Verteilte Systeme

Oktober - November 2023

4. Vorlesung – 26.10.2023

Kurs: TINF21AI1

Dozent: Tobias Schmitt, M.Eng.

Kontakt: d228143@

student.dhbw-mannheim.de

Wiederholungsfragen

- Welche Arten der Kommunikation k\u00f6nnen Sie unterscheiden?
- .Was ist die Idee hinter einem RPC (Remote Procedure Call)?
- •Welche Schwierigkeiten finden bei der Kommunikation zwischen verschiedenen Rechnern (Diensten) in heterogenen Umgebungen?
- Nennen Sie Beispiele für flüchtige und persistente Kommunikation.

Wiederholungsfragen

- ·Was verstehen Sie im Zusammenhang mit Datenstreams unter einer
 - asynchronen Übertragung?
 - synchronen Übertragung?
 - isochronen Übertragung?
- •Durch welche Möglichkeiten kann man die limitierte Dienstgüteregulierung durch die Verwendung des IP-Protokolls ausgleichen?
- •Warum werden bei Multicast-Kommunikation in Peer-to-Peer-Netzwerken selten optimale Routen verwendet?
- •Welche Metriken zur Einschätzung der Qualität eines Multicast-Baumes kennen Sie?
- Was können Sie zur Informationsverbreitung in Peer-to-Peer-Netzwerken basierend auf endemisches Verhalten sagen?

Themenüberblick

.Benennung und Namenssysteme

- Begrifflichkeiten
- Lineare Benennung
- Hierarchische Benennung
- (Attributbasierte Benennung)
- Synchronisierung

Benennung und Namenssysteme Einstiegsfragen

Was verstehen Sie unter den Begriffen Entität, Adresse und Bezeichner?

Wie findet man den Rechner zu einer IP-Adresse in lokalen Netzwerken?

- Wie kann das Finden von Entitäten implementieren?
 - Auf einem Rechner?
 - Auf einem Rechner unter Verwendung externer Laufwerke?
 - In verteilten Systemen?
 - Im Internet?

Benennung und Namenssysteme

.Begrifflichkeiten

- Entitäten
 - Bsp.: Hosts, Drucker, Festplatten, Dateien, ...
- Adresse einer Entität
 - Entspricht Zugriffspunkt auf eine Entität

Benennung und Namenssysteme

.Begrifflichkeiten

- Bezeichner einer Entität
 - Unabhängig von Adresse → ortsunabhängig
 - Einfachere und flexiblere Handhabung
 - Echter Bezeichner, wenn
 - Jeder Bezeichner verweist auf höchstens eine Entität.
 - Auf jede Entität verweist höchstens ein Bezeichner.
 - Ein Bezeichner verweist immer auf die gleiche Entität.
- Benutzerfreundliche Namen
 - Auf die Benutzung von Menschen zugeschnitten

Benennung und Namenssysteme

- Bezeichner sind zufällige Bit-Folgen (unstrukturierter / linearer Name)
- Bezeichner enthält keine Info, wie Zugriffspunkt lokalisiert werden kann
- Bsp.: MAC-Adresse

- Verwendung von Broadcast
 - Nur für lokale Netzwerke praktikabel
 - Aufforderung an alle, wer den Bezeichner der Entität als Inhalt hat
 - Nur Rechner, die Zugriffspunkt für Entität bieten, melden sich zurück
 - Bsp.: Address Resolution Protocol (ARP)
 - (Welcher Rechner im Netzwerk besitzt eine bestimmte IP-Adresse?)

- Verwendung von Multicast
 - Nur für lokale Netzwerke praktikabel, wenn Multicast-Unterstützung auf Netzwerkschicht (Sicherungsschicht) vorhanden
 - z.B. mobile Host: neben Zuweisung der IP, Eintrag in Multicast-Gruppe

•Annahme:

Die Lokalisierung eines Dienstes ist bekannt.

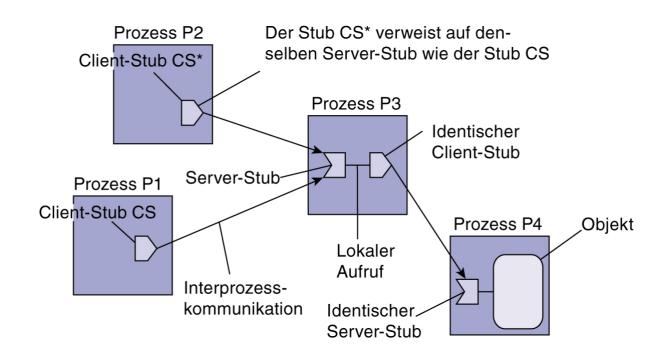
•Wie kann man damit umgehen, wenn dieser Dienst umzieht (ggf. mehrfach)?

