Kapazitätsplanung und Leistungsbewertung verteilter Systeme

Spezialvorlesung im Hauptstudium

Peter Buchholz
Informatik IV (Modellierung & Simulation)
Raum: GB V R 406a
Tel: (0231) 755 4746
Email: peter.buchholz@udo.edu
Sprechstunde Do 10.00-11.30 und n.V.

Daten und Orte:

Vorlesung:	Mo 8.15-10.00 GB IV, R 113
Vorlesung/Übung:	Do 14.15-16.00 GB IV, R 113

Material und Literatur zur Vorlesung

PDF-Dateien mit Folien im Internet:

http://ls4-www.cs.uni-dortmund.de/Lehre/05-42163.html

Folien vorlesungsbegleitend im Netz!!

Literatur (kleine Auswahl, weitere Literatur siehe www)

- D. A. Menasce, V.F. Almeida, L. W. Dowdy. Performance by Design. Prentice Hall 2004
- N. J. Gunther. Analyzing Computer System Performance with Perl::PDQ. Springer 2005.
- R. Jain. The art of computer systems performance analysis. Wiley 1991.
- E.D. Lazowska et al. Quantitative System Performance - Computer System Analysis Using Queueing Network Models. Prentice Hall 1984 (im Netz verfügbar!)

Tools die wir verwenden können:

- PDQ Sammlung von Perl/C-Prozeduren zur Analyse von Warteschlangennetzen (über www verfügbar)
- Excel Berechnungsprogramme aus dem Menasce Buch (über www verfügbar)
- HIT umfangreiches Modellierungs- und Analysewerkzeug

Gliederung

1.	Einführende Beispiele und Problemdefinition
2.	Kapazitätsplanung, Struktur des Umfelds
3.	Benutzerverhalten, Lastmodellierung
4.	Modellierung ausgewählter Komponenten
5.	Einfache Gesetzte der Leistungsanalyse,
6.	Ein Ansatz zur Kapazitätsplanung
7.	Einfache Warteschlangenmodelle
8.	Offene Warteschlangennetze
9.	Geschlossene Warteschlangennetze (eine Klasse)
10.	Geschlossene Warteschlangennetze (mehrere Klassen)
11.	Anwendungsbeispiele
12.	Zuverlässigkeit und Verfügbarkeit
13.	Aggregierungstechniken
14.	Approximative Mittelwertanalyse
15.	Hierarchische Netze
16.	Charakterisierung der Systemlast
17.	Benchmarks
18.	Messung
19.	Lastvorhersage
20.	Resümee

Ergänzung durch andere Vorlesungen, Seminare und Projektgruppen aus dem Gebiet der Modellierung und Simulation

Wahlpflichtveranstaltung:
Modellgestützt Analyse und Optimierung

Themenbereiche für Spezialvorlesungen, Seminare und PGs:

- Leistungs- und Zuverlässigkeitsanalyse von Rechen-, Kommunikations- und Materialflusssystemen
- Software zur Modellierung und Analyse
- Formale Modelle
- Modellierung und Simulation
- Computernumerik und paralleles Rechnen

Schwerpunktgebiete:
2 Rechnerarchitektur, eingebettete Systeme, Simulation &
3 Verteilte Systeme

Ziele der Vorlesung:

- Vermittlung eines ingenieurwissenschaftlichen Vorgehens zur quantitativen Analyse von verteilten Systemen (und speziell Client-Server Systemen)

- Einsatz der Techniken zur Planung von Installationen

Was wird dazu benötigt?

- Definition von Metriken, die Leistung umfassen

- Messung von Leistung

- Analyse der Messergebnisse

- Modellierung der Systeme

- Analyse der Modelle

- Entwicklung von Alternativen unter Kosten- und Leistungsgesichtspunkten

- Vorhersagemethoden für zukünftige Lasten

1.1 Einige Beispiele

Beispiele aus Mena02

Migration von einem Mainframe auf ein Client-Server System

Szenario Autovermietung mit

- 500000 PKWs an 3500 Standorten
- Reservierung an den Standorten oder telefonisch bei 1800 Mitarbeitern (rund um die Uhr)
- im Durchschnitt 360000 Reservierung pro Tag
- davon 60% (216000) während 12 Stunden
  d.h. 21667 Transaktionen per Std bzw.
  6 Transaktionen pro Sekunde

Jetzige Konfiguration

- Mainframe mit 1800 Terminals
- pro Standort Terminals per Konzentrator und Standleitung mit Mainframe verbunden

Geplante Migration auf Client-Server System

- lokales Netz und lokaler Server für jeden Standort
- Verbindung über ein WAN zur Zentrale

Folgende Anfragen existieren

- lokale Reservierung
- telefonische Reservierung
- technische Hilfe
- Abholung/Rückgabe von Autos

Leistungsanforderungen

- Antworten dürfen im Mittel nicht länger als 3 Sekunden dauern

Fragestellungen bei der Migration

- welcher Server in den Zweigstellen?
- welcher Transaktionsmonitor?
- welche Datenbank und welche Rechnerkonfiguration in der Zentrale?
- welche lokale Netzwerktechnologie?
- welche Bandbreite für das WAN?

Zusammensetzung der Antwortzeit (Bottleneckanalyse)

Komponente	Prozentsatz
Client Workstation	5
LAN	5
Lokaler Server	25
WAN	10
LAN Zentrale	4
DB Server	51

Planung für die Zukunft

durch Werbemaßnahmen Erweiterung des Umsatzes

Szenarien. Steigerung um 5%, 10% und 15%

Wie verändern sich die Antwortzeiten bei gleicher
IT Infrastruktur?

Transaktion	akt.	+5%	+10%	+15%
lokale Reservierung	1.28	1.67	2.45	5.06
telefonische Reserv.	0.64	0.87	1.37	3.20
technische Hilfe	0.64	0.76	0.94	1.23
Abholung/Rückgabe	0.85	1.16	1.82	4.24

Ergebnis: Steigerungen um 5% und 10% sind verkraftbar
Steigerung von 15% führt bei 3 von 4
"Lastklassen" zur Überschreitung des erlaubten
Limits (System *saturiert*)

Es fällt auf:
Kein lineares Wachstum der Antwortzeit mit der Last!

Warum werden die Antwortzeiten zu groß?

Antwort aus den Modellen
(die wir im Verlauf der Vorlesung kennenlernen)

DB Server verbraucht die meiste Zeit

Abhilfe: Leistungsverbesserung am DB Server

z.B. zusätzliche Platten, größerer Cache, ...

Leistung von Web Servern

Gebrauchtwagenverkauf über das Internet

- PKW nach bestimmten Kriterien suchen
- Liste der verfügbaren PKWs versenden
- Detailinformation abrufen
- Kauf tätigen bzw. Wagen reservieren

Beobachtung: Suchen ist die kritische Transaktion

- Wartezeit >4 Sek. Abbruch 60% Transaktionen
- Wartezeit >6 Sek. Abbruch 95% Transaktionen

Annahme 5% aller Suchtransaktionen führen zum Kauf
mit einem durchschnittlichen Umsatz von 1800 $

Auswirkungen Steigerung der Anfragen?

	akt.	+10%	+20%	+30%
Suchtr. pro Tag	92448	101693	110.938	120182
Antwortzeit in Sek.	2.86	3.80	5.67	11.28
Abbruch Trans. %	0	0	60	95
Verkäufe pro Tag	4622	5085	2219	300
Umsatz pro Tag	83203	91524	39938	5408
pot. Umsatz p. T.	83203	91524	99844	108164
pot. Verlust p. T.	-	-	59906	102756

Umsatz bricht bei Steigerung um 20% ein

Leistung eines Intranets

Firma mit 60000 Beschäftigten

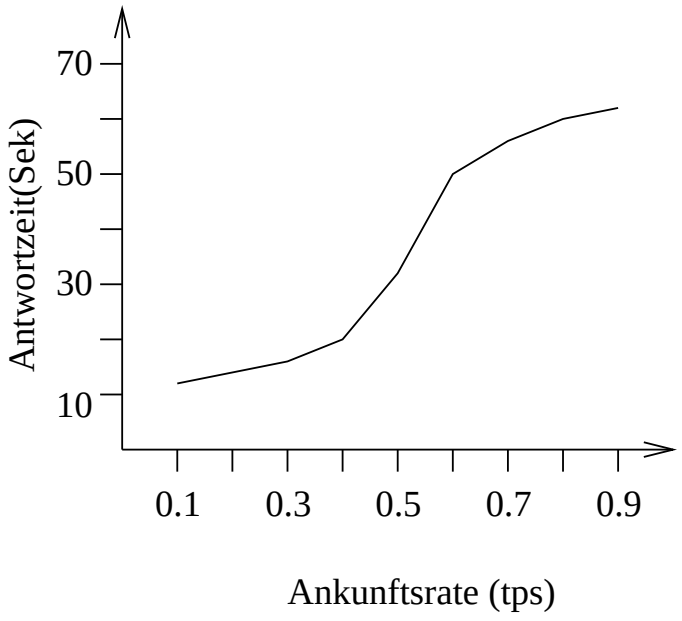
- jeder Beschäftigte arbeitet am PC
- Intranet wird installiert, eine Aufgabe Unterstützung bei PC Problemen
  - ▷ DB mit Antworten auf FAQs
  - ▷ Beschreibung von Problemen, die nicht in der FAQ-DB beantwortet werden

10% der Beschäftigten stellen eine Anfrage pro Tag

70% der Anfragen von 10-12 und 14-16 Uhr

d.h. 0.29 Anfragen pro Sek. an den Server

Frage: Was passiert nach Installation einer neue BS Version, wenn die Zahl der Anfragen steigt?



1.2 Quantitative Maße in verteilten Systemen

Quantitative Anforderungen an verteilte Systeme werden heute oft als Dienstqualität (Quality of Service QoS) subsumiert

Darunter fast man sehr unterschiedliche Maße, wie

- Antwortzeit
- Durchsatz
- Verfügbarkeit
- Zuverlässigkeit
- Sicherheit
- Skalierbarkeit
- Erweiterbarkeit
- ...

System erreicht QoS, falls verschiedene dieser Maße gewisse Werte erreichen (d.h. je nach Maß gewisse Werte unter- oder überschreiten)

Antwortzeiten

Aus Benutzersicht wichtigste Metrik:

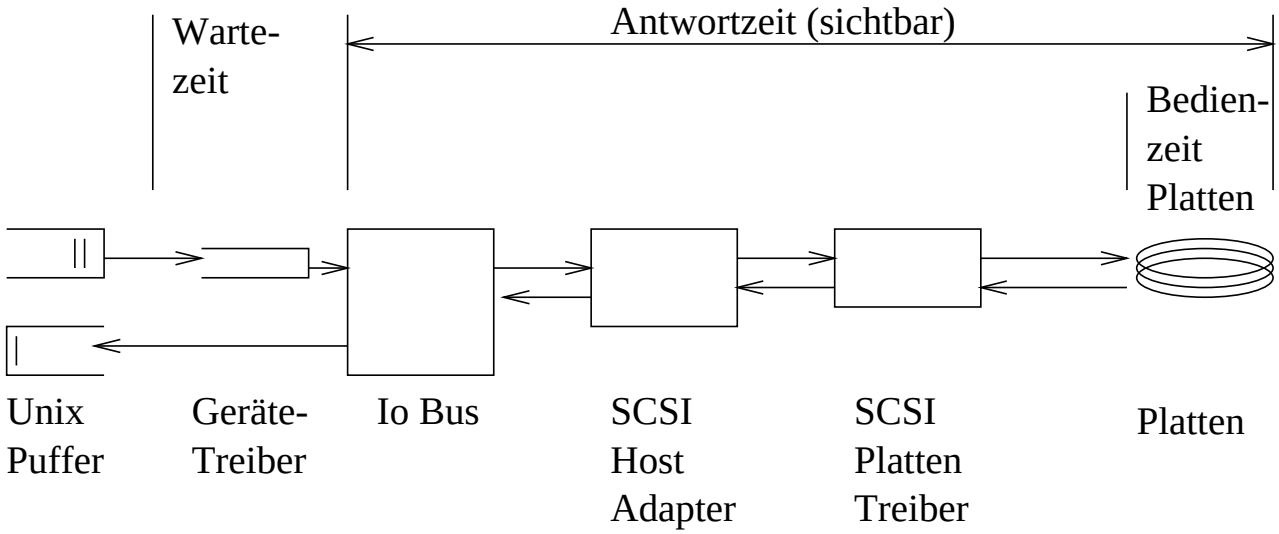
Zeit zwischen Absetzen einer Anfrage und
deren Beantwortung

Typischerweise setzt sich Antwortzeit aus
mehreren Komponenten zusammen,
die jeweils wieder aus Bedienzeiten und Wartezeiten
bestehen

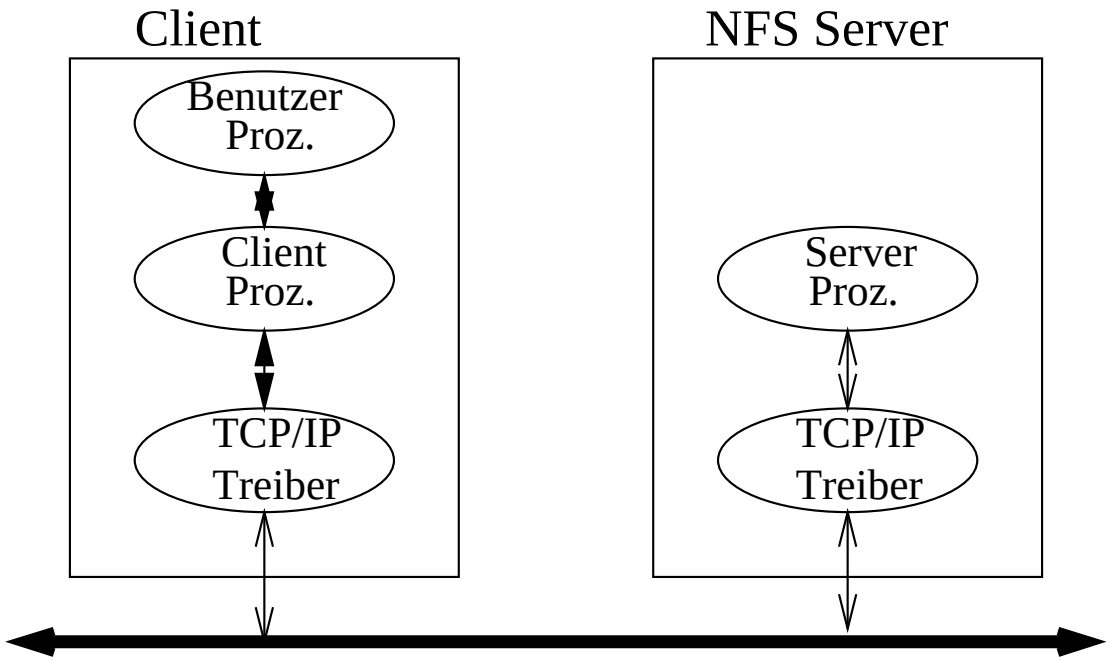
Erste Schritte zur Analyse:

Zerlegen der Antwortzeit in ihre Bestandteile und
Analyse der Bestandteile zur Bestimmung
von Flaschenhälsen

Bsp. SCSI Schreiboperation



NFS Leseoperation



Messungen haben folgende Werte ergeben

- Übertragungszeit Anfrage Ethernet 0.16 ms
- Bearbeitungszeiten (IP, UDP, RPC, NFS) je Richtung 3.5ms
- Dateibearbeitung je Richtung 1.5 ms
- Ein-/Ausgabetreiber je Richtung 0.37 ms
- Plattenzugriff 29.5 ms
- Rückübertragung Ethernet 6.8ms

$\Rightarrow$ Flaschenhals Plattenzugriff
(u.U. weiter zerlegen und analysieren)

Durchsatz

Rate mit der Aufgaben (Jobs, Tasks, Transaktionen,...) erledigt werden (in Anzahl pro Zeiteinheit)

Beispiele

System	Metrik
OLTP System	Transaktionen pro Sek. (tps)
Web-Seite	HTTP Anfrage/Sek.
	Seitenzugriffe / Sek.
	Bytes / Sek.
E-Comm. Site	Web Interaktionen /sek (WIPS)
	Session / Sek.
	Searches / Sek.
Router	Pakete / Sek.
	MByte / Sek.
CPU	Millionen Instr. / Sek. (MIPS)
	Fließkommaoper. / Sek. (FLOPS)
Platte	I/O-Operationen /Sek.
	Übertragungsmenge KByte /Sek.

Praktisches Problem oft Definition einer Lasteinheit

z.B. Was ist eine Transaktion ?

Festlegung teilweise durch Definition von Standardfällen (siehe auch Benchmarks später in der Vorlesung)

Zeiten in Systemen mit Ausfall und Reparatur

Neben der Leistung eines Systems ausgedrückt durch Antwortzeiten und ähnliche Maße hat die Verfügbarkeit eine große Bedeutung

- $MTTF$ Mean Time To Failure
  (mittlere Zeit bis zum Auftreten des nächsten Fehlers)

- $MTBF$ Mean Time Between Failure
  (mittlere Zeit zwischen zwei Fehlern)

- $MTTR$ Mean Time To Repair
  (mittlere Reparaturzeit)

- Verfügbarkeit (oder Availability) $A = \frac{MTBF}{MTBF+MTTR}$

- Zuverlässigkeit (oder Reliability) Wahrscheinlichkeit, dass ein System über einen Zeitraum korrekt arbeitet

Oftmals wird auch die Kombination von Leistung und Zuverlässigkeit betrachtet (performability)

Was leistet ein System unter Berücksichtigung von Ausfällen und Reparatur?

Weitere (nur teilweise quantitative) Maße

- Sicherheit
  Garantie der Einhaltung von Zugriffsbeschränkungen, Datenintegrität, Authentifizierung, ...

  Quantitative Aspekte durch

  – Aufwandsabschätzung für zusätzliche Sicherungsmechanismen
  – Berechnung der Wahrscheinlichkeit von Sicherheitsverletzungen

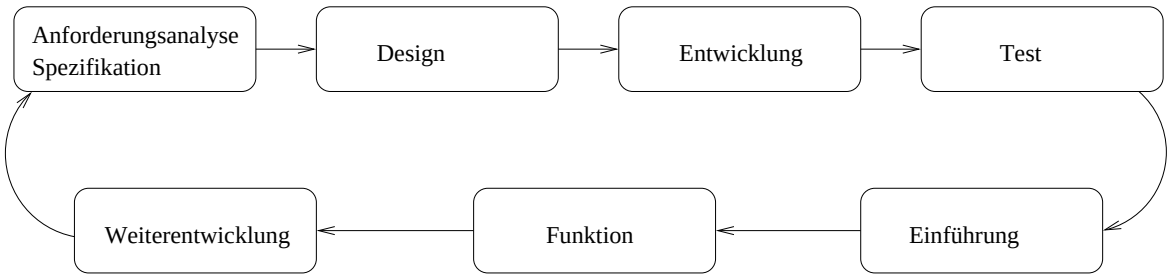
- Skalierbarkeit
  Anpassung eines Systems an wachsende Lasten

- Erweiterbarkeit
  Möglichkeit der Einführung neuer Funktionen, ohne Leistungsverluste bei bisherigen Funktionen

1.3 Lebenzyklus komplexer Systeme

Übliches Vorgehen beim Entwurf, Bau und Betrieb von Systemen (IT-Systemen, aber auch in anderen Bereichen)

Idealisierte/Vereinfachte Darstellung:



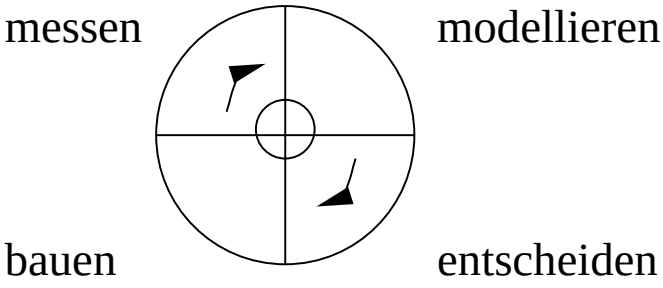
QoS ist Teil der Spezifikation
(siehe SLAs im nächsten Abschnitt)

Leistungsanalyse oft erst nachträglich, wenn QoS nicht erreicht wird (damit teure und aufwändige Redesigns)

Besser:

- Leistungsanalyse als Teil des Entwurfsprozesses
  (siehe folgende Folien)
- Leistungsanalyse als Teil des Betriebs
  (noch in den Anfängen)
- Leistungsanalyse als Teil der gesamten Lebenszyklus
  (wichtiger Teil ist Kapazitätsplanung, siehe nächster Abschnitt)

Vorgehen beim Entwurf von Systemen nach Gunther91

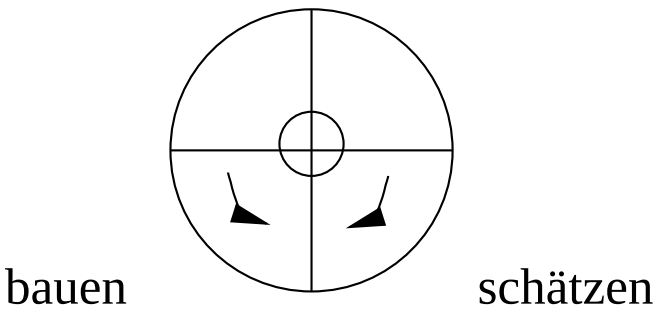


- Messung am existierenden Objekt
- Realisierung eines (Simulations-)Modells
- Entscheidung über Designparameter
- Bau des Systems (und Vergleich mit Modellergebnissen)
- Verbesserung und Tuning des Systems
- Markteinführung

Entwicklung der letzten Jahre führte zu

- immer kürzeren Innovationszyklen
- einem enormen Druck Software und Hardware schnell auf den Markt zu bringen
- Entwicklungszeiten (um fast) jeden Preis verkürzen

Folgen für den Entwurfszyklus



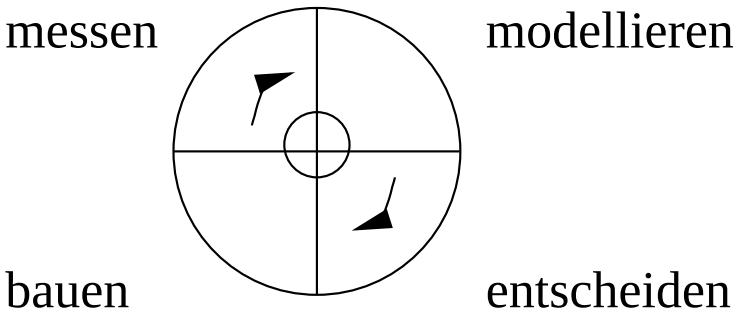
Leistungsverbesserung und Fehlerkorrektur
nach Markteinführung

Leistungsanalyse wird als zu aufwändig und
teuer angesehen

Aber ad hoc Vorgehen führt zu

- frustrierten Kunden
- teuren Fehlern
- fehlenden Planungsdaten

Deshalb wird benötigt:



Unterschied zur vorherigen Folie?

- der Geschwindigkeit angepasste Modellierungstechniken
- einfache, schnell zu erstellende und
schnell zu analysierende Modelle
- dies bedeutet i.a. keine Simulationsmodelle
(zumindest nicht für erste Analysen)

$\Rightarrow$ *performance by design*

1.4 Ein Referenzmodell zur Kapazitätsplanung

Die sind meisten heutigen verteilten Anwendungen sind
Client-Server Anwendungen

Anwendung wird in zwei Prozesstypen unterteilt

- in Clients, die Anfragen stellen
- und Server, die Anfragen beantworten

Prozesse laufen meist auf unterschiedlichen Rechnern

Vorlesung handelt primär von der Analyse solcher Systeme

⇒ Basisarchitektur und zugehörige Protokolle
müssen eingeführt werden, dies geschieht

- an Hand von Beispielen
- unter dem Gesichtspunkt der quantitativen Analyse

Basis für die quantitative Analyse

- Warteschlangennetze, die mit analytischen Methoden
analysierbar sind

Zur Ermittlung der Eingabedaten werden

- Messtechniken und
- Auswertungstechniken der Messungen benötigt

Schließlich erfordert die Planung neuer Installationen

- die Vorhersage zukünftiger Lastszenarien und
- deren Auswirkungen auf das Systemverhalten

⇒ Kapazitätsplanung

Was ist Kapazitätsplanung?

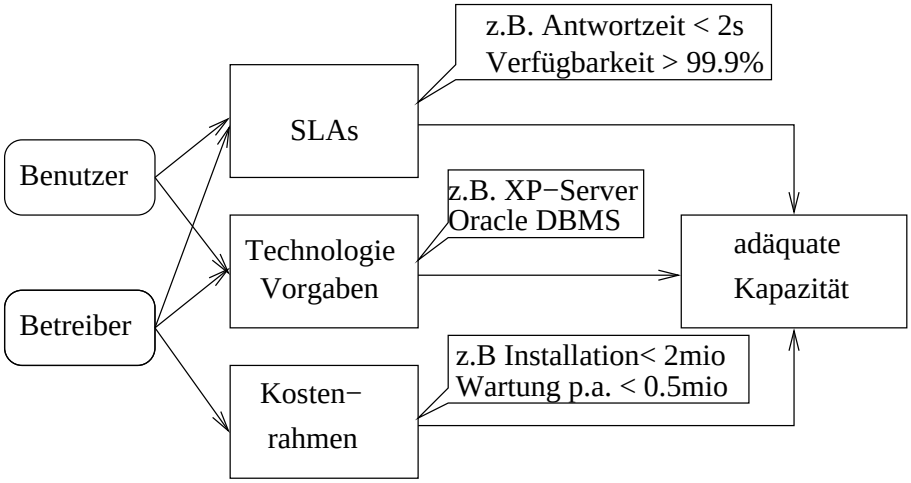
Definition (nach Menasce, Almeida)

Kapazitätsplanung ist der Prozess der Prognose, wann eine zukünftige Last ein System saturiert, und der Bestimmung eines kosten-effektiven Weges, diese Situation möglichst lange hinauszuzögern.

Bemerkungen:

- Ein System ist saturiert, wenn es die Leistungsgrenze erreicht, d.h. die vereinbarte Dienstgüte wird nicht mehr eingehalten.
- Die Angabe, wann ein System saturiert erfordert die Vorhersage zukünftiger Lasten, die existierende und neue Anwendungen entstehen können.

Definition der adäquaten Kapazität eines Systems:



Kapazitätsplanung als Prozess

Vorhersage als zentraler Aspekt der Planung
Gefragt ist ein systematischer, kontinuierlicher Prozess der Kapazitätsplanung, also nicht erst planen, wenn Probleme aufgetreten sind oder größere Änderungen anstehen damit

- finanzielle Verluste vermieden werden
- Kundenzufriedenheit erreicht wird
- Zeit für die Lösung von Leistungsproblemen bleibt, denn oft reichen einfache Veränderungen aus, um die Systemleistung zu erhöhen (Verlagern von Servern in LAN-Segmenten, Restrukturierung von DBs, ...)

Ziel ist es, IT-Produktionskapital optimal zu nutzen
Kapazitätsplanung leistet dabei
Entscheidungsunterstützung, damit

- das Budget wirtschaftlich eingesetzt wird
- aktuelle und zukünftige Leistungsanforderungen im Rahmen von Service Level Agreements erfüllt werden
- Zuverlässigkeit und Verfügbarkeit gegeben sind
- neue Techniken integrierbar sind

Konkrete Schritte

- Prognose der zukünftigen Last
- Prognose der notwendigen Systemleistung für die zukünftige Last
- Identifikation potentieller Engpässe
- Planung einer kostengünstigen Infrastruktur zur Beseitigung der Engpässe

Auch hier wieder zentraler Aspekt der Prognose

Leistungsbewertung allein untersucht die Leistung eines gegebenen Systems unter einer gegebenen Last

Randbedingungen der eingesetzten Methodik

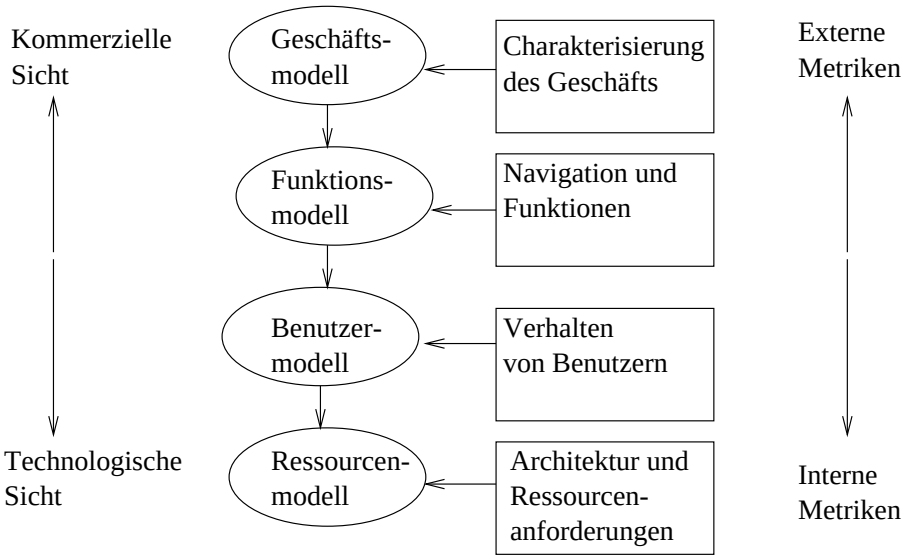
- Einsetzbarkeit für heterogene Systeme
- Skalierbarkeit für große Systeme
- Durchführbarkeit mit vertretbarem Aufwand

Typisches Vorgehen bei
der Leistungsanalyse und Kapazitätsplanung:

Analyse eines Systems bzgl. seiner angebotenen Funktionen auf Basis eines Modells der vorhandenen Ressourcen und deren Belastung

⇒ unterschiedliche Sichten

⇒ Transformationen sind notwendig



Die beiden oberen Ebenen betrachten primär
wirtschaftliche Aspekte

Die beiden unteren Ebenen betrachten das quantifizierte
Kundenverhalten und daraus resultierende
Ressourcenbelastungen

Metriken bewerten das Verhalten
auf den verschiedenen Ebenen

Geschäftsmodell:

- Beschreibung der Geschäftsziele
- Darstellung der Geschäftsabläufe
  inkl. der Quantifizierung von Größen
- Festlegung der Geschäftsteilnehmer und ihrer Rollen
- Beschreibung der Architektur des Produkt-, Service- und Informationsflusses

Verschiedene Modelle je nach Geschäftsbereich

Beispiel elektronisches Auktionshaus:

Ziel

▷ Durchführung von Auktionen unterschiedlichster Güter

Geschäftsablauf:

▷ Klassifizierung der gehandelten Güter nach Kategorien
▷ keine Kosten für Kunden aber Gebühren für Verkäufer
▷ keine Einschränkung bei den Geschäftszeiten

Teilnehmer:

▷ Käufer, Verkäufer und Auktionator

Käufer:

▷ melden sich an

▷ wählen Kategorie des gesuchten Gutes

▷ blättern im Katalog

▷ geben Gebote ab

▷ bezahlen gekaufte Güter

Verkäufer

▷ bieten an Gut an

▷ bezahlen Gebühren

▷ wählen u.U Kategorie aus

▷ legen Mindestgebote fest

▷ übergeben das Gut

Auktionator

▷ wacht über Regeln

▷ aktualisiert Kataloge

▷ erteilt Käufer den Zuschlag

▷ regeln Austausch von Gut und Geld

Quantitative Informationen:

▷ Zahl der Auktionen pro Tag

▷ mittlere Zahl der Käufer

▷ Anzahl Kategorien

Funktionsmodell:

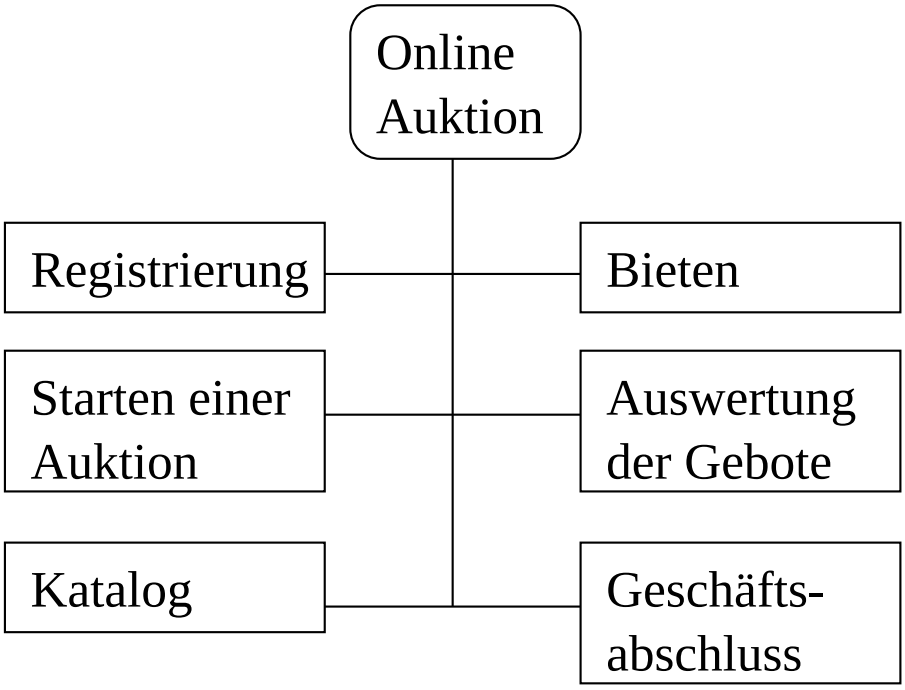
Aufteilung der Geschäftsprozesse in die verwendeten Funktionen und deren Interaktion

Zuordnung Funktionen zu Teilnehmern

Generierung durch $top\text{-}down$ Analyse des Geschäftsmodells

Beschreibung durch Flussdiagramme, Aktivitätsdiagramme, Entity-Relationship Modelle etc.

Beispiel Funktionsmodell des Auktionshauses



Basis für die Definition des "typischen" Teilnehmerverhaltens

Benutzermodell:

- Benutzer interagieren mit dem Anbieter durch Sequenzen von Funktionsaufrufen
- Benutzermodell beschreibt das (typische) Verhalten von Benutzern
- (probabilistische) Beschreibung der Zugriffe auf verschiedene Funktionen und der Zugriffsdauern
- Definition verschiedener Benutzerklassen

Ressourcenmodell:

- Ressourcen werden durch Funktionen belastet
- Funktionsaufruf eines Benutzers bedingt Belastung der Ressourcen
- Ressourcenmodell berechnet aus der Zahl der Benutzer und deren Verhalten die Belastung der Ressourcen ($\Rightarrow$ Lastmodell)
- Ressourcenmodell enthält quantitative Beschreibung der vorhandenen Ressourcen ($\Rightarrow$ Leistungsmodell)
- Analyse von Leistungs- + Lastmodell liefert Leistungsmaße des Systems!

Leistungsmaße werden auf Basis
des Ressourcenmodells ermittelt
benötigt werden aber

- Resultate auf Benutzerebene
  (z.B. nicht Durchsatz an der CPU sondern Durchsatz von Kundenanfragen, nicht Antwortzeit der Festplatte sondern Zugriffszeit auf eine Katalogseite)

- Resultate bzgl. der Geschäftsprozesse
  (oftmals quantifiziert bzgl. Kosten und Erlösen)

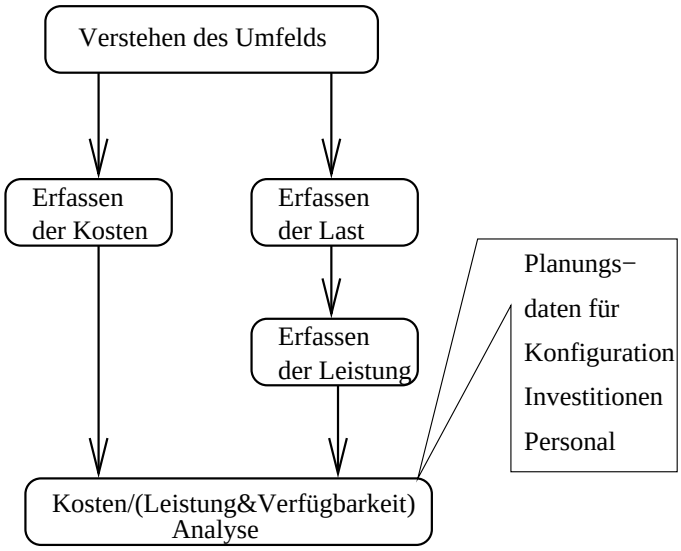
⇒ Transformation der Leistungsmaße vom
Ressourcenmodell zum Benutzer- und/oder
Geschäftsmodell

Typischerweise werden Dienstqualitäten vorgegeben
(z.B. Verfügbarkeit, Antwortzeit)
und sind Restriktionen zu berücksichtigen
(z.B. Kostenrahmen, technologische Vorgaben)

⇒ Kapazitätsplanung muss sicherstellen,
dass Dienstqualität unter den gegebenen Restriktionen
auch in Zukunft garantiert wird

Methodisches Vorgehen bei der Leistungsanalyse und Kapazitätsplanung

1. Schritt Verständnis des Umfeldes



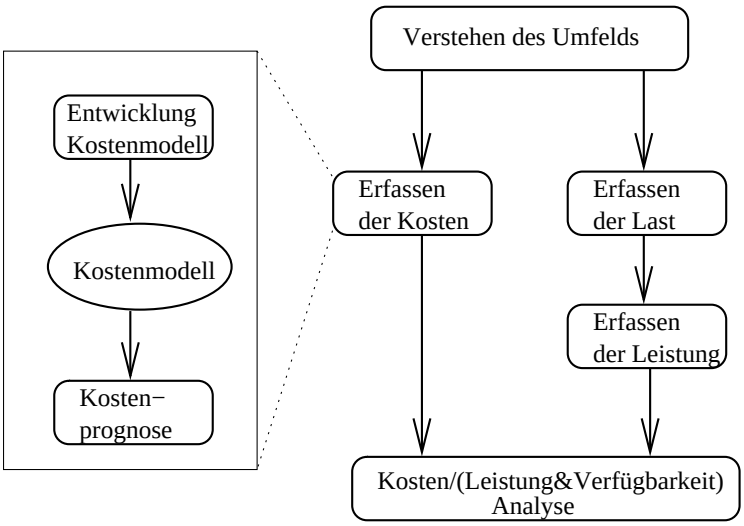
Verstehen des Umfeldes:

- Hardware
- Software
- Netzwerke und Protokolle
- Management Strukturen
- Service-Level Agreements
- Nutzungszeiten

Informationen mit Hilfe von

- Meetings, Umfragen, Interviews
- Planungen
- Aufzeichnungen

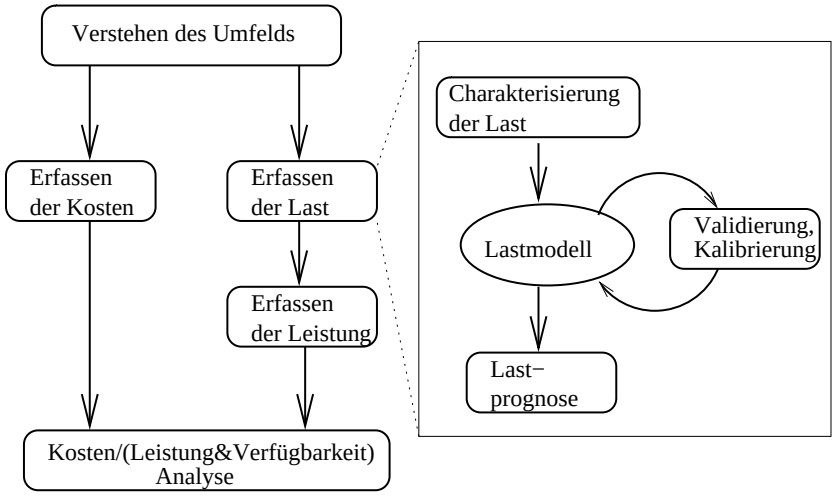
Methodik, detaillierter



Quantifizierung der Kosten für die Bereitstellung und den Betrieb von Ressourcen

- Hardwarekosten
  - Anschaffung, Betrieb, Wartung
- Softwarekosten
  - Anschaffung, Lizensierung, Updates
- Personal
- Verwaltung

hier nicht im Detail untersucht, da Gegenstand klassischer betriebswirtschaftlicher Betrachtungen



Lastmodell umfasst Umfang und Intensität der Ressourcenbelastung für jede Komponente der Gesamtlast über einen repräsentativen Zeitraum

Grundfragen

- Welche Aufgaben/Tasks gibt es?
- Wieviel Bearbeitungszeit erfordert die Bearbeitung einer Anfrage eines Typs an einer Ressource?
- Wie oft tritt die Anfrage eines Typs auf?

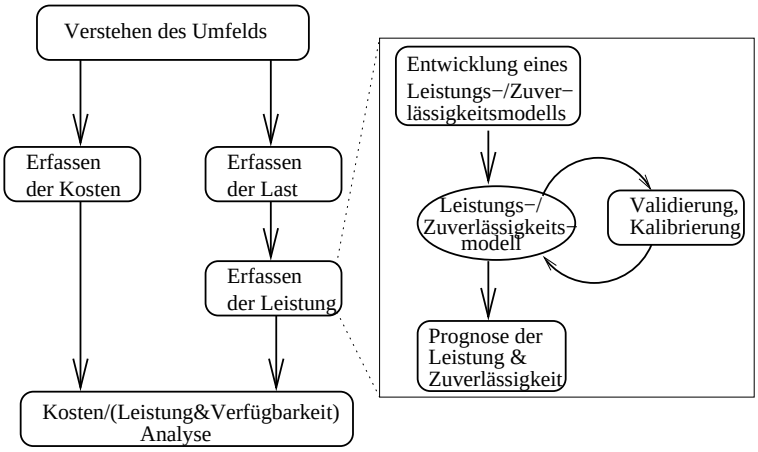
Datenbeschaffung

- Auswertung von Messdaten (z.B. HTTP-Logs)
- Aggregierung der Anfragedaten

Lastmodell beschrieben durch

- Benutzerverhaltensgraph (BVG)
- Client-Server-Interaktionsdiagramm (CSID)

basierend im wesentlichen auf Mittelwerten



Leistung/Zuverlässigkeit als Funktion der Last

- Maschinenmodell
  Hardware (CPU, Platten, Netze), Software (Threads)

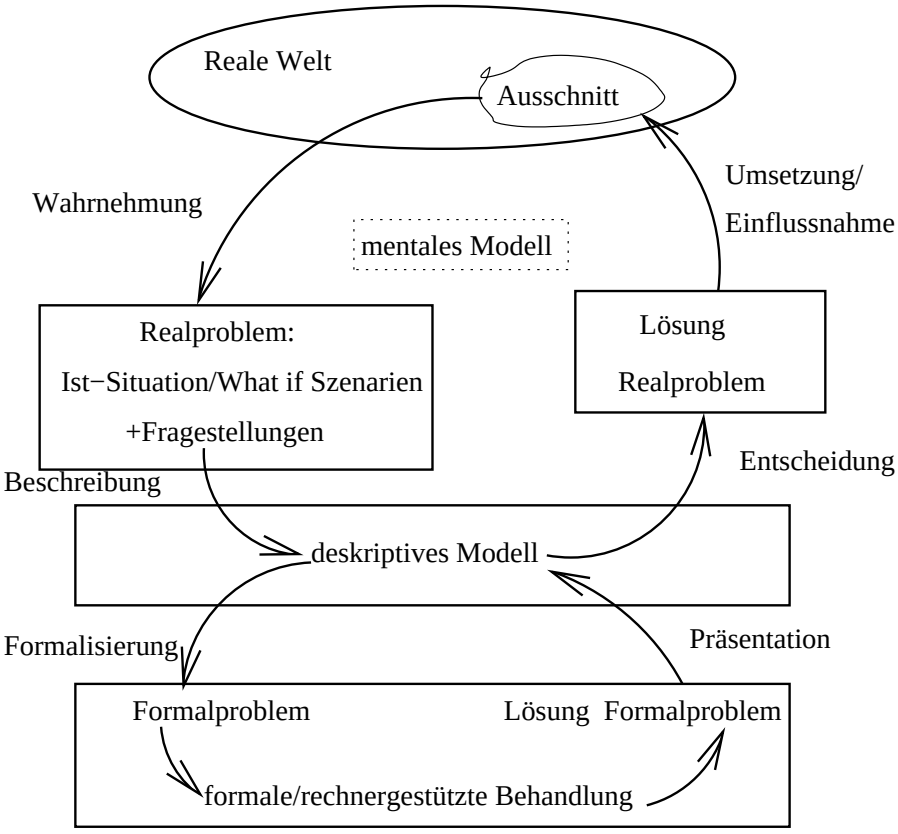
Ziel der Leistungs-/Zuverlässigkeitsanalyse

▷ Bestimmung von Verzögerungen auf Grund von Wartesituationen an geteilten Ressourcen

Methodik

- experimentell
  reale Last/reales System oder synth. Last/reales System
  statistische Auswertung
- simulativ
  reale Last/Maschinenmodell oder synth. Last/Maschinenmod.
  statistische Auswertung
- analytisch (hier primär verwendet)
  Lastmodell/Maschinenmodell
  Berechnung der Ergebnisgrößen

Modellierung ⇒ Modellgestützte Lösung von Problemen



Grundregeln

- Ein Modell ist einfacher als das Original
- Modell dient zur Beantwortung einer Frage

⇒ Komplexitätsreduktion & Zielorientierung

⇒ Nur die Aspekte des Originalsystems, die zur Beantwortung der Fragestellung relevant sind, müssen bei der Modellierung berücksichtigt werden

Sichtweise der Kapazitätsplanung

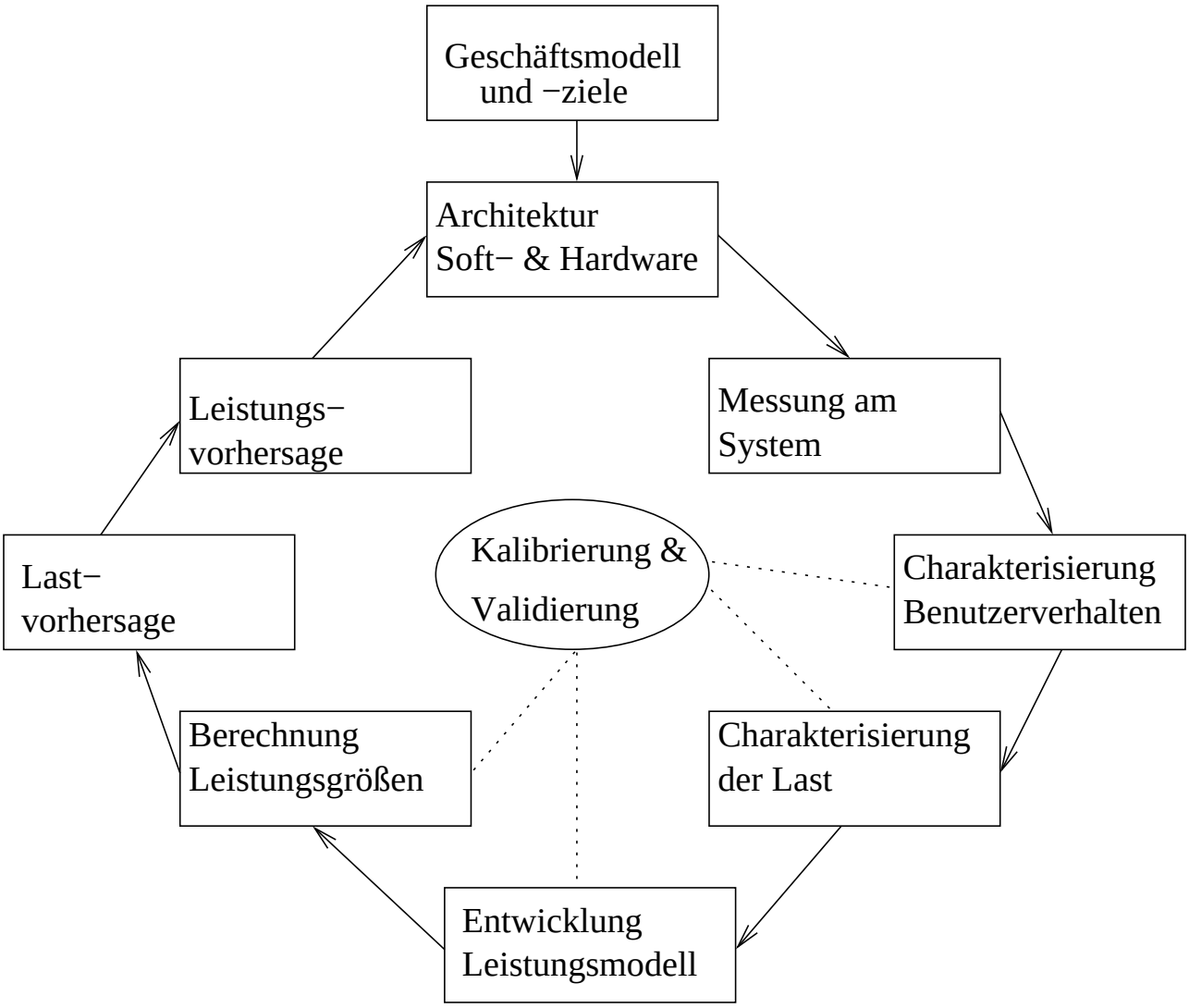
Last: Anwender nutzen Software und verursachen
Berechnungs- und Kommunikationsvorgänge

Arbeit wird durch IT-Infrastruktur erbracht

Leistung lässt sich beobachten, messen, berechnen

Eingangsgrößen	**Resultate**
Lastbeschreibung, Lastentwicklung	Leistungskenngrößen
• Intensität • Umfang • Typ	• Typ, Mittelwerte Antwortzeiten Auslastungen, Durchsätze
Systemparameter	Saturierungspunkte
• CPUs, Platten • Netzverbindungen, Bandbreite max. Anzahl Verbindungen	• Bei welcher Last wird die erforderliche Dienstgüte nicht mehr erreicht?
Gefortderte Dienstgüte	Kosteneffektive Alternativen
• Antwortzeit $< 8s$ • Durchsatz $> 12tps$	Verhältnis Leistung/Kosten in Szenarien (z.B. tps per \$)

Übersicht über das gesamte Vorgehen



einzelne Schritte werden in der Vorlesung erläutert