtuellen Betriebsmittels in Abhängigkeit der Anzahl der virtuellen Betriebsmittel auf einem physischen Betriebsmittel bestimmt, lässt sich die Ausnutzung des Kapazitätsangebots weiter verbessern. Hierzu muss jedoch die Voraussetzung einer festen Größe der zu verpackenden Gegenstände des Bin-Packing-Problems aufgegeben werden. Für eine Untersuchung, welche Auswirkung dies auf die Kompetitivität der Bin-Packing-Algorithmen hat besteht weiterer Forschungsbedarf. Im folgenden Abschnitt wird die Durchführung eines VRRP-Laufs an einem einfachen Beispiel erläutert.

6.1.3 Beispiel einer Zuordnung virtualisierter Betriebsmittel

Ein Rechenzentrum bietet drei IT-Dienstleistungen A, B und C an. Für jede IT-Dienstleistung wird ein virtueller Server mit einem spezifischen technischen Leistungsvermögen benötigt:

- IT-Dienstleistung A: virtueller Server V^A mit $c^A = 1000$ MIPS,
- IT-Dienstleistung B: virtueller Server V^B mit $c^B = 2000$ MIPS und
- IT-Dienstleistung C: virtueller Server V^C mit $c^C = 3000$ MIPS.

Die virtuellen Server werden in Losen Q zu 10 Stück eingerichtet. Der Meldebestand s beträgt zwei Stück:

$$\begin{aligned} Q^A = Q^B = Q^C = 10 \\ s^A = s^B = s^C = 2 \end{aligned} \tag{6.56}$$

Physische Server werden in Losen zu 10 Stück mit einem technischen Leistungsvermögen von 5000 MIPS beschafft. Ein neues Los physischer Server

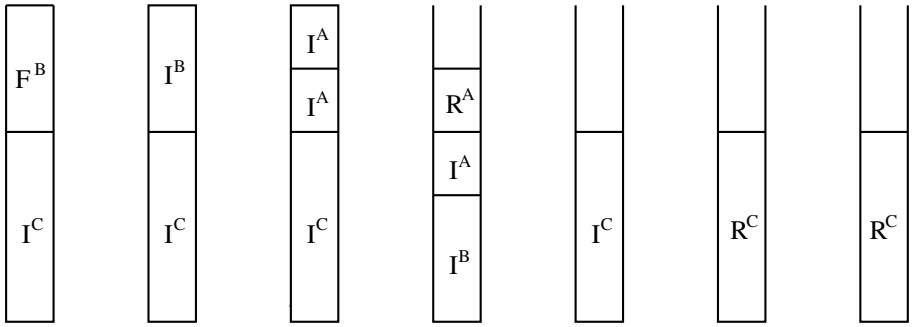

Abbildung 6.7: Zuordnung der virtuellen Server nach der Substitution

wurde beschafft und steht für die Produktion zur Verfügung. Nach dem letzten VRRP-Lauf ist der Bestand an Installationen I, die nicht in Fertigungsaufträgen verwendet werden, und Reservierungen R:

$$\begin{aligned} I^A &= 3 & R^A &= 1 \\ I^B &= 3 & R^B &= 0 \\ I^C &= 4 & R^C &= 2 \end{aligned} \tag{6.57}$$

Es wird ein neuer Fertigungsauftrag zur Produktion der IT-Dienstleistung B angelegt. Anschließend wird das Planbetriebsmittel, das den virtuellen Server beschreibt, durch einen virtuellen Server V^B substituiert. Eine Installation I^B wird in einen Fertigungsauftrag F^B überführt. Abbildung 6.7 zeigt die Zuordnung der virtuellen Server nach der Substitution.
Nachdem das Planbetriebsmittel im Fertigungsauftrag substituiert wurde, ist zu überprüfen, ob der Bestand an Installationen I für einen virtuellen Server den Meldebestand s erreicht oder unterschritten hat. Die Überprüfung liefert folgendes Ergebnis:

$$\begin{aligned} I^A = 3 &> s^A \\ I^B = 2 &\leq s^B \\ I^C = 4 &> s^C \end{aligned} \tag{6.58}$$

Für die virtuellen Server V^B wird der Meldebestand s^B erreicht. Es ist ein VRRP-Lauf anzustoßen. Als Erstes werden die physischen Kapazitäten reserviert. Um das Bin-Packing für die Reservierungen vornehmen zu können,

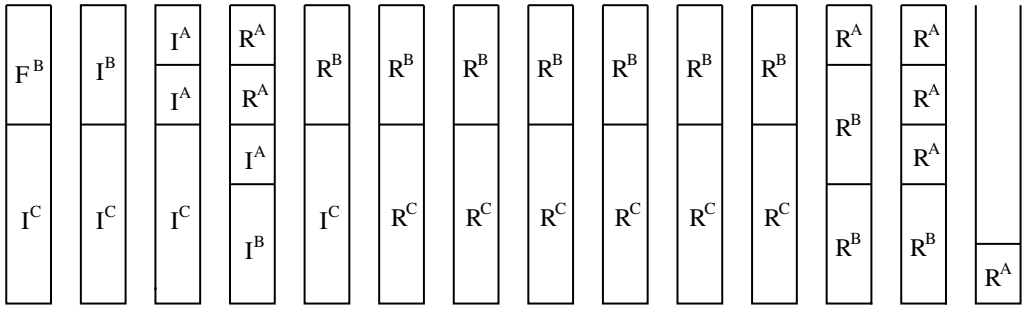

Abbildung 6.8: Zuordnung der virtuellen Server nach der Ausführung von Best-Fit-Decreasing

muss die Menge an neuen Reservierungen ermittelt werden. Für die virtuellen Server V^B beträgt die Menge an vorzunehmenden Reservierungen nach Formel 6.54:

$$\begin{aligned} R^B_{neu} &= Q^B - R^B_{alt} \\ &= 10 - 0 \\ &= 10 \end{aligned} \tag{6.59}$$

Nach Formel 6.55 beträgt die Menge an vorzunehmenden Reservierungen für die virtuellen Server V^A:

$$\begin{aligned} R^A_{neu} &= Q^A - I^A_{alt} - R^A_{alt} \\ &= 10 - 3 - 1 \\ &= 6 \end{aligned} \tag{6.60}$$

Für die virtuellen Server V^C beträgt die Menge an vorzunehmenden Reservierungen nach Formel 6.55:

$$\begin{aligned} R^C_{neu} &= Q^C - I^C_{alt} - R^C_{alt} \\ &= 10 - 4 - 2 \\ &= 4 \end{aligned} \tag{6.61}$$

Die Reservierungen sollen durch einen Offline-Algorithmus vorgenommen werden. Da das technische Leistungsvermögen von keinem virtuellen Server $\frac{1}{6}$ des technischen Leistungsvermögens eines physischen Servers unterschreitet, ist nach Formel 6.31 der Einsatz des Best-Fit-Decreasing-Algorithmus

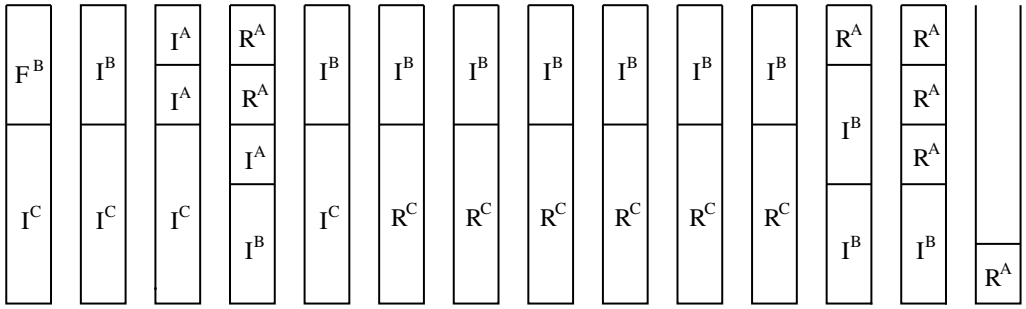

Abbildung 6.9: Zuordnung der virtuellen Server nach der Einrichtung der Betriebsmittel V^B

sinnvoll. Hierzu sind die Reservierungen absteigend nach ihrem technischen Leistungsvermögen zu sortieren:

$$c^C > c^B > c^A \tag{6.62}$$

Die Reservierungen der virtuellen Server V^C werden als erste nach dem Best-Fit-Algorithmus zugeordnet, da sie das größte technische Leistungsvermögen haben. Im Anschluss werden die virtuellen Server V^B zugeordnet. Als Letztes werden die Zuordnungen der Reservierungen der virtuellen Server V^A ausgeführt. Die Liste L der Reservierungen R, die den physischen Servern zuzuordnen sind, enthält:

$$\begin{aligned} L = (\quad & R^C, R^C, R^C, R^C, \\ & R^B, R^B, R^B, R^B, R^B, R^B, R^B, R^B, R^B, R^B, \\ & R^A, R^A, R^A, R^A, R^A, R^A \quad) \end{aligned} \tag{6.63}$$

Als Nächstes wird die Liste L nach dem Best-Fit-Algorithmus den physischen Servern zugeordnet. Abbildung 6.8 zeigt das Ergebnis der Zuordnung.
Im nächsten Schritt werden die virtuellen Betriebsmittel V^B eingerichtet. Die Reservierungen R^B werden hierbei in Installationen I^B überführt. Abbildung 6.9 zeigt das Ergebnis dieses Vorgangs. Nach der Durchführung des VRRP-Laufs steht eine ausreichende Menge an virtuellen Betriebsmitteln zur Produktion der IT-Dienstleistung B zur Verfügung.
Werden die Planbetriebsmittel in einem Fertigungsauftrag durch virtuelle

Betriebsmittel ersetzt, stehen mehrere Ressourcen für die Substitution zur Auswahl. Im nächsten Abschnitt wird beschrieben, wie die Auswahl eines geeigneten virtuellen Betriebsmittels vorgenommen wird.

6.2 Betriebsmittelauswahl

Als Ergebnis der Zuordnung der virtuellen Betriebsmittel stehen auf verschiedenen physischen Betriebsmitteln mehrere virtuelle Betriebsmittel mit gleichem technischen Leistungsvermögen zur Verfügung. Diese virtuellen Betriebsmittel sollen eingesetzt werden, um die Planbetriebsmittel in den Fertigungsaufträgen zu substituieren. Die Kapazitäten der physischen Betriebsmittel sollen hierbei möglichst effizient eingesetzt werden. In Abschnitt 4.2.3 wurde nachgewiesen, dass bei sinkender Korrelation der Kapazitätsbedarfe der virtuellen Server auch das in Anspruch genommene Kapazitätsangebot des physischen Servers sinkt. Diese Eigenschaft soll bei der Betriebsmittelauswahl berücksichtigt werden.

6.2.1 Korrelationskoeffizient von Bravais-Pearson

Gegeben sind zwei Merkmale X und Y. Für definierte Zeitpunkte werden Beobachtungswerte der Merkmale erhoben. Existieren für wachsende Beobachtungswerte des Merkmals X auch tendenzmäßig größere Beobachtungswerte des Merkmals Y zum selben Zeitpunkt, wird ein Zusammenhang zwischen diesen Merkmalen vermutet. Ein Maß für die Stärke dieses Zusammenhangs ist der Bravais-Pearson-Korrelationskoeffizient (vgl. Fahrmeir, 2004, S. 134). Dieser wird wie folgt ermittelt (vgl. Cramer und Kamps, 2008, S. 109):

$$r = r(XY) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad \text{mit} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6.64}$$

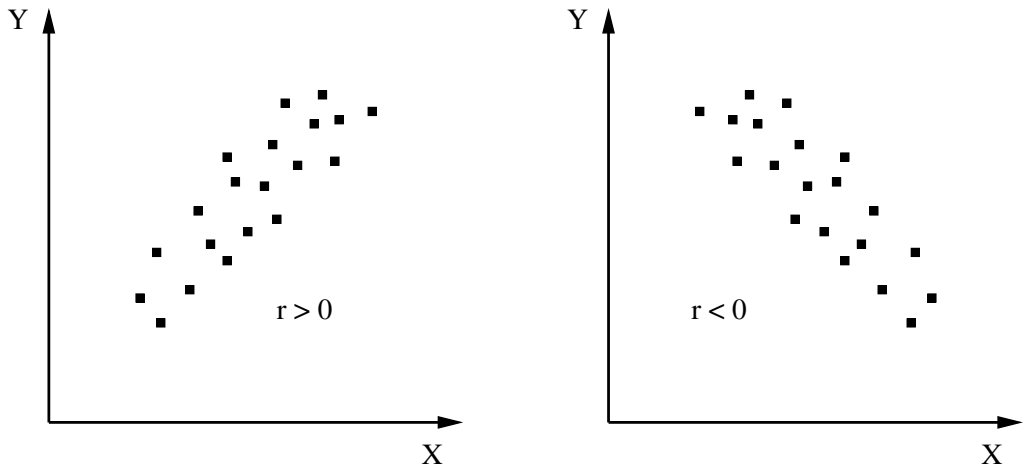

Abbildung 6.10: Bravais-Pearson-Korrelationskoeffizient für Merkmale X und Y (nach Fahrmeir, 2004, S. 138)

Durch den Term im Nenner der Formel wird eine Normierung des Korrelationskoeffizienten r vorgenommen. Der Bravais-Pearson-Korrelationskoeffizient liegt im Intervall (vgl. Fahrmeir, 2004, S. 139):

$$-1 \leq r \leq 1 \tag{6.65}$$

Für Ausprägungen des Bravais-Pearson-Korrelationskoeffizient gilt:

- $r > 0$
  Zwischen den Merkmalen X und Y liegt eine positive Korrelation vor. Es besteht ein gleichsinniger linearer Zusammenhang.

- $r < 0$
  Zwischen den Merkmalen X und Y liegt eine negative Korrelation vor. Es besteht ein gegensinniger linearer Zusammenhang.

- $r = 0$
  Die Merkmale X und Y sind unkorreliert. Es besteht ein kein linearer Zusammenhang.

Abbildung 6.10 veranschaulicht das Vorliegen positiver und negativer Korrelation zwischen zwei Merkmalen X und Y.Anhand des Korrelationskoef-

fizienten von Bravais-Pearson kann die Betriebsmittelauswahl vorgenommen werden.

6.2.2 Betriebsmittelauswahl anhand des Korrelationskoeffizienten von Bravais-Pearson

Im Rahmen des VRRP-Laufs wurden mehrere virtuelle Betriebsmittel mit identischem technischen Leistungsvermögen auf verschiedenen physischen Betriebsmitteln angelegt. Auf den physischen Betriebsmitteln befinden sich weitere virtuelle Betriebsmittel, die bereits in Fertigungsaufträgen verwendet werden. Diese Betriebsmittel nehmen das Kapazitätsangebot der physischen Betriebsmittel als Kapazitätsbedarf in Anspruch. Der Kapazitätsbedarf der virtuellen Betriebsmittel ist nicht konstant sondern unterliegt Schwankungen im Zeitverlauf. Ein virtuelles Betriebsmittel hat einen Grundbedarf an technischem Leistungsvermögen, der im Tagesverlauf nie unterschritten wird. Als Mittelbedarf wird der Bedarf an technischem Leistungsvermögen bezeichnet, der im Tagesverlauf über den Grundbedarf hinaus für einen längeren Zeitraum in Anspruch genommen wird. Kurzzeitig auftretende Bedarfe an technischem Leistungsvermögen, die über den Mittelbedarf hinaus gehen, werden als Spitzenbedarf bezeichnet. Der Kapazitätsbedarf eines virtuellen Betriebsmittels lässt sich in einem Lastprofil abbilden. Das technische Leistungsvermögen eines physischen Betriebsmittels wird kontinuierlich in Anspruch genommen. Bei der Erstellung der Lastprofile wird diese kontinuierliche Inanspruchnahme in Stundenintervallen diskretisiert. Ein Lastprofil besteht aus 24 Beobachtungswerten. Abbildung 6.11 zeigt ein Lastprofil für den Kapazitätsbedarf in MIPS eines virtuellen Servers. Die Daten der Lastprofile werden durch Monitoringwerkzeuge erhoben. Für das virtuelle Betriebsmittel, das in einem neuen Fertigungsauftrag verwendet werden soll, liegen noch keine Beobachtungswerte vor. Ausgehend von der Annahme, dass für identische IT-Dienstleistungen auch ähnliche Lastprofile der virtuellen Betriebsmittel vorliegen, wird aus den Lastprofilen der virtuellen Betriebs-

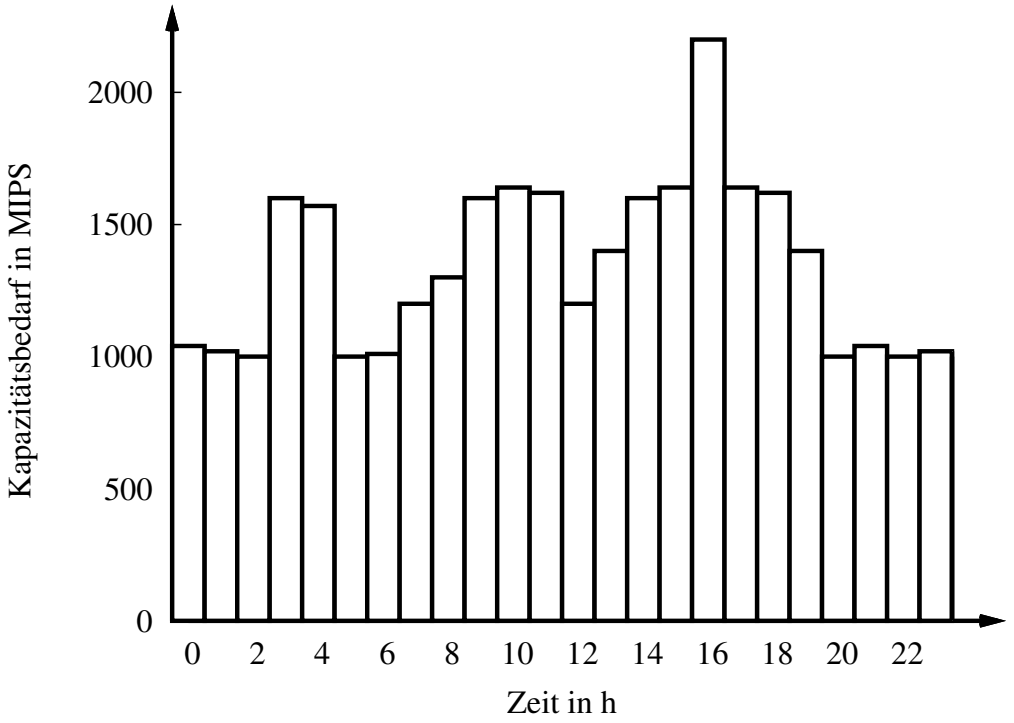

Abbildung 6.11: Lastprofil eines virtuellen Servers

mittel, die bereits in Fertigungsaufträgen für die IT-Dienstleistung verwendet werden, ein erwartetes Lastprofil erstellt. Für jedes Zeitintervall wird ein Durchschnittswert aus den vorliegenden Lastprofilen ermittelt. Hierbei ist die Zeitzone, in der sich der Kunde befindet, der die IT-Dienstleistung in Anspruch nimmt, zu berücksichtigen. Die Lastprofile werden entsprechend der Zeitverschiebung auf der Zeitachse verschoben.
Auf Basis der Lastprofile der virtuellen Betriebsmittel wird ein Lastprofil des physischen Betriebsmittels erstellt. Hierbei werden alle Lastprofile der virtuellen Server, die bereits in Fertigungsaufträgen verwendet werden, aufsummiert. Im nächsten Schritt werden die Korrelationskoeffizienten zwischen dem erwarteten Lastprofil des neuen virtuellen Betriebsmittels und den physischen Betriebsmitteln bestimmt. Der Fertigungsauftrag wird auf dem physischen Betriebsmittel mit dem niedrigsten Korrelationskoeffizienten ausgeführt.

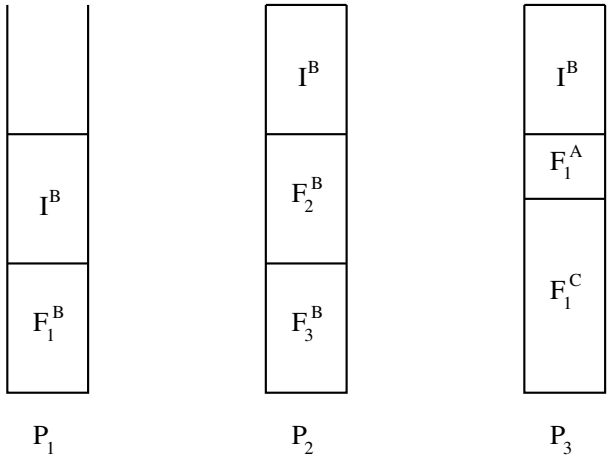

Abbildung 6.12: Zuordnung der Fertigungsaufträge

Durch die Berücksichtigung des Korrelationskoeffizienten werden Reservekapazitäten auf den physischen Servern geschaffen. Diese Reservekapazitäten stehen für unerwartet auftretende Spitzenbedarfe des technischen Leistungsvermögens zur Verfügung. Je geringer der Bestand an eingerichteten virtuellen Betriebsmitteln ist, um so geringer sind auch die Auswahlmöglichkeiten für die Zuordnung der Fertigungsaufträge. Würden in einem VRRP-Lauf alle virtuellen Betriebsmittel nicht nur reserviert sondern auch installiert, ständen stets mehr Auswahlmöglichkeiten zur Verfügung. Es könnte jedoch nicht so flexibel auf Betriebsstörungen reagiert werden. An dieser Stelle besteht ein Zielkonflikt zwischen der Flexibilität bei der Reaktion auf Betriebsstörungen und einer effizienten Kapazitätswirtschaft. Im nächsten Abschnitt wird die Betriebsmittelauswahl an einem einfachen Beispiel erläutert.

6.2.3 Beispiel einer Betriebsmittelauswahl

Ein Rechenzentrum in Berlin bietet drei IT-Dienstleistungen A, B und C an. Für jede IT-Dienstleistung wird ein virtueller Server mit einem spezifischen

technischen Leistungsvermögen benötigt:

- IT-Dienstleistung A: virtueller Server V^A mit $c^A = 1000$ MIPS,
- IT-Dienstleistung B: virtueller Server V^B mit $c^B = 2000$ MIPS und
- IT-Dienstleistung C: virtueller Server V^C mit $c^C = 3000$ MIPS.

Die virtuellen Server werden auf physischen Servern eingerichtet, deren Nettokapazität als technisches Leistungsvermögen 6000 MIPS beträgt. Es wird eine Virtualisierungstechnik unterstellt, die eine dynamische Zuordnung der Kapazität des physischen Servers auf die virtuellen Server zulässt. Betrachtet werden drei physische Server P_1, P_3 und P_3. Auf den physischen Servern werden bereits Fertigungsaufträge F für IT-Dienstleistungen A, B und C ausgeführt. Auf jedem physischen Server wurde im letzten VRRP-Lauf ein virtueller Server I zur Erbringung der IT-Dienstleistung B eingerichtet. Abbildung 6.12 veranschaulicht die Zuordnungen. Für einen Kunden in Berlin liegt ein neuer Fertigungsauftrag F^B zur Produktion der IT-Dienstleistung B vor. Es ist zu entscheiden, welcher virtuelle Server I^B für den Fertigungsauftrag eingesetzt wird. Die Fertigungsaufträge werden für Kunden in Berlin, London und Moskau ausgeführt. Durch geeignete Monitoringwerkzeuge wurden Lastprofile für die virtuellen Server der Fertigungsaufträge ermittelt. Tabelle A.3 im Anhang zeigt die Lastprofile im Zeitverlauf aus Sicht des Kunden und lokal aus Sicht des Rechenzentrums.
Die Lastprofile der physischen Server werden aus den Lastprofilen der Fertigungsaufträge ermittelt, die auf den physischen Servern ausgeführt werden. Hierzu werden die lokalen Lastprofile aus Sicht des Rechenzentrums verwendet. Für das Zeitintervall von 0:00 Uhr bis 1:00 Uhr beträgt der Kapazitätsbedarf des physischen Servers P_2 in MIPS:

$$\begin{aligned} P_2 &= F_2^B + F_3^B \\ &= 1019 + 1013 \\ &= 1032 \end{aligned} \tag{6.66}$$

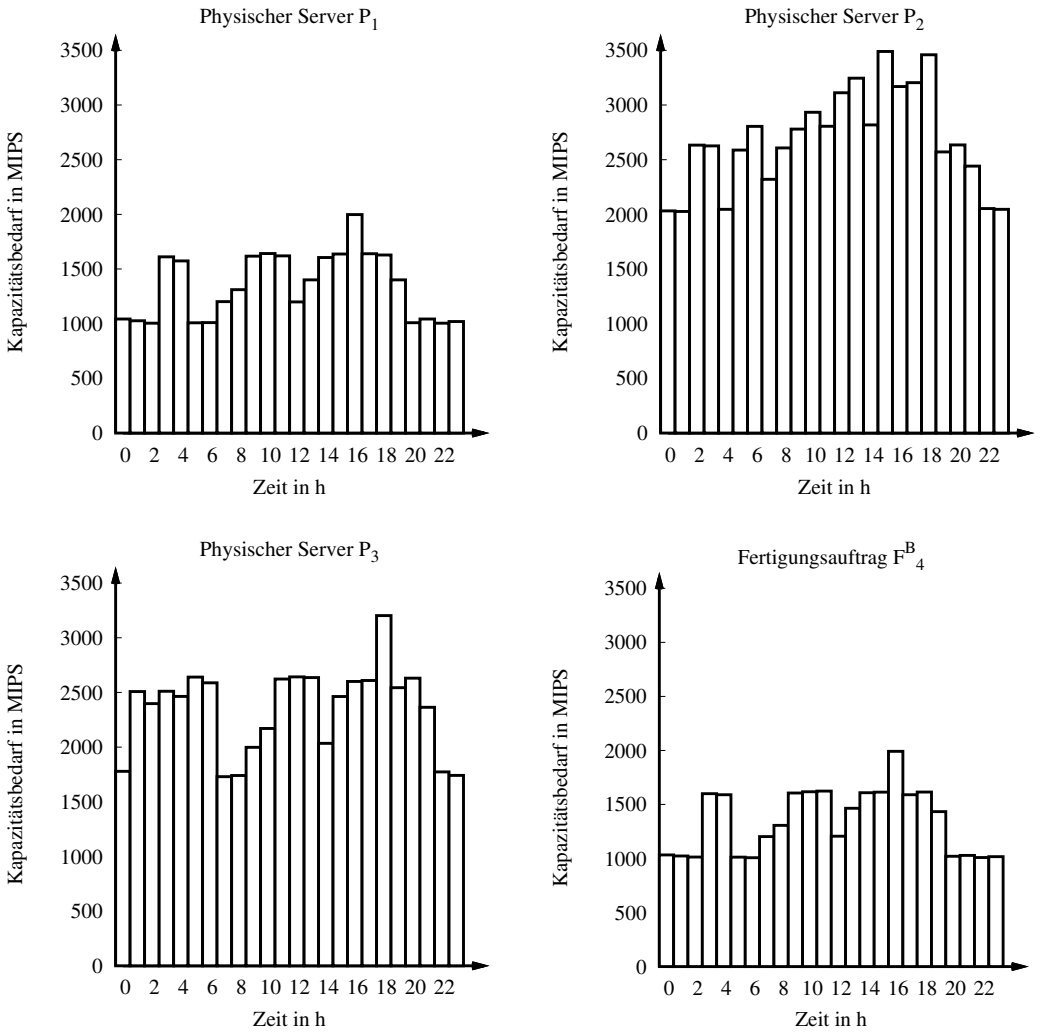

Abbildung 6.13: Grafische Darstellung der Lastprofile

Die weiteren Ergebnisse sind in Tabelle A.4 im Anhang zusammengefasst und in Abbildung 6.13 grafisch dargestellt. Aus den Daten der Lastprofile der Fertigungsaufträge F_1^B, F_2^B und F_3^B wird das erwartete Lastprofil des neuen Fertigungsauftrags F_4^B ermittelt. Hierzu werden die Durchschnittswerte aus den Kapazitätsbedarfen eines Zeitintervalls berechnet. Bei der Ermittlung der Durchschnittswerte werden die Lastprofile aus zeitlicher Sicht des Kunden verwendet.Für das Zeitintervall von 0:00 Uhr bis 1:00 Uhr beträgt der

erwartete Kapazitätsbedarf des Fertigungsauftrags F_4^B in MIPS:

$$\begin{aligned} F_4^B &= \frac{F_1^B + F_2^B + F_3^B}{3} \\ &= \frac{1043 + 1023 + 1031}{3} \\ &\approx 1033 \end{aligned} \tag{6.67}$$

Das vollständige Lastprofil ist in Tabelle A.4 im Anhang aufgeführt und in Abbildung 6.13 grafisch dargestellt. Im nächsten Schritt kann der Korrelationskoeffizient von Bravais-Pearson zwischen den physischen Servern P_1, P_2, P_3 und dem Fertigungsauftrag F_4^B berechnet werden. Nach Formel 6.64 lassen sich folgende Korrelationskoeffizienten ermitteln:

$$\begin{aligned} r(P_1F_4^B) &= 1 \\ r(P_2F_4^B) &= 0,6 \\ r(P_3F_4^B) &= 0,28 \end{aligned} \tag{6.68}$$

Für die Korrelationskoeffizienten gilt:

$$r(P_1F_4^B) > r(P_2F_4^B) > r(P_3F_4^B) \tag{6.69}$$

Der physische Server P_1 weist den größten linearen Zusammenhang zum Fertigungsauftrag F_4^B auf, da bereits ein Fertigungsauftrag für eine identische IT-Dienstleistung in derselben Zeitzone ausgeführt wird. Deshalb ist dieser Server ungeeignet für die Ausführung des Fertigungsauftrags F_4^B, obwohl er die geringste Auslastung mit Fertigungsaufträgen hat. Der niedrigste Korrelationskoeffizient besteht zwischen dem physischen Server P_3 und dem Fertigungsauftrag F_4^B. Der Fertigungsauftrag wird dem physischen Server P_3 zugeordnet. Im Fertigungsauftrag wird das Planbetriebsmittel durch den physischen Server P_3 substituiert. Nach der Freigabe des Fertigungsauftrags kann die Produktion begonnen werden.

6.3 Zusammenfassung

Aus Sicht der Kapazitätswirtschaft ist es stets sinnvoll, die virtualisierten Betriebsmittel auf Vorrat bereitzustellen. Für den Fall, dass sich das technische Leistungsvermögen eines virtuellen Betriebsmittels anhand einer Kapazitätsrestriktion abbilden lässt, konnte die Hypothese durch ein mathematisch-formales Modell bestätigt werden.
Wird das technische Leistungsvermögen eines virtuellen Betriebsmittels durch mehrere Kapazitätsrestriktionen beschrieben, wird die Gültigkeit der Hypothese ebenfalls vermutet. Es besteht jedoch noch weiterer Forschungsbedarf für den Beweis dieser Vermutung.
Die Zuordnung von virtualisierten Betriebsmitteln zu physischen Betriebsmitteln lässt sich als Bin-Packing-Problem formulieren. Heuristiken zur Lösung des Bin-Packing-Problems können zur Automatisierung der Betriebsmittelzuordnung eingesetzt werden. Die in diesem Kapitel gewonnenen Erkenntnisse vervollständigen somit das entwickelte Referenzmodell zur Abbildung der Produktion von IT-Dienstleistungen in Rechenzentren.

Kapitel 7

Zusammenfassung

Die Verfahren der Produktionsplanung und -steuerung lassen sich auf den Betrieb eines Rechenzentrums anwenden. Die hergestellten Endprodukte eines Rechenzentrums sind IT-Produkte in Form von IT-Dienstleistungen. Der Rechenzentrumsbetrieb ist eine Produktionsform der auftragsorientierten Serienfertigung. Der Arbeitsplan zur Produktion einer IT-Dienstleistung lässt sich in die Phasen Bereitstellung und Betrieb der Informationsinfrastruktur unterteilen. Die Phase des Betriebs der Informationsinfrastruktur stellt die zu produzierende Serie dar.
Für die Kapazitätsbedarfsplanung können Prognoseverfahren eingesetzt werden, um den zukünftigen Bedarf an Betriebsmitteln zu extrapolieren. Die Steuerung des Kapazitätsangebots lässt sich auf Basis der (s, Q)-Politik durchführen. Durch Rückwärtsterminierung der Fertigungsaufträge wird sichergestellt, dass alle Arbeitsschritte zur Bereitstellung der Informationsinfrastruktur rechtzeitig abgeschlossen sind, um die Erbringung der IT-Dienstleistung in einem vorgegebenen zeitlichen Rahmen zu ermöglichen. Der Einsatz von Virtualisierungstechniken ermöglicht die Einrichtung von virtuellen Betriebsmitteln entsprechend den Fertigungsaufträgen.
Die beschriebenen Datenstrukturen und Methoden lassen sich in SAP ERP implementieren. Die Produktion von IT-Dienstleistungen wird hierbei als kundenauftragsorientierte Prozessfertigung abgebildet. Die bei der Ausführung eines Prozessauftrags entstehenden Bewegungsdaten ermöglichen eine qualitative Bewertung der Produktionsplanung und -steuerung.
Aus Sicht der Kapazitätswirtschaft lässt sich beschreiben, wie gut oder

schlecht die IT-Dienstleistung die Forderung an ihre Beschaffenheit erfüllt. Die Beschreibung erfolgt anhand der Qualitätsmerkmale Bereitstellung zum vereinbarten Termin, zeitliche Verfügbarkeit und technische Verfügbarkeit während der Leistungserstellung. Es ist sinnvoll, die Qualitätsmerkmale aus Anbieter- und Kundensicht zu betrachten, da beide Gruppen für die Qualitätsbewertung der Kapazitätswirtschaft relevant sind und Qualitätsmerkmale aus verschiedenen Sichten verschiedenen Qualitätsstufen zugeordnet werden können. Die beschriebenen Qualitätsmerkmale lassen sich auf die Qualitätsdimensionen Potentialqualität, Prozessqualität und Ergebnisqualität abbilden. Da bei Dienstleistungen Produktion und Konsumtion simultan stattfinden, werden dieselben Qualitätsmerkmale den Dimensionen Prozessqualität und Ergebnisqualität zugeordnet. Aus Kundensicht beschreiben die technische und zeitliche Verfügbarkeit die Ergebnisqualität während diese Qualitätsmerkmale aus Anbietersicht die Prozessqualität beschreiben. Zur Abbildung der Qualitätsmerkmale auf Qualitätsstufen eignet sich eine Ordinalskala.

Eine Bewertung der Qualität der Kapazitätswirtschaft ist anhand der beschriebenen Qualitätsmerkmale möglich. Für eine Gesamtbewertung der Qualität der IT-Dienstleistung ist eine isolierte Betrachtung der Qualität der Kapazitätswirtschaft jedoch nicht ausreichend, da IT-Dienstleistungen in der Praxis häufig als Dienstleistungsbündel angeboten werden. So werden in der Regel Infrastrukturdienstleistungen des Rechenzentrums gemeinsam mit Unterstützungsdienstleistungen angeboten (vgl. Zarnekow, 2007, S. 11). Für eine umfassende Bewertung der Qualität einer IT-Dienstleistung besteht weiterer Forschungsbedarf.

Die Qualität der Kapazitätswirtschaft lässt sich durch die Planung des Kapazitätsangebots beeinflussen. Das Kapazitätsangebot eines Betriebsmittels beschreibt sein technisches Leistungsvermögen. Die Inanspruchnahme des technischen Leistungsvermögens entspricht dem Kapazitätsbedarf. Die Auslastung eines Betriebsmittels ist das Verhältnis aus Kapazitätsbedarf und Kapazitätsangebot. Anhand eines Warteschlangenmodells lässt sich zeigen, dass das Kapazitätsangebot eines Servers nicht in vollem Umfang in Anspruch genommen werden kann. Für eine gegebene Ankunftsrate λ

von Serveranfragen lässt sich die kritische Auslastung eines Servers für ein gefordertes Antwortzeitverhalten $\overline{T}$ bestimmen.
Das Niveau der Ankunftsrate λ ist nicht im Zeitverlauf konstant, sondern unterliegt statistischen Schwankungen. Ist die Verteilungsfunktion des Niveaus der Ankunftsrate bekannt, lässt sich anhand des Erwartungswerts und der Varianz der Kapazitätsbedarf ermitteln, der für einen bestimmten Anteil an Serveranfragen nicht überschritten wird. Aus dem Kapazitätsbedarf und der kritischen Serverauslastung lässt sich das notwendige Kapazitätsangebot bestimmen.
Wird dem Niveau der Ankunftsrate λ eine durch die Weibullverteilung approximierte Normalverteilung unterstellt, lässt sich zeigen, dass durch den Einsatz von Virtualisierung Kapazitätsersparnisse erzielt werden. Bei negativer Korrelation des Kapazitätsbedarfs zweier virtueller Systeme ist die Kapazitätsersparnis höher als im unkorrelierten Fall.
Die Kapazität der Hardwarekomponenten eines Rechenzentrums wächst exponentiell. Der Rebound-Effekt beschreibt den Umstand, dass diese zusätzlichen Kapazitäten vom Konsumenten auch verbraucht werden.
Lineare Regressionsrechnung, exponentielle Glättung zweiter Ordnung und das Verfahren von Holt lassen sich zur Prognose des durchschnittlichen Kapazitätsbedarfs der Hardwarekomponenten eines Rechenzentrums einsetzen. Als Datenbasis dienen die Beobachtungsdaten der durchschnittlichen Kapazitätsauslastung der vorausgegangenen Perioden. Das Kapazitätsangebot der Hardwarekomponenten wird üblicherweise so dimensioniert, dass diese nicht voll ausgelastet werden. Die Kapazitätsauslastungen der Hardwarekomponenten eignen sich deshalb als Beobachtungsdaten. Die Datenerhebung erfolgt durch den Einsatz geeigneter Messwerkzeuge.
Das Ergebnis der Prognose ist der zu erwartende durchschnittliche Kapazitätsbedarf der folgenden Periode. Der Einsatz von Prognoseverfahren zur Ermittlung des durchschnittlichen Kapazitätsbedarfs an Personal ist nicht sinnvoll, da dieser Potentialfaktor in der Regel voll ausgelastet ist. Die Ergebnisse der Kapazitätsplanung können jedoch als Entscheidungsgrundlage der Personalplanung dienen.
Auf Basis der Bedarfsprognose werden die physischen Betriebsmittel für

die kommende Planungsperiode beschafft. Auf den physischen Betriebsmitteln sind die virtualisierten Betriebsmittel einzurichten. Die Zuordnung von virtuellen zu physischen Betriebsmitteln lässt sich als Bin-Packing-Problem beschreiben. Für die Zuordnung der virtuellen Betriebsmittel können Online- und Offline-Bin-Packing Algorithmen eingesetzt werden. Sollen die virtuellen Betriebsmittel für jeden Fertigungsauftrag einzeln angelegt werden, entspricht dieser Vorgang einem Online-Bin-Packing. Die Vorratsbereitstellung virtueller Betriebsmittel lässt sich als Offline-Bin-Packing darstellen. Werden die virtuellen Betriebsmittel nur durch ein Kapazitätsmerkmal beschrieben, lässt sich zeigen, dass die Vorratsbereitstellung aus Sicht der Kapazitätswirtschaft bessere Ergebnisse erzielt. Um nachzuweisen, dass die Vorratsbereitstellung bessere Ergebnisse liefert, wenn die virtuellen Betriebsmittel durch mehrere Kapazitätsmerkmale beschrieben werden, besteht weiterer Forschungsbedarf.

Durch geeignete Monitoringwerkzeuge lassen sich Lastprofile der in Fertigungsaufträgen verwendeten virtuellen Betriebsmittel erstellen. Auf Grundlage dieser Lastprofile können erwartete Lastprofile für neue Fertigungsaufträge derselben IT-Dienstleistung erstellt werden. Die Auswahl des physischen Servers auf dem ein neuer Fertigungsauftrag ausgeführt werden soll, kann anhand dieser Lastprofile vorgenommen werden. Hierzu werden die Korrelationskoeffizienten von Bravais-Pearson zwischen dem erwarteten Lastprofil des neuen Fertigungsauftrags und den Lastprofilen der physischen Server ermittelt. Der Fertigungsauftrag wird auf dem physischen Server mit dem niedrigsten Korrelationskoeffizienten ausgeführt. Durch die Berücksichtigung der Korrelation werden Reservekapazitäten geschaffen, die für unerwartet auftretende Spitzenbedarfe des technischen Leistungsvermögens zur Verfügung stehen.

Forschungsziel der Arbeit war es, das Einsatzgebiet von Systemen zur Produktionsplanung und -steuerung zu erweitern. Es wurde aufgezeigt, dass es möglich ist, diese Systeme zur Abbildung der Produktion von IT-Dienstleistungen in Rechenzentren einzusetzen. Ein neues Anwendungsgebiet der Wirtschaft für ein existierendes Informations- und Kommunikationssys-

tem konnte erschlossen werden. Im Rahmen der Arbeit wurden folgende Hypothesen bestätigt:

- IT-Dienstleistungen lassen sich in einem System zur Produktionsplanung und -steuerung abbilden.
- Der Einsatz von Virtualisierungstechniken führt zu Einsparungen des Kapazitätsangebots.
- Der Kapazitätsbedarf an Potentialfaktoren wächst auch bei konstanter Nachfrage nach IT-Dienstleistungen exponentiell.
- Aus Sicht der Kapazitätswirtschaft ist es stets sinnvoll, die virtualisierten Betriebsmittel auf Vorrat bereitzustellen.

Zur Bestätigung der ersten Hypothese wurde ein Referenzmodell zur Abbildung der Produktion von IT-Dienstleistungen in einem System zur Produktionsplanung und -steuerung entwickelt. Das Referenzmodell wurde anhand eines Prototypen evaluiert. Anhand des Referenzmodells und des Prototypen wurden weitere Forschungsfragen generiert. Diese Forschungsfragen wurden durch mathematisch-formale Modelle beantwortet.
Zum Forschungsthema Produktion von IT-Dienstleistungen in Rechenzentren besteht weiterer Forschungsbedarf. Nachdem gezeigt wurde, dass der Einsatz von Systemen zur Produktionsplanung und -steuerung sinnvoll ist, soll im nächsten Schritt die Einsatzmöglichkeit von Advanced-Planning-Systemen (vgl. Fleischmann et al., 2007, S. 81 ff.) überprüft werden.
Im Rahmen einer weiteren Dissertation wird eine Fallstudie zur Einsparung des Kapazitätsangebots durch den Einsatz von Virtualisierungstechniken durchgeführt. Das Modell zur Planung des Kapazitätsangebots soll anhand der Ergebnisse der Fallstudie überprüft werden.
Zu den Themen Wartung und Störungsmanagement im Rechenzentrum wird ebenfalls eine weitere Dissertation erstellt. Die Forschungsergebnisse dieser Arbeit sind mit den Erkenntnissen der vorliegenden Arbeit zu vergleichen.

Anhang

Vorgang	**Phase**	**Ressource**	**Bezeichnung**	**Dauer**	**Mengenabhängig**
0010		SE-00000	Vorgang Server		
	0020	SE-00000	Installation Betriebssystem	1 h	nein
	0030	SE-00000	Installation Datenbankmanagementsystem	1 h	nein
	0040	SE-00000	Installation SAP BW	2 h	nein
	0050	SE-00000	Betrieb SAP BW	24 h	ja (Basismenge 1 TAG)
0060		ST-00000	Vorgang Storagepartition		
	0070	ST-00000	Installation Betriebssystem	1 h	nein
	0080	ST-00000	Installation Datenbankmanagementsystem	1 h	nein
	0090	ST-00000	Installation SAP BW	2 h	nein
	0100	ST-00000	Betrieb SAP BW	24 h	ja (Basismenge 1 TAG)
0110		VL-00000	Vorgang VLAN		
	0120	VL-00000	Installation Betriebssystem	1 h	nein
	0130	VL-00000	Installation Datenbankmanagementsystem	1 h	nein
	0140	VL-00000	Installation SAP BW	2 h	nein
	0150	VL-00000	Betrieb SAP BW	24 h	ja (Basismenge 1 TAG)
0160		ADMIN	Vorgang Administrator		
	0170	ADMIN	Installation Betriebssystem	1 h	nein
	0180	ADMIN	Installation Datenbankmanagementsystem	1 h	nein
	0190	ADMIN	Installation SAP BW	2 h	nein

Tabelle A.1: Planungsrezept für Bereitstellung und Betrieb eines SAP BW

Vorgang/ Phase		Ressource	Bezeichnung	Spätester Starttermin	Spätester Endtermin
0010		SE-00002	Vorgang Server		
	0020	SE-00002	Installation Betriebssystem	16.3.2009 13:00 Uhr	16.3.2009 14:00 Uhr
	0030	SE-00002	Installation Datenbankmanagementsystem	16.3.2009 14:00 Uhr	16.3.2009 15:00 Uhr
	0040	SE-00002	Installation SAP BW	16.3.2009 15:00 Uhr	16.3.2009 17:00 Uhr
	0050	SE-00002	Betrieb SAP BW	17.3.2009 00:00 Uhr	21.3.2009 24:00 Uhr
0060		ST-00002	Vorgang Storagepartition		
	0070	ST-00002	Installation Betriebssystem	16.3.2009 13:00 Uhr	16.3.2009 14:00 Uhr
	0080	ST-00002	Installation Datenbankmanagementsystem	16.3.2009 14:00 Uhr	16.3.2009 15:00 Uhr
	0090	ST-00002	Installation SAP BW	16.3.2009 15:00 Uhr	16.3.2009 17:00 Uhr
	0100	ST-00002	Betrieb SAP BW	17.3.2009 00:00 Uhr	21.3.2009 24:00 Uhr
0110		VL-00001	Vorgang VLAN		
	0120	VL-00001	Installation Betriebssystem	16.3.2009 13:00 Uhr	16.3.2009 14:00 Uhr
	0130	VL-00001	Installation Datenbankmanagementsystem	16.3.2009 14:00 Uhr	16.3.2009 15:00 Uhr
	0140	VL-00001	Installation SAP BW	16.3.2009 15:00 Uhr	16.3.2009 17:00 Uhr
	0150	VL-00001	Betrieb SAP BW	17.3.2009 00:00 Uhr	21.3.2009 24:00 Uhr
0160		ADMIN	Vorgang Administrator		
	0170	ADMIN	Installation Betriebssystem	16.3.2009 13:00 Uhr	16.3.2009 14:00 Uhr
	0180	ADMIN	Installation Datenbankmanagementsystem	16.3.2009 14:00 Uhr	16.3.2009 15:00 Uhr
	0190	ADMIN	Installation SAP BW	16.3.2009 15:00 Uhr	16.3.2009 17:00 Uhr

Tabelle A.2: Ergebnis der Terminierung und der Ressourcenauswahl

Zeit in h	Fertigungsauftrag F_1^A		Fertigungsauftrag F_1^B		Fertigungsauftrag F_2^B		Fertigungsauftrag F_3^B		Fertigungsauftrag F_1^C	
	Moskau UTC+3	Lokal UTC+1	Berlin UTC+1	Lokal UTC+1	London UTC+0	Lokal UTC+1	Moskau UTC+3	Lokal UTC+1	Moskau UTC+3	Lokal UTC+1
0	852	259	1043	1043	1023	1019	1031	1013	1546	1519
1	976	981	1027	1027	1019	1008	1023	1018	1534	1527
2	923	852	1005	1005	1008	1602	1027	1031	1540	1546
3	261	976	1611	1611	1602	1603	1586	1023	2379	1534
4	202	923	1574	1574	1603	1019	1591	1027	2386	1540
5	212	261	1008	1008	1019	1002	1012	1586	1518	2379
6	228	202	1010	1010	1002	1213	1008	1591	1512	2386
7	209	212	1202	1202	1213	1308	1193	1012	1789	1518
8	217	228	1311	1311	1308	1599	1302	1008	1953	1512
9	221	209	1618	1618	1599	1587	1601	1193	2401	1789
10	208	217	1642	1642	1587	1631	1623	1302	2434	1953
11	205	221	1620	1620	1631	1204	1621	1601	2431	2401
12	212	208	1199	1199	1204	1489	1215	1623	1822	2434
13	208	205	1401	1401	1489	1624	1503	1621	2254	2431
14	203	212	1605	1605	1624	1603	1598	1215	2397	1822
15	206	208	1637	1637	1603	1986	1602	1503	2403	2254
16	214	203	1998	1998	1986	1571	1992	1598	2988	2397
17	205	206	1640	1640	1571	1602	1559	1602	2338	2403
18	209	214	1628	1628	1602	1467	1614	1992	2421	2988
19	216	205	1401	1401	1467	1012	1432	1559	2148	2338
20	212	209	1009	1009	1012	1021	1041	1614	1561	2421
21	207	216	1043	1043	1021	1009	1023	1432	1534	2148
22	259	212	1005	1005	1009	1012	1013	1041	1519	1561
23	981	207	1020	1020	1012	1023	1018	1023	1527	1534

Tabelle A.3: Lastprofile der virtuellen Server

Zeit in h	physischer Server P_1	physischer Server P_2	physischer Server P_3	Fertigungsauftrag F_4^B	
	Lokal UTC+1	Lokal UTC+1	Lokal UTC+1	Berlin UTC+1	Lokal UTC+1
0	1043	2032	1778	1033	1033
1	1027	2026	2508	1023	1023
2	1005	2633	2398	1014	1014
3	1611	2626	2510	1600	1600
4	1574	2046	2463	1590	1590
5	1008	2588	2640	1013	1013
6	1010	2804	2588	1007	1007
7	1202	2320	1730	1203	1203
8	1311	2607	1740	1307	1307
9	1618	2780	1998	1606	1606
10	1642	2933	2170	1618	1618
11	1620	2805	2622	1624	1624
12	1199	3112	2642	1206	1206
13	1401	3245	2636	1465	1465
14	1605	2818	2034	1609	1609
15	1637	3489	2462	1614	1614
16	1998	3169	2600	1992	1992
17	1640	3204	2609	1590	1590
18	1628	3459	3202	1615	1615
19	1401	2571	2543	1434	1434
20	1009	2635	2630	1021	1021
21	1043	2441	2364	1029	1029
22	1005	2053	1773	1009	1009
23	1020	2046	1741	1017	1017

Tabelle A.4: Lastprofile der physischen Server und des Fertigungsauftrags F_4^B

Literaturverzeichnis

Adam, D. (1998): Produktions-Management. Wiesbaden: Gabler.

Adam, S. (1989): Optimierung der Anlageninstandhaltung, Verfügbarkeitsanforderung, Ausfallfolgekosten und Ausfallverhalten als Bestimmungsgrößen wirtschaftlich sinnvoller Instandhaltungsstrategien. Berlin: Erich-Schmidt-Verlag.

Alpar, P. (2000): Anwendungsorientierte Wirtschaftsinformatik : eine Einführung in die strategische Planung, Entwicklung und Nutzung von Informations- und Kommunikationssystemen. Braunschweig u. a.: Vieweg.

Arbeitskreis Publikation ITIL Version 3 Translation Project (2007): ITIL V 3 - Glossar (Englische Basisversion: 3.1.24). IT Service Managament Forum Deutschland e.V.

Balzert, H. (1998): Software-Management, Software-Qualitätssicherung, Unternehmensmodellierung. Heidelberg u. a.: Spektrum Akad. Verl.

Benz, J. und Höflinger, M. (2008): Logistikprozesse mit SAP: eine anwendungsbezogene Einführung. Wiesbaden: Vieweg + Teubner.

Bichler, M., Setzer, T. und Speitkamp, B. (2006): *Capacity Planning for Virtualized Servers.* In Workshop on Information Technologies and Systems (WITS), Milwaukee, Wisconsin, USA, 2006. München: Technische Universität.

Bolch, G. (1998): Queueing networks and Markov chains: modeling and performance evaluation with computer science applications. New York, NY u. a.: Wiley.

Bronshtein, I. N., Semendyayev, K. A., Musiol, G. und Muehlig, H. (2007): Handbook of Mathematics. Berlin u. a.: Springer.

Brown, R. G. (1984): Materials management systems: a modular library. Malabar, Fla.: Krieger.

Brown, R. G. und Meyer, R. F. (1960): *The Fundamental Theorem of Exponential Smoothing.* Operations research: the journal of the Operations Research Society of America, Band 9, Nr. 5, S. 673–685.

Bruhn, M. (2006): Qualitätsmanagement für Dienstleistungen: Grundlagen, Konzepte, Methoden. Berlin u. a.: Springer.

Buchsein, R., Victor, F., Günther, H. und Machmeier, V. (2008): IT-Management mit ITIL V3: Strategien, Kennzahlen, Umsetzung. Wiesbaden: Vieweg+Teubner / GWV Fachverlage GmbH, Wiesbaden.

Böhmann, T. und Krcmar, H. (2006): *Modulare Servicearchitekturen.* In H.-J. Bullinger und A.-W. Scheer (Hrsg.), Service Engineering: Entwicklung und Gestaltung innovativer Dienstleistungen. Berlin, Heidelberg: Springer.

Caprara, A., Kellerer, H. und Pferschy, U. (2002): *Approximation schemes for ordered vector packing problems.* Naval Research Logistics, Band 50, Nr. 1, S. 58–69.

Coffman, E. G., Garey, M. R. und Johnson, D. S. (1984): *Approximation Algorithms for Bin-Packing – An Updated Survey.* In G. Ausiello, M. Lucertini und P. Serafini (Hrsg.), Algorithm design for computer system design. Wien u. a.: Springer.

Coffman, E. G., Garey, M. R. und Johnson, D. S. (1997): *Approximation Algorithms for Bin Packing: A Survey.* In D. S. Hochbaum (Hrsg.), Approximation Algorithms for NP-Hard Problems. Boston, Mass. u. a.: PWS Publ.

Corsten, H. und Gössinger, R. (2007): Dienstleistungsmanagement. München u. a.: Oldenbourg.

Cramer, E. und Kamps, U. (2008): Grundlagen der Wahrscheinlichkeitsrechnung und Statistik. Berlin, Heidelberg: Springer-Verlag.

Dadam, P. (1996): Verteilte Datenbanken und Client/Server-Systeme: Grundlagen, Konzepte und Realisierungsformen. Berlin u. a.: Springer.

DeLurgio, S. A. (1998): Forecasting principles and applications. Boston, Mass. u. a.: McGraw-Hill.

Dempe, S. und Schreier, H. (2006): Operations Research: Deterministische Modelle und Methoden. Wiesbaden: B.G. Teubner Verlag / GWV Fachverlage GmbH, Wiesbaden.

Dickmann, P. (2007): Schlanker Materialfluss: mit Lean Production, Kanban und Innovationen. Berlin u. a.: Springer.

Donabedian, A. (1980): Explorations in quality assessment and monitoring Vol. 1: The definition of quality and approaches to its assessment. Ann Arbor, MI: Health Administration Press.

Dubey, S. Y. D. (1967): *Normal and Weibull distributions.* Naval Research Logistics Quarterly, Band 14, Nr. 1, S. 69–79.

Fahrmeir, L. (1996): Multivariate statistische Verfahren. Berlin u. a.: de Gruyter.

Fahrmeir, L. (2004): Statistik: der Weg zur Datenanalyse. Berlin u. a.: Springer.

Fleischmann, B., Meyr, H. und Wagner, M. (2007): *Advanced Planning.* In H. Stadtler und C. Kilger (Hrsg.), Supply Chain Management and Advanced Planning: Concepts, Models, Software, and Case Studies. Berlin, Heidelberg: Springer.

Frietzsche, U. und Maleri, R. (2006): *Dienstleistungsproduktion.* In H.-J. Bullinger und A.-W. Scheer (Hrsg.), Service Engineering: Entwicklung und Gestaltung innovativer Dienstleistungen. Berlin, Heidelberg: Springer.

Galambos, G. und Woeginger, G. J. (1995): *On-line bin packing - A restricted survey.* Mathematical Methods of Operations Research, Band 42, Nr. 1, S. 25–45.

Gardner, E. S. (1984): *The Fundamental Theorem of Exponential Smoothing.* Interfaces: an international journal of the Institute for Operations Research and the Management Sciences, Band 14, Nr. 3, S. 47–50.

Garey, M. R., Graham, R. L., Johnson, D. S. und Yao, A. (1976): *Resource Constrained Scheduling as Generalized Bin Packing.* Journal of combinatorial theory – Series A, Band 21, S. 257–298.

Garey, M. R., Graham, R. L. und Ullman, J. D. (1972): *Worst-case analysis of memory allocation algorithms.* In Proceedings of the fourth annual ACM symposium on Theory of computing. New York: ACM.

Gehweiler, J. und auf der Heide, F. M. (2008): *Bin Packing oder Wie bekomme ich die Klamotten in die Kisten?* In B. Vöcking, H. Alt, M. Dietzfelbinger, R. Reischuk, C. Scheideler, H. Vollmer und D. Wagner (Hrsg.), Taschenbuch der Algorithmen. Berlin, Heidelberg: Springer-Verlag.

Geiger, W. und Kotte, W. (2008): Handbuch Qualität: Grundlagen und Elemente des Qualitätsmanagements: Systeme – Perspektiven. Wiesbaden: Vieweg Sohn.

Ghorpade, S. R. und Limaye, B. V. (2006): A Course in Calculus and Real Analysis. New York, NY: Springer Science+Business Media, LLC.

Grawe, T. und Fähnrich, K.-P. (2008): *Service Engineering bei IT-Dienstleistern.* In K.-P. Fähnrich und C. Husen (Hrsg.), Entwicklung IT-basierter Dienstleistungen: Co-Design von Software und Services mit ServCASE. Heidelberg: Physica-Verlag.

Gutenberg, E. (1983): Grundlagen der Betriebswirtschaftslehre: Die Produktion. Berlin u. a.: Springer.

Haasis, H.-D. (2008): Produktions- und Logistikmanagement: Planung und Gestaltung von Wertschöpfungsprozessen. Wiesbaden: Gabler / GWV Fachverlage GmbH, Wiesbaden.

Hanisch, L., Osterburg, S. und Pinnow, A. (2009): *Forecasting Demand of Potential Factors in Data Centers*. Informatica Economica, Band 49, Nr. 1, S. 9–15.

Heinrich, L. J. (2005): *Forschungsmethodik einer Integrationsdisziplin: Ein Beitrag zur Geschichte der Wirtschaftsinformatik*. N.T.M. Internationale Zeitschrift für Geschichte und Ethik der Naturwissenschaften, Technik und Medizin, Band 13, S. 104–117.

Heinrich, L. J., Heinzl, A. und Roithmayr, F. (2004): Wirtschaftsinformatik-Lexikon. München: Oldenbourg.

Heinrich, L. J. und Lehner, F. (2005): Informationsmanagement: Planung, Überwachung und Steuerung der Informationsinfrastruktur. München u. a.: Oldenbourg.

Holt, C. C. (2004): *Forecasting seasonals and trends by exponentially weighted moving averages*. International journal of forecasting, Band 20, S. 5–10.

Hromkovič, J. (2007): Theoretische Informatik – Formale Sprachen, Berechenbarkeit, Komplexitätstheorie, Algorithmik, Kommunikation und Kryptographie. Wiesbaden: B.G. Teubner Verlag / GWV Fachverlage GmbH.

ISO/IEC 9126-1 (2001): Software Engineering - Product Quality - Part 1: Quality Model.

Johnson, D. S., Garey, M. R., Graham, R. L., Demers, A. und Ullman, J. D. (1974): *Worst Case Performance Bounds for Simple One-Dimensional Packing Algorithms*. SIAM journal on computing: a publication of the Society for Industrial and Applied Mathematics, Band 3, Nr. 4, S. 299–325.

Kern, W. (1992): Industrielle Produktionswirtschaft. Stuttgart: Poeschel.

Klein, R. (2005): Algorithmische Geometrie: Grundlagen, Methoden, Anwendungen. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Klimant, H., Piotraschke, R. und Schönfeld, D. (2006): Informations- und Kodierungstheorie. Stuttgart u. a.: Teubner.

Kollerer, H. (1978): Die betriebswirtschaftliche Problematik der Betriebsunterbrechungen: Planungsgrundlagen zur Berücksichtigung von Betriebsunterbrechungen im Rahmen der Unternehmenspolitik. Berlin: Erich Schmidt Verlag.

Kolmogoroff, A. (1936): *Zur Theorie der Markoffschen Ketten.* Mathematische Annalen, Band 113, Nr. 1, S. 155–160.

Korte, B. und Vygen, J. (2008): Kombinatorische Optimierung: Theorie und Algorithmen. Berlin, Heidelberg: Springer-Verlag.

Kuhlen, R. (2004): *Wissensökologie.* In R. Kuhlen, T. Seeger und D. Strauch (Hrsg.), Grundlagen der praktischen Information und Dokumentation. München: Saur.

Kurbel, K. (2005): Produktionsplanung und -steuerung im Enterprise Resource Planning und Supply Chain Management. München, Wien: Oldenbourg.

Langendörfer, H. (1992): Leistungsanalyse von Rechensystemen: Messen, Modellieren, Simulation. München u. a.: Hanser.

Lassmann, W. (2006): Wirtschaftsinformatik : Nachschlagewerk für Studium und Praxis. Wiesbaden: Betriebswirtschaftlicher Verlag Dr. Th. Gabler.

Lee, C. C. und Lee, D. T. (1985): *A simple online bin-packing algorithm.* Journal of the ACM, Band 32, Nr. 3, S. 562–572.

Little, J. D. (1961): *A proof for the queuing formula:* $L = \lambda W$. Operations research: the journal of the Operations Research Society of America, Band 9, Nr. 3, S. 383–387.

Maassen, A., Schoenen, M., Frick, D. und Gadatsch, A. (2006): Grundkurs SAP R/3: Lern- und Arbeitsbuch mit durchgehendem Fallbeispiel – Konzepte, Vorgehensweisen und Zusammenhänge mit Geschäftsprozessen. Wiesbaden: Friedr. Vieweg & Sohn Verlag/GWV Fachverlage GmbH.

Makridakis, S., Wheelwright, S. C. und McGee, V. E. (1983): Forecasting: methods and applications. New York u. a.: Wiley.

Malerie, R. und Frietzsche, U. (2008): Grundlagen der Dienstleistungsproduktion. Berlin, Heidelberg: Springer.

Mansouri-Samani, M. und Sloman, M. (1992): *Monitoring Distributed Systems (A Survey).* Technischer Bericht DOC92/23, Dept. of Computing, Imperial College, London, UK.

Mertens, P. (2004): Integrierte Informationsverarbeitung 1 — Operative Systeme in der Industrie. Wiesbaden: Gabler.

Miklitz, T., Buxmann, P. und Röddiger, A. (2006): *Standortplanung für Anbieter von IT-Services.* Wirtschaftsinformatik, Band 48, Nr. 6, S. 397–406.

Mißbach, M. (2005): Adaptive Hardware-Infrastrukturen für SAP: Lösungen und Kostenplanung. Bonn: Galileo Press.

Mosler, K. und Schmid, F. (2006): Wahrscheinlichkeitsrechnung und schließende Statistik. Berlin u. a.: Springer.

Murthy, D. N. P., Xie, M. und Jiang, R. (2003): Weibull models. Hoboken, NJ: Wiley-Interscience.

Márquez, A. C. (2007): The Maintenance Management Framework: Models and Methods for Complex Systems Maintenance. London: Springer-Verlag London Limited.

Neter, J., Kutner, M. H., Nachtsheim, C. J. und Wasserman, W. (1996): Applied linear statistical models. Boston, Mass. u. a.: McGraw-Hill.

OGC (2007a): Service design (SD): ITIL. London: TSO (The Stationery Office).

OGC (2007b): The Official Introduction to the ITIL Service Lifecycle. London: TSO (The Stationery Office).

Olbrich, A. (2008): ITIL kompakt und verständlich: Effizientes IT Service Management – Den Standard für IT-Prozesse kennenlernen, verstehen und erfolgreich in der Praxis umsetzen. Wiesbaden: Friedr. Vieweg & Sohn Verlag / GWV Fachverlage GmbH, Wiesbaden.

Osterburg, S. und Pinnow, A. (2009): *Produktion von IT-Dienstleistungen in Rechenzentren – Ein Implementierungsansatz für die Prozessfertigung in SAP ERP.* PPS-Management: Zeitschrift für Produktionsplanung und -steuerung, Ausgabe 1/2009, S. 24–26.

Osterburg, S., Pinnow, A., Rautenstrauch, C. und Winter, M. (2009a): *Neue Computing-Grundlagen für das Rechenzentrum.* Informatik-Spektrum: Organ der Gesellschaft für Informatik e.V., Band 32, Nr. 2, S. 118–126.

Osterburg, S., Pinnow, A. und Winter, M. (2008): *Das Rechenzentrum als Produktionsstätte für IT-Dienstleistungen: ein Kapazitätsmodell für Potentialfaktoren.* In M. Bichler (Hrsg.), Multikonferenz Wirtschaftsinformatik. Berlin: Gito-Verlag.

Osterburg, S., Pinnow, A. und Winter, M. (2009b): *Das Rechenzentrum als Produktionsstätte für IT-Dienstleistungen.* IM – Fachzeitschrift für Information Management & Consulting, Ausgabe 2/2009, S. 65–70.

Radermacher, F. J. (2000): *Building the Information Society: Labour Market Pressures, Globalisation, and the Political Goal of Sustainability as Challenges to the Regions in Europe.* In R. Sturm und G. Weinmann (Hrsg.), The information society and the regions in Europe: a British-German comparison. Baden-Baden: Nomos-Verl.-Ges.

Rasch, B., Friese, M., Hofmann, W. und Naumann, E. (2006): Quantitative Methoden: Einführung in die Statistik. Berlin u. a.: Springer Medizin Verlag.

Rautenstrauch, C. (1997): Effizienter Einsatz von Arbeitsplatzsystemen – Konzepte und Methoden des Persönlichen Informationsmanagements. Bonn u. a.: Addison-Wesley Longman.

Saake, G. und Heuer, A. (1999): Datenbanken: Implementierungstechniken. Bonn: MITP.

SAP (2007): *Produktionsplanung und -steuerung (PP).* In SAP-Bibliothek, Release 6.0 SR1. http://help.sap.com.

Schneider, H.-J. (1998): Lexikon Informatik und Datenverarbeitung. München u. a.: Oldenbourg.

Schuh, G. (2006): Produktionsplanung und -steuerung : Grundlagen, Gestaltung und Konzepte. Berlin, Heidelberg: Springer.

Schönsleben, P. (2007): Integrales Logistikmanagement: Operations und Supply Chain Management in umfassenden Wertschöpfungsnetzwerken. Berlin, Heidelberg: Springer-Verlag.

Seiden, S. S. (2002): *On the online bin packing problem.* Journal of the ACM, Band 49, Nr. 5, S. 640–671.

Smith, J. E. und Nair, R. (2005): Virtual machines: versatile platforms for systems and processes. Amsterdam u. a.: Elsevier.

Tanenbaum, A. S. (1999): Structured computer organization. Englewood Cliffs u. a.: Prentice-Hall Internat.

Tempelmeier, H. (2006): Material-Logistik: Modelle und Algorithmen für die Produktionsplanung und -steuerung in Advanced-Planning-Systemen. Berlin u. a.: Springer.

van Husen, C., Fuchs, B., Böttcher, M. und Meyer, K. (2008): *Ermittlung von Problemfeldern bei der Entwicklung IT-basierter Dienstleistungen.* In K.-P. Fähnrich und C. Husen (Hrsg.), Entwicklung IT-basierter Dienstleistungen: Co-Design von Software und Services mit ServCASE. Heidelberg: Physica-Verlag.

von Bertalanffy, L. (1980): General system theory : foundations, development, applications. New York: Braziller.

Weber, K. (1990): Wirtschaftsprognostik. München: Vahlen.

Weiß, C. (2005): Basiswissen Medizinische Statistik. Berlin u. a.: Springer Medizin Verlag.

Wilde, T. und Hess, T. (2007): *Forschungsmethoden der Wirtschaftsinformatik – Eine empirische Untersuchung.* Wirtschaftsinformatik, Band 49, Nr. 4, S. 280–287.

Winter, R. (2009): *Was ist eigentlich Grundlagenforschung in der Wirtschaftsinformatik?* Wirtschaftsinformatik, Band 51, Nr. 2, S. 223–231.

Wunsch, G. und Schreiber, H. (2006): Stochastische Systeme. Berlin u. a.: Springer.

Yao, A. C.-C. (1980): *New Algorithms for Bin Packing.* Journal of the ACM, Band 27, Nr. 2, S. 640–671.

Yue, M. (1991): *A simple proof of the inequality* $FFD(L) \leq 11/9OPT(L)+1$, $\forall L$ *for the FFD bin-packing algorithm.* Acta Mathematicae Applicatae Sinica (English Series), Band 7, Nr. 4, S. 321–331.

Zarnekow, R. (2007): Produktionsmanagement von IT-Dienstleistungen: Grundlagen, Aufgaben und Prozesse. Berlin u. a.: Springer.

Zarnekow, R., Brenner, W. und Pilgram, U. (2006): Integrated Information Management: Applying Successful Industrial Concepts in IT. Berlin, Heidelberg: Springer.

Zarnekow, R., Hochstein, A. und Brenner, W. (2005): Serviceorientiertes IT-Management: ITIL-Best-Practices und -Fallstudien. Berlin u. a.: Springer.

Ziemeck, H. A. I. (2006): Kunden- und mitarbeiterorientierte Organisationsgestaltung industrieller Dienstleistungsunternehmen. Wiesbaden: Dt. Univ.-Verl.

Zikun, W. und Xiangqun, Y. (1992): Birth and death processes and Markov chains. Berlin u. a.: Springer.

Zäpfel, G. (2001): Grundzüge des Produktions- und Logistikmanagement. München, Wien: Oldenbourg.