Zeiger zur Weiterleitung

- Ansatz zur Lokalisierung mobiler Entitäten
- Bei Verschiebung einer Entität → Hinweis (Zeiger) auf neuen Ort wird hinterlassen
- Client: Nach Lokalisierung der Entität → Folge den Zeigern

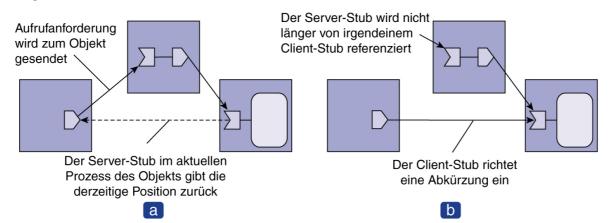
Frage:

Was sind die Nachteile dieses Vorgehens?



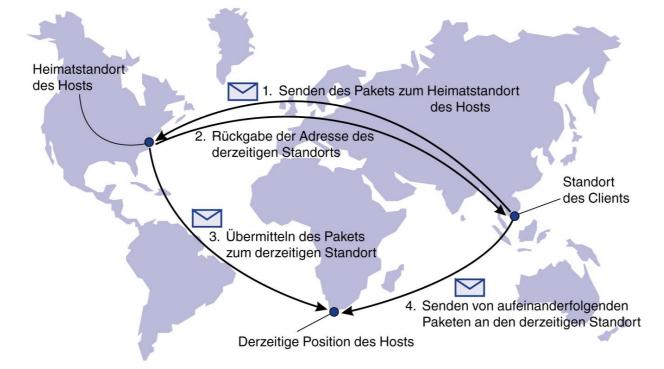
- •Zeiger zur Weiterleitung Teil 2
 - Nachteile:
 - Kette kann sehr lang werden (bei hochgradig mobilen Entitäten)
 - Weiterleitung muss so lange wie möglich gehalten werden
 - Anfälligkeit für gebrochene Verknüpfungen
 - Zielsetzung:
 - Ketten kurz halten und Stabilität sicherstellen

- Zeiger zur Weiterleitung Teil 3
 - Lösungsmöglichkeit:
 - Neuer Standort wird zurück übermittelt
 - Direkt an auslösenden Client-Stub
 - Zurück über die Kette an den auslösenden Client-Stub
 - Schwierigkeit:
 - Was passiert mit den Weiterleitern?



·Heimatgestützte Ansätze

- Unterstützung bei mobilen Entitäten
- Einführung des Heimatstandortes einer Entität
- Bewegung der Entität
 - Care-of-Adresse wird beim Heimatagenten hinterlegt



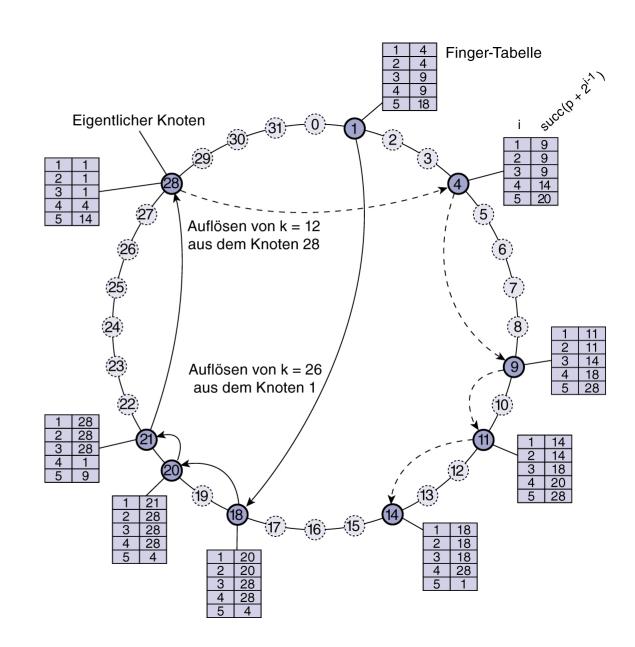
•Hintergrund: Peer-to-Peer-Netzwerke

•Wie ließen sich Entitäten in solchen Netzwerken finden?

•Was könnte eine Zielsetzung hinsichtlich der Verteilung der Entitäten sein?

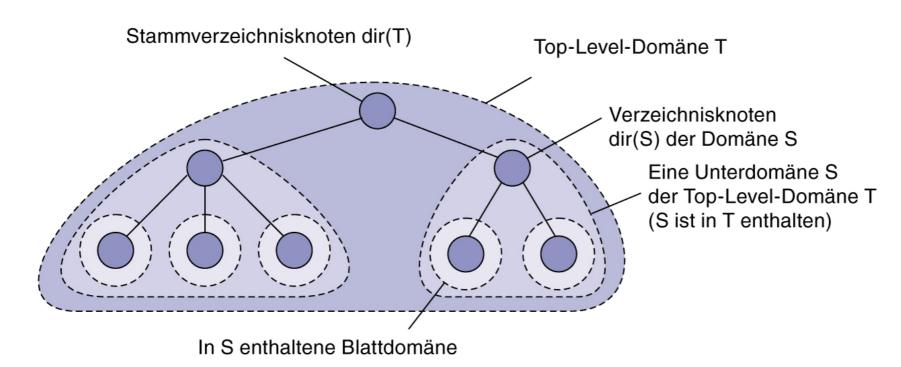
- Verteilte Hashtabelle (Distributed Hash Table, DHT)
 - z.B. zwecks Speicherung des Speicherorts von Dateien
 - Prinzip:
 - Zuweisung der Datenobjekte von Schlüsseln (im linearen Wertebereich, 1 Datenobjekt = 1 Schlüssel)
 - Jeder Knoten ist zuständig für Teilbereichs des Schlüsselraums.
 - Zuständigkeiten können sich dynamisch ändern.
 - Routing zu einem zuständigen Knoten über Verantwortlichkeiten der Knoten für Teilbereiche möglich.

- Beispiel für Verteilte
- Hashtabellen



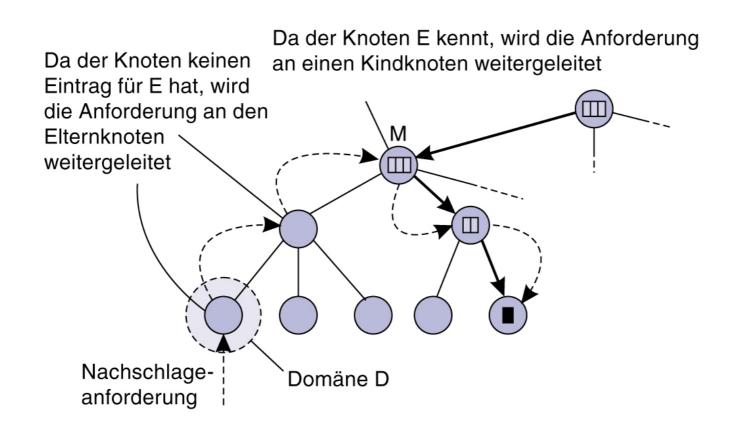
·Hierarchische Ansätze

- Standorteintrag einer Entität nur auf unterster Domäne
- Darüber liegende Domänen haben zur Entität nur Referenz auf die tieferliegende Domäne



Hierarchische Ansätze

Nachschlagen eines Standorts



Themenüberblick

Benennung und Namenssysteme

- Begrifflichkeiten
- Lineare Benennung
- Hierarchische Benennung
- (Attributbasierte Benennung)
- Synchronisierung

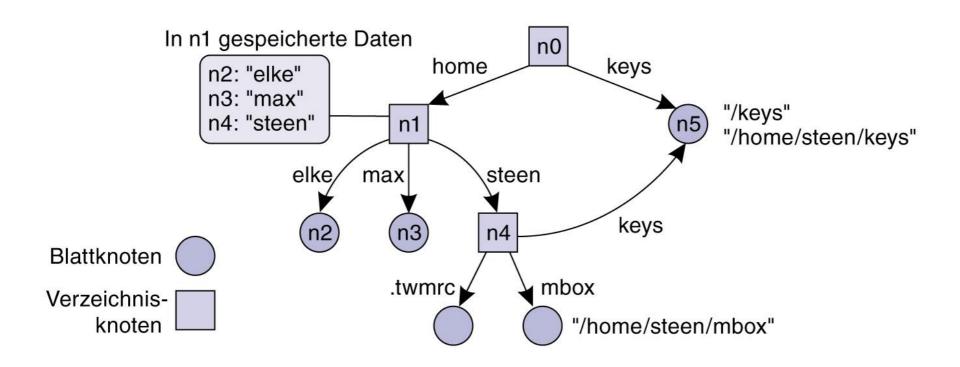
Nennen Sie mindesten 2 Beispiele für hierarchische Benennungsschemata!

- •Was verstehen Sie unter einem Nameserver?
- •Welche Anforderungen werden an Nameserver gestellt?

•Auf welche Weise kann man die Namensauflösung mit Hilfe von Nameservern realisieren?

- •Hierarchische Namen → beschriftete, gerichtete Graphen mit
 - Blattknoten
 - Verzeichnisknoten
- Blattknoten
 - Keine ausgehenden Kanten
 - Steht für eine Entität und enthält allg. Infos (z.B. Adresse)
- Verzeichnisknoten
 - Besitzt gewisse Anzahl ausgehender Kanten mit Namen
 - Ausgehende Kanten in einer Tabelle
 - Verzeichnisknoten ebenso Entität mit Bezeichner

Beispiel eines allgemeinen Namensgraphen mit einem einzelnen Wurzelknoten



- Nutzung von Pfadnamen
 - "absoluter Pfad", wenn erster Knoten im Pfandnamen auf Wurzel referenziert
 - "relativer Pfad", andernfalls (ausgehend vom aktuellen Ort)
- •Pfadnamen als Zeichenkette mit Trennzeichen (z.B. Schrägstrich oder ..., z.B. /root/home/....)
 - Aber: Knoten kann durch verschiedene Pfadnamen ausgedrückt werden (siehe Hard Links, Symbolic Links)

- •Namensauflösung:
- •N: < Beschriftung₁, Beschriftung₂, ..., Beschriftung_n>
 - Start bei Knoten N und