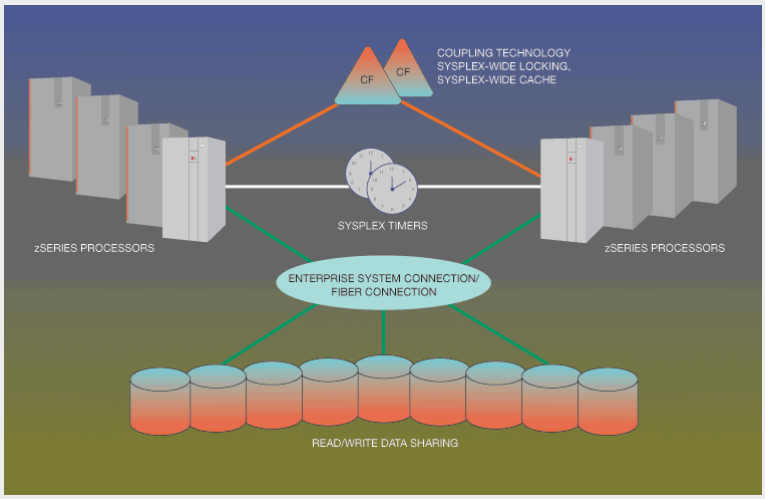
Lastverteilung in z/OS Workload Manager

Wintersemester 2009/10

Martin Bogdan
bogdan@informatik.uni-leipzig.de

Universität Leipzig
Technische Informatik

Lastverteilung



z/OS SYSPLEX

Lastverteilung: Problemstellung

- **Problem**
  - **Viele Anfragen auf ein System**
    - **Beispiel: WWW-Cluster**
      - **Kann aus mehreren tausend Systemen bestehen**
        - **Relativ einfach: Einheitliche Anwendung**
  - **ABER: Ressourcen sind begrenzt**
    - **Nicht alle Anfragen können sofort bearbeitet werden**
- **Fragestellung:**
  - **Wie können die vorhanden Ressourcen so verteilt werden, daß sie optimal genutzt werden können?**
    - **Kann man auch steuernd eingreifen?**

Lastverteilung

- **Klassische Verfahren**
  - **Run-To-Completion**
  - **Zeitscheibensteuerung, z.B. exponentiell**
  - **Anzahl aktiver Prozesse**
  - **Multithreading**
    - **Anzahl von Subsystem Instanzen**
  - **Zuordnung von Anwendungen auf physikalische Server**
  - **Zuordnung der E/A Kanäle**
  - **Prioritäten**
    - **etc.**

Lastverteilung

- **Komplexität**
  - Mit wachsender Größe des zu steuernden Systems wächst die Komplexität der Steuerung
    - Belastungsschwankung
    - Zuordnung Anwendung → reale CPU
    - Unterschiedliche E/A-Anforderungen
- **Lösungsansätze**
  - Hardware ist billig → Unterschiedliche Anwendung auf unabhängige Server
    - Schlechte Auslastung
    - Viele Server
    - Komplexe Kommunikation
    - Hoher Administrationsaufwand
    - Heterogene Serverlandschaft

ODER:

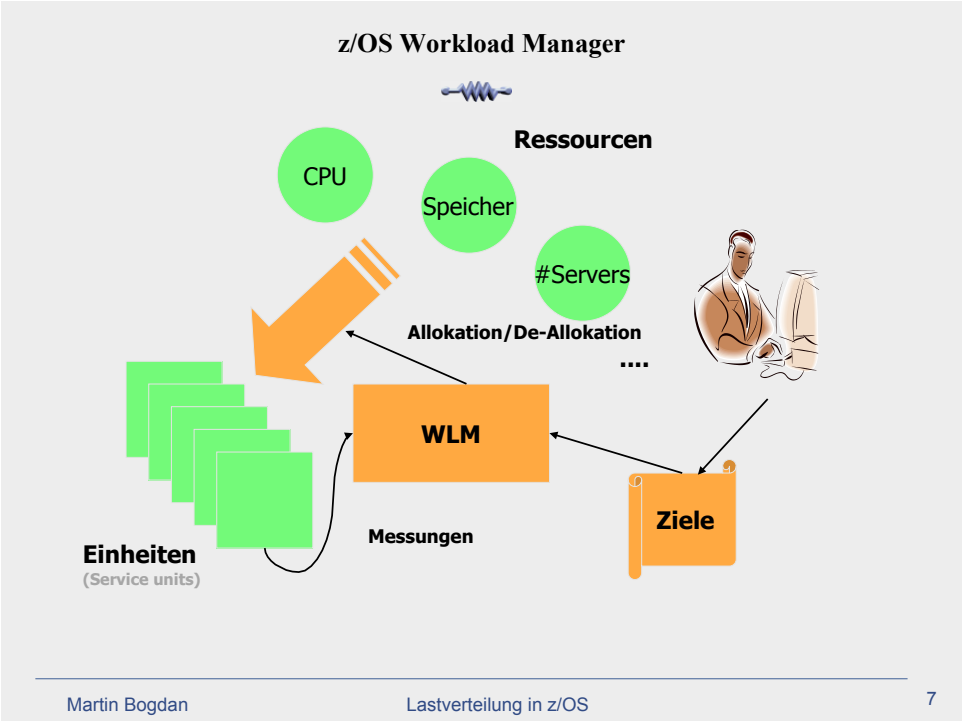
Workload Manager

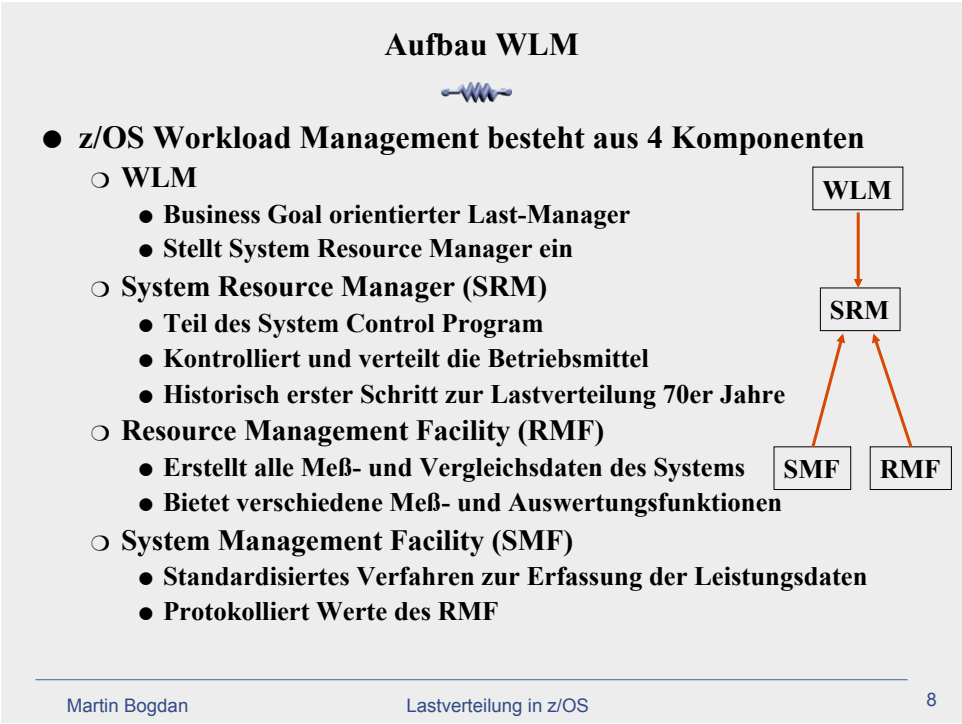
Martin Bogdan Lastverteilung in z/OS 5

Lastverteilung

- **z/OS Workload Manager (WLM)**
  - Verteilung der Ressourcen
    - Automatisch
    - Dynamisch
  - Grundlage: „Business Goals“
    - WLM setzt Vorgaben um
  - Vorteile
    - Reduzierter Aufwand für Administration
    - Reduzierung Detailwissen möglich
      - ACHTUNG: Zu wenig Detailwissen kann zur Verschlechterung führen, da eine Vielzahl von Einstellungsmöglichkeiten gegeben wird!

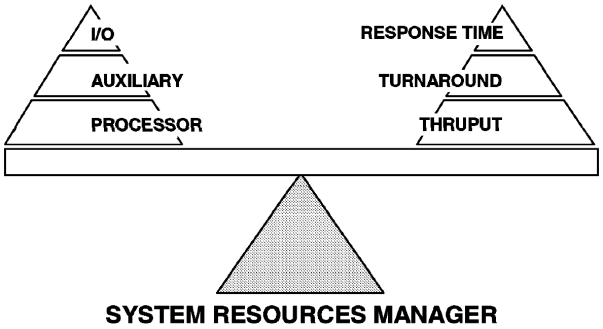
Martin Bogdan Lastverteilung in z/OS 6


z/OS Workload Manager
Ressourcen
CPU
Speicher
#Servers
Allokation/De-Allokation
....
WLM
Ziele
Messungen
Einheiten
(Service units)
Martin Bogdan
Lastverteilung in z/OS
7


Aufbau WLM
z/OS Workload Management besteht aus 4 Komponenten
WLM
Business Goal orientierter Last-Manager
Stellt System Resource Manager ein
System Resource Manager (SRM)
Teil des System Control Program
Kontrolliert und verteilt die Betriebsmittel
Historisch erster Schritt zur Lastverteilung 70er Jahre
Resource Management Facility (RMF)
Erstellt alle Meß- und Vergleichsdaten des Systems
Bietet verschiedene Meß- und Auswertungsfunktionen
System Management Facility (SMF)
Standardisiertes Verfahren zur Erfassung der Leistungsdaten
Protokolliert Werte des RMF
WLM
SRM
SMF
RMF
Martin Bogdan
Lastverteilung in z/OS
8

System Resource Manager SRM

- SRM
  - Hauptaufgabe: Verteilung der Betriebsmittel
    - CPU Auslastung
    - Hauptspeicher Nutzung
    - E/A Belastung



Martin Bogdan Lastverteilung in z/OS 9

System Resource Manager SRM

- SRM-Zeit
  - Im MVS ist Zeit ≠ konventionelle Zeit!
    - SRM-Zeit ist Basis für Meß- und Entscheidungsintervalle
    - Idee: Vergleichbarkeit des Aufwandes auf unterschiedlichen Rechnern
  - „SRM-Sekunde ist diejenige Zeiteinheit, die bestimmt, wann Algorithmen im System ablaufen“ (Teuffel/Vaupel)
    - Auf langsameren Rechner läuft eine SRM-Sekunde entsprechend langsamer ab
    - Systemaufwand bleibt aber konstant, egal ob schneller oder langsamer Rechner!

Martin Bogdan Lastverteilung in z/OS 10

System Resource Manager SRM

- **Weitere SRM-Zeitabstände**
  - **SRM-Meßintervall**
    - **Basis für alle SRM-Algorithmen**
    - **Vielfaches der SRM-Sekunde**
  - **Vielfaches des Meßintervalls („Meßdatenzeit“)**
    - **Über diesen Zeitraum werden die Meßdaten summiert**
    - **Entspricht etwa 2 Sekunden echter Zeit**
      - **Variiert in Abhängigkeit von der CPU-Geschwindigkeit**
  - **Transaktionsantwortzeit**
    - **Gemessene Zeit von Beginn bis Ende einer Transaktion**
    - **Messung durch MVS-Komponenten**
  - **„Adreßraumzeit“**
    - **Zeit, wie lange Adreßräume Speicherraum belegen, ausgelagert sind oder keine Anforderungen gestellt werden**

Martin Bogdan Lastverteilung in z/OS 11

System Resource Manager SRM

- **Service Units**
  - **Transaktionen im MVS nutzen Betriebsmittel des Systems**
    - **CPU-Zyklen**
    - **Pages im Hauptspeicher**
    - **E/A-Operationen**
  - **Systeme unterscheiden sich stark durch vorhandene Betriebsmittel**
  - **Standardisiertes Maß für den Betriebsmittelverbrauch sind die Service Units**
    - **Unterschiedlich definiert für CPU, Speicher und E/A**

Martin Bogdan Lastverteilung in z/OS 12

System Resource Manager SRM

- Service Units CPU
  - CPU-Verbrauch wird auf zwei Arten gemessen
    - Anwendungen, die in Adreßräumen unter Task Control Blocks (TCB) laufen
    - Inanspruchnahme des Betriebssystems → Laufen unter Service Request Blocks (SRB) in einem Adreßraum ab
  - CPU-Verbrauch wird in TCB- und SRB-Zeit aufgeschlüsselt
    - Erlaubt gesonderte Analyse
    - Verbessert Einstellungsmöglichkeiten der Ressourcen des Systems
    - Anpassung an unterschiedliche CPU-Geschwindigkeiten über Multiplikation einer Konstante, die sich in INITTUNE befindet
- Service Units Speicher
  - Anzahl der Speicherseiten einer Anwendung, solange Anwendung CPU-Zeit nutzt
    - Speicherbelegung: Main Storage Occupancy (MSO)
    - Speicher Service Units: Main Storage Units (MSU)

Martin Bogdan Lastverteilung in z/OS 13

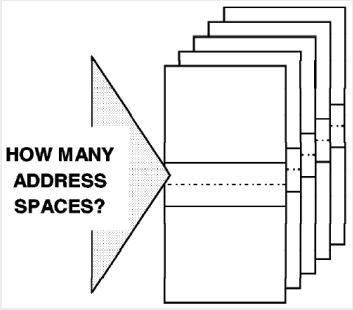
System Resource Manager SRM

- Service Unit E/A
  - Zwei Verfahren von Anwender wählbar
    - Anzahl der abgeschickten Execute Channel Programme (EXCP)
      - Nachteil: E/A mit großen Datenmengen werden nicht korrekt erfaßt
    - Alternativ E/A-Transfer-Zeit
      - Nachteil: Virtuelle E/A wird nicht erfaßt
    - Heute wird meistens die erste Alternative gewählt
- Verbrauchserfassung
  - SRM-Algorithmen hängen von der Service-Rate ab
    - Service-Rate: Summe der Services pro Sekunde
    - Gewichtung der Service-Komponenten untereinander
      - Definition Coefficients in SYS1.PARMLIB
      - Anpassung an Benutzerbedürfnisse und unterschiedliche Installationen

Martin Bogdan Lastverteilung in z/OS 14

System Resource Manager SRM

- **SRM-Transaktionen**
  - **Transaktion aus Sicht des Betriebssystems: Bearbeitungsvorgang mit definiertem Anfang und Ende**
  - **In z/OS ist das zugehörige Basiskonstrukt der Adreßraum**
    - **Adreßraum wird beim Start erzeugt**
    - **Adreßraum wird beim Ende gelöscht**



Martin Bogdan Lastverteilung in z/OS 15

System Resource Manager SRM

- **SRM-Transaktionen**
  - **Problem: Betriebssystem erkennt Transaktion nur durch Auf- bzw. Abbau des Adreßraumes!**
    - **Für reine Batch-Verarbeitung in Ordnung**
    - **Forderung bei TSO: Zeitraum, bei dem der Anwender kein Anforderungen stellt, zu erkennen!**
      - **Beginnt mit Eingabe, endet mit Antwort auf Eingabe**
  - **SRM erkennt Adreßräume für**
    - **JES (Batch-Jobs)**
    - **TSO-Benutzer**
    - **Rest: Standard-Tasks**

Martin Bogdan Lastverteilung in z/OS 16

System Resource Manager SRM

- **Betriebsmittelsteuerung**
  - **SRM verwaltet CPU, Speicher, Platten des E/A-Subsystems**
  - **SRM kann auch Adreßräume verwalten**
  - **Einteilung der Transaktionen durch SysAdmin in**
    - **STC (Standard Tasks): alle Adreßräume des Systems**
    - **JES2, JES3 für Job Entry Subsysteme**
    - **TSO für Time Sharing Option**
    - **APPC für Advanced Program-to-Program Communication**
    - **OMVS für Unix System Services**
    - **CICS für Costumer Information Control System**
  - **Basierend auf der Einteilung können Performance-Gruppen und darausfolgende Domainen definiert werden**

Martin Bogdan Lastverteilung in z/OS 17

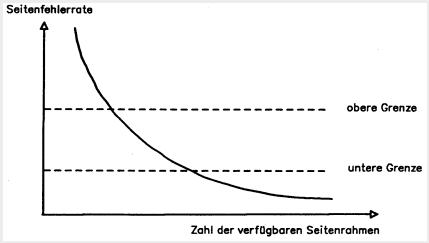
System Resource Manager SRM

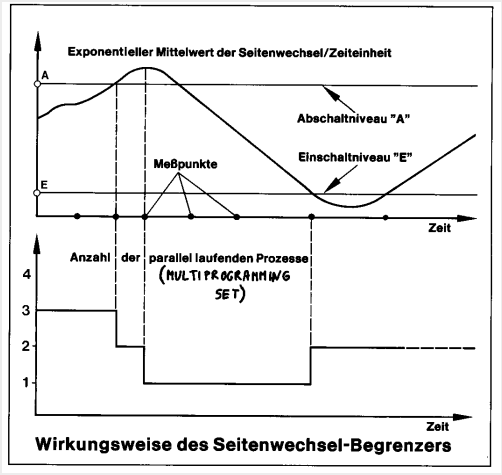
- **Steuerung des Betriebsmittelverbrauch**
  - **Domaine legt fest, wieviel Adreßräume gleichzeitig im System sein dürfen**
    - **Zahl der gleichzeitig im System vertretenen Adreßräume: Multi Programing Level (MPL)**
    - **MPL Teil des Swapping**
  - **Swapping**
    - **Auslagern von Adreßräumen**
      - **Bereitstellung von genügendem Hauptspeicher**
      - **Vermeidung von Page Faults**
      - **Ermöglichung von Swap-Ins**
  - **Paging**
    - **Lagert die ältesten, nicht mehr genutzten Speicherseiten aus**

Martin Bogdan Lastverteilung in z/OS 18

System Resource Manager SRM

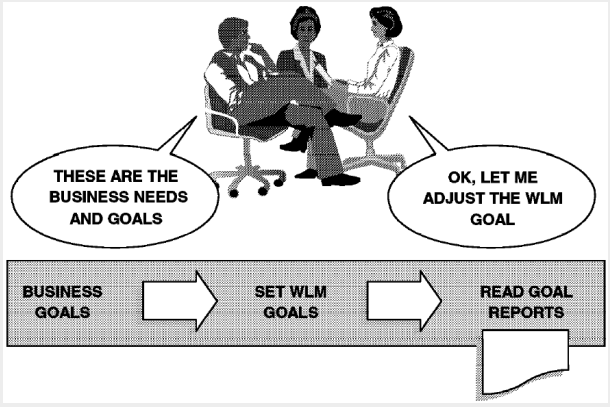
- Swapping/Paging





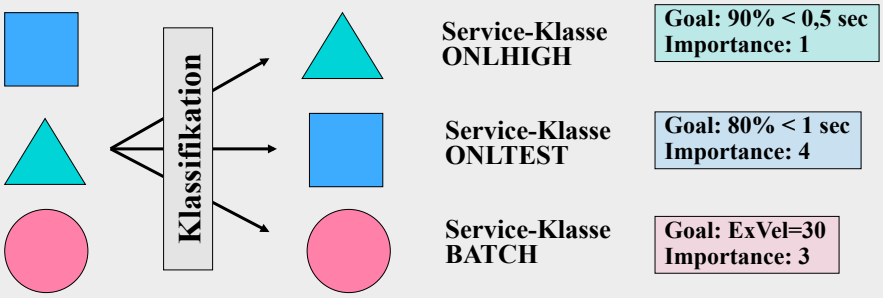
Workload Manager WLM

- Workload Manager (WLM)
  - „Übersetzt“ Business Goals für SRM
  - WLM steuert Parameter des SRM auf Grund der Business Goals



Workload Manager WLM

- Service-Klassen und Klassifizierung
  - Service-Klasse ist Grundkonstrukt im Goal Mode
    - Klassifizierung der Workloads mit unterschiedlichen Merkmalen
      - Art der Arbeit (Anwendung/Subsystem)
      - Laufzeit-Charakterisitken
      - Hauptkriterium: Wichtigkeit des Jobs!



Martin Bogdan Lastverteilung in z/OS 21

Workload Manager WLM

- Zielvorgaben für die Service-Klassen
  - Response Time
  - Execution Velocity
  - Discretionary
    - Business Importance (für jede Service-Klasse)
- Response Time Goals
  - Average Response Time Goal
    - Durchschnittliche Ausführungsdauer in Sekunden
      - Z.B.: 1,5 sec
  - Percentile Response Time Goal
    - Prozentsatz aller Transaktion, die innerhalb einer Zeit beendet werden
      - Z.B.: 90% < 0,5 sec

Martin Bogdan Lastverteilung in z/OS 22

Workload Manager WLM

- **Execution Velocity Goal**
  - WLM sammelt alle 250 msec Daten aus aktuellem Betriebszustand
  - Für jede Arbeitseinheit wird ermittelt, ob Betriebsmittel verwendet wird oder darauf gewartet wird
  - Daraus wird die Execution Velocity berechnet

$$ExecutionVeloctiy = 100 \bullet \frac{\text{Total Using}}{\text{Total Using + Total Delays}}$$

  - Zeiten, an denen keine Betriebsmittel genutzt und auch nicht auf sie gewartet wurde, werden nicht berücksichtigt

Martin Bogdan Lastverteilung in z/OS 23

Workload Manager WLM

- **Discretionary**
  - Keine Zielvorgabe existent
  - Erhält nur dann Betriebsmittel, wenn ausreichend vorhanden
  - Bei Engpässen sofortiger Entzug der Betriebsmittel
- **Business Importance**
  - Erhält jede Service-Klasse vom Benutzer zugeteilt
  - Gibt an, wie wichtig dieses Ziel im Verhältnis zu anderen ist
  - 1 (am wichtigsten), 2, 3, 4, 5 (am wenigsten wichtig)
  - Discretionary erhält Business Importance 6

Martin Bogdan Lastverteilung in z/OS 24

Workload Manager WLM

- **Service Definition**
  - **Konstrukt, in dem Benutzerdefinition für die Systemsteuerung beschrieben wird**
  - **Wird benötigt um ein System im Goal Mode zu betreiben**
  - **Definition mit WLM Administrative Application**
    - **Als TSO-Benutzer unter ISPF**

Workload Manager WLM

- **Service Definition beinhaltet:**
  - **Allgemeiner Teil**
  - **Service Policy**
    - **Umfaßt Service Klassen mit bestimmten Zieldefinitionen**
  - **Workload**
    - **Gruppierungsmechanismus für die Arbeit des Systems**
  - **Service-Klassen-Periode**
    - **Einheiten von Arbeiten mit gleichartigen Charakteristiken für die Business Importance und Ziele definiert sind**
  - **Classification Rules**
    - **Zuordnung von Arbeitseinheiten zu Service-Klassen**
  - **Resource Groups**
    - **Legt obere und untere Grenzen für Verbrauch von Service Units fest**
  - **Report Klassen**
    - **Unterstützen Berichtswesen durch weitere Klassifizierung der Arbeit neben Service-Klassen**
  - **Application Environments**
    - **Erlaubt dem WLM Arbeit von Anwendungen an WLM-gesteuerte Adreßräume weiterzuleiten**
  - **Scheduling Environments**
    - **Erlaubt Festlegung von Affinitäten zwischen Anwendungen, Systemen und Workloads**

Workload Manager WLM

- **Kontrolle der Zielvorgaben**
  - **Performace Index PI**

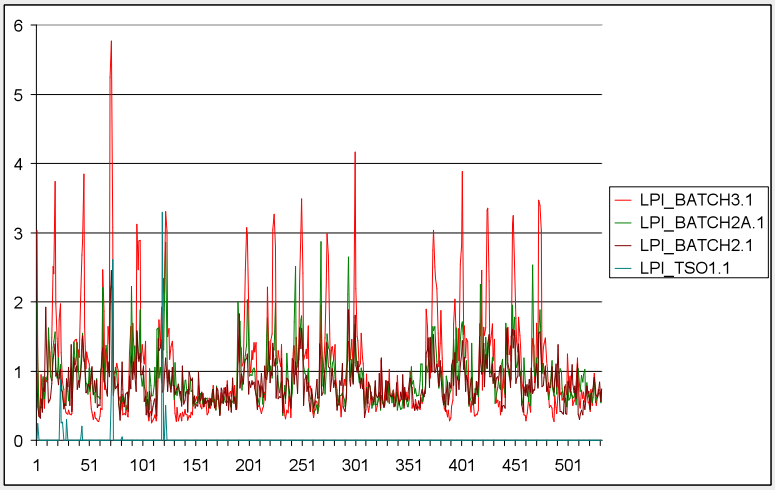
$$PI = \frac{\text{Aktuelle Response Time}}{\text{Response Time Goal}} \qquad PI = \frac{\text{ExecVelocity Goal}}{\text{Aktuell erreichte ExeVelocity}}$$

  - **PI > 1 : Zielvorgabe verfehlt**
  - **PI = 1 : Zielvorgabe exakt erfüllt**
  - **PI < 1 : Zielvorgabe übererfüllt**

Workload Manager WLM

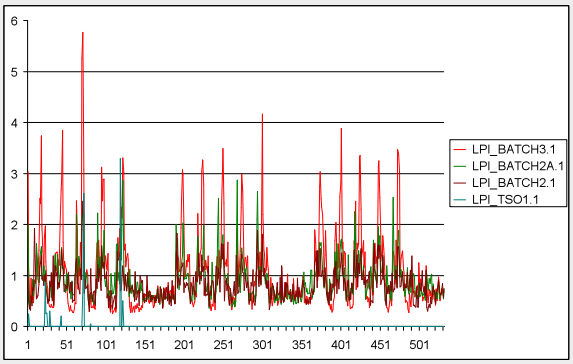
- **Policy Adjustment**
  - **Strategie zur Berechnung neue Vorgabe der Betriebsmittel**
  - **Ermittelt Empfänger**
    - **Abhängig von Business Importance**
    - **PI**
    - **Definition in Ressource Group (untere Grenze erfült?)**
  - **Feststellen des Engpasses**
    - **CPU Delays, Paging Delays, ...**
  - **Ermittlung Geber**
    - **Welche Service-Klasse benutzt benötigte Betriebsmittet?**
    - **Berechnung der Veränderung durch Übergabe der Betriebsmittel**
  - **Betriebsmittel werden alle 10 sec neu verteilt (wenn nötig)**

Gemessener Performance Index von 4 Service-Klassen



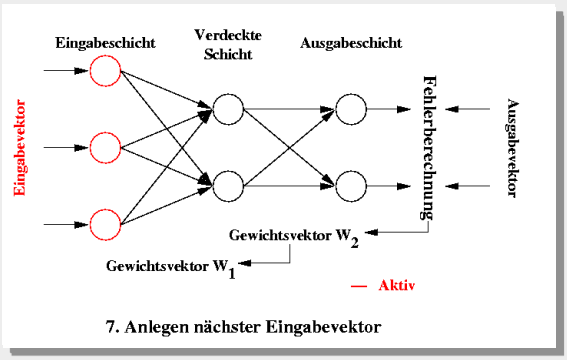
Aktuelle Fragestellung für WLM

- **Problem: Ist Workload spezifisch vorhersagbar?**
- **Ziel: Rechtzeitige Bereitstellung von Ressourcen**
- **Idee: Einsatz Künstlicher Neuronaler Netze zur Vorhersage**



Neuronale Netze und Algorithmen

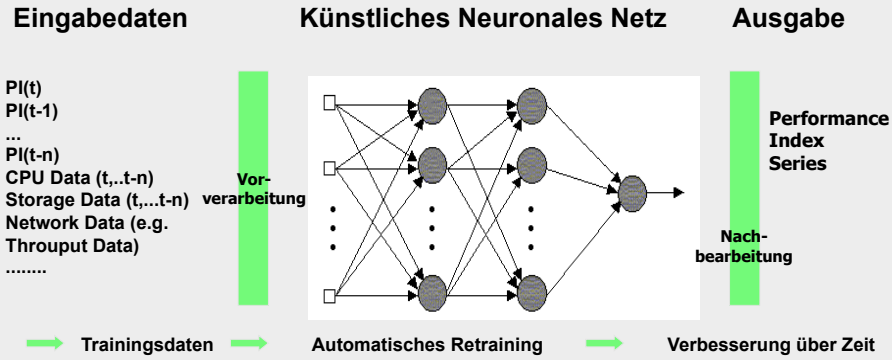
- Beispiel: Backpropagation

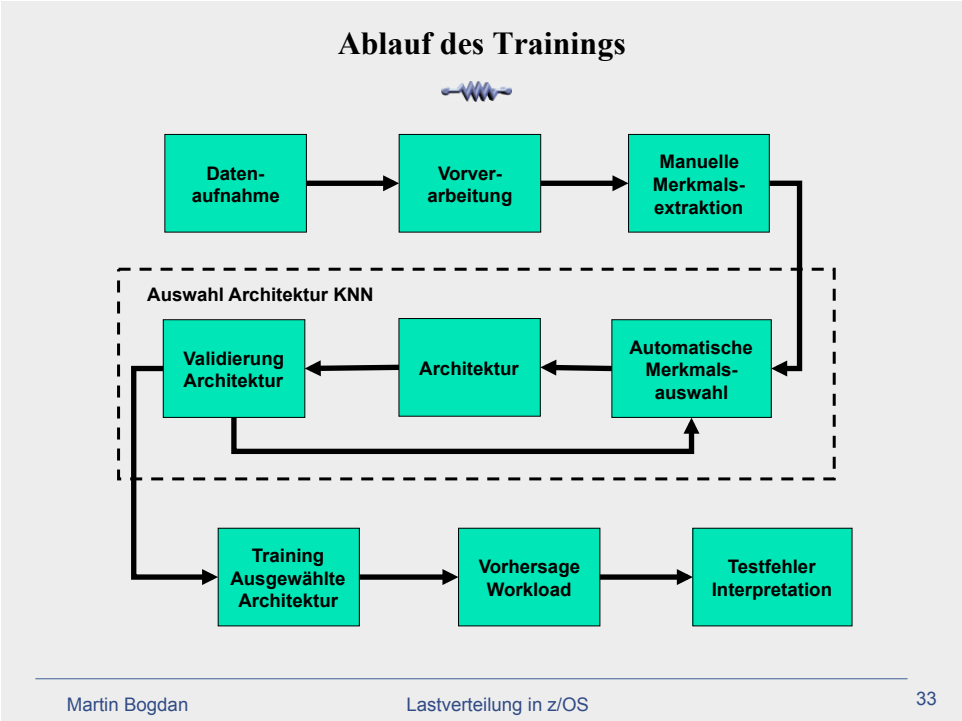


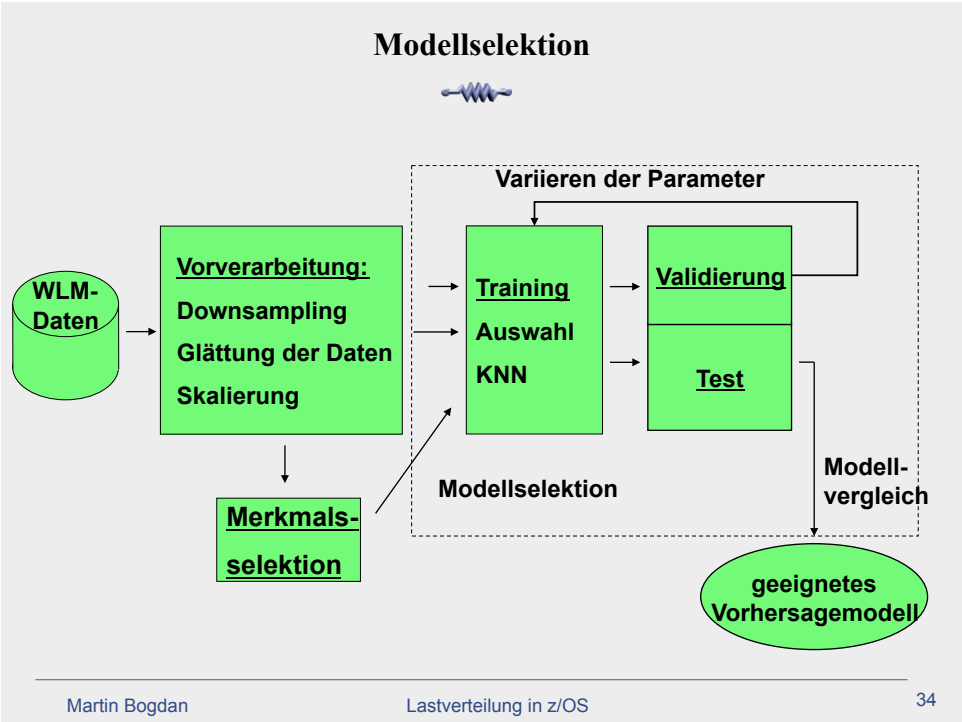
$$\Delta w_{ij} = \eta o_i \delta_{ij}^{back} \qquad \delta_{ij}^{back} = \frac{\partial o_j}{\partial \sum_i o_i w_{ij}} (t_j - o_j) \qquad \delta_{ij}^{back} = \frac{\partial o_j}{\partial \sum_i o_i w_{ij}} \sum_k \delta_k w_{jk}$$

Für o_j Ausgabeneuron — Für o_j verdecktes Neuron

KNN zur Vorhersage des Performance Index

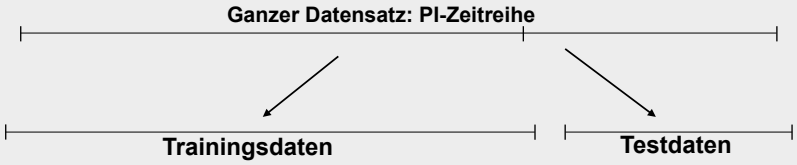



Ablauf des Trainings
Daten-aufnahme
Vorver-arbeitung
Manuelle Merkmals-extraktion
Auswahl Architektur KNN
Validierung Architektur
Architektur
Automatische Merkmals-auswahl
Training Ausgewählte Architektur
Vorhersage Workload
Testfehler Interpretation
Martin Bogdan
Lastverteilung in z/OS
33


Modellselektion
Variieren der Parameter
WLM-Daten
Vorverarbeitung:
Downsampling
Glättung der Daten
Skalierung
Training
Auswahl
KNN
Validierung
Test
Modell-vergleich
Modellselektion
Merkmals-selektion
geeignetes Vorhersagemodell
Martin Bogdan
Lastverteilung in z/OS
34

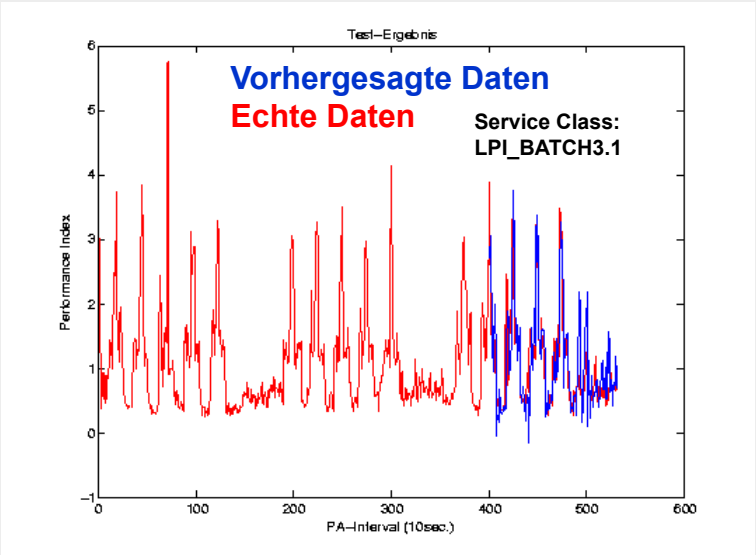
Training

- **Kurzzeit Datensatz:**
  - **1 Vorhersageschritt = 100s**
  - **50 Vorhersageschritte ≈ 1.4 h**
- **Langzeit Datensatz:**
  - **1 Vorhersageschritt = 500s**
  - **50 Vorhersageschritte ≈ 7h**
- **Training mit ¾ des Datensatzes**
- **Test mit ¼ des Datensatzes**



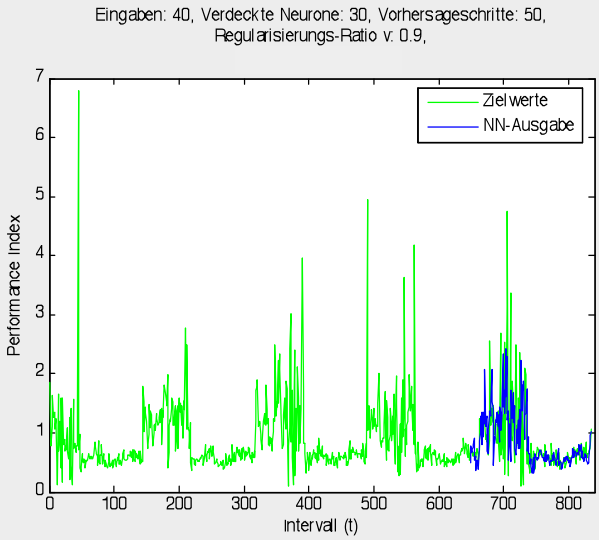
Martin Bogdan Lastverteilung in z/OS 35

Vorhergesagter Performance Index durch KNN



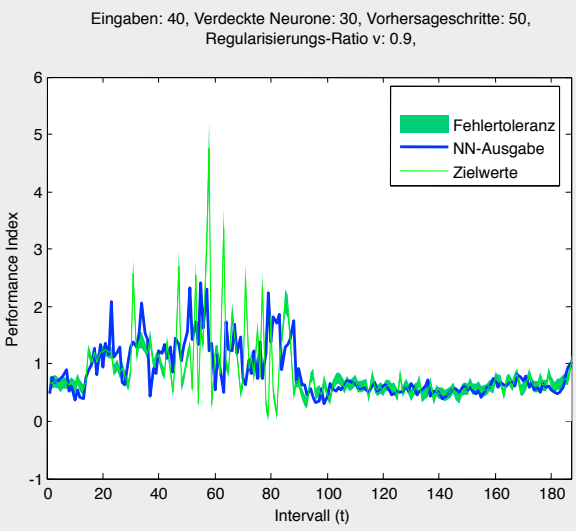
Martin Bogdan Lastverteilung in z/OS 36

Ergebnisse…

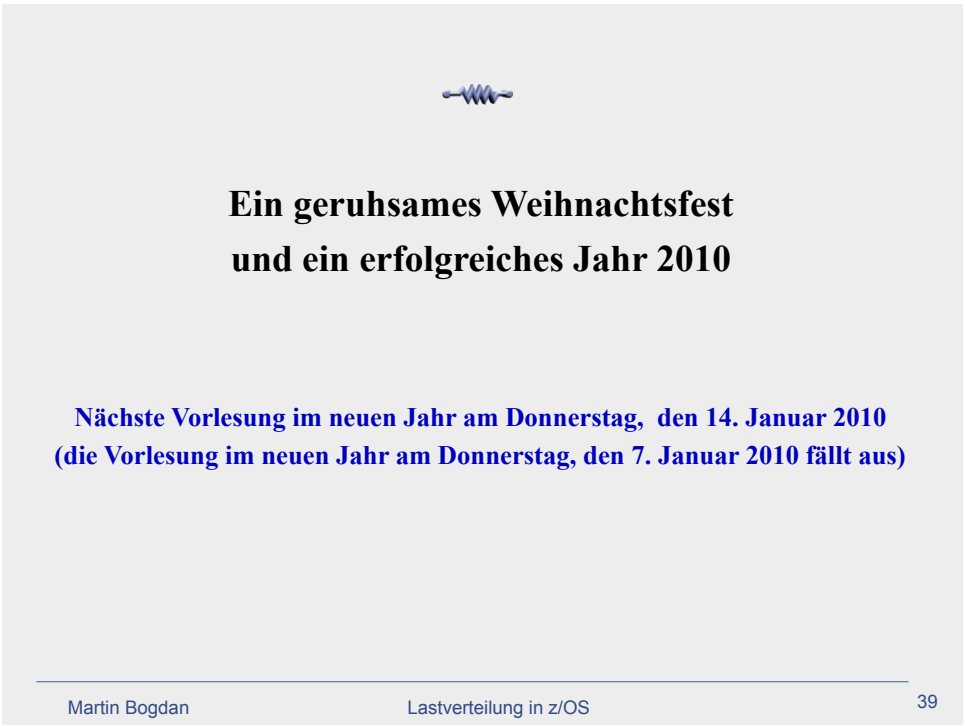


Martin Bogdan Lastverteilung in z/OS 37

… Ergebnisse



Martin Bogdan Lastverteilung in z/OS 38


Ein geruhsames Weihnachtsfest
und ein erfolgreiches Jahr 2010
Nächste Vorlesung im neuen Jahr am Donnerstag, den 14. Januar 2010
(die Vorlesung im neuen Jahr am Donnerstag, den 7. Januar 2010 fällt aus)
Martin Bogdan
Lastverteilung in z/OS
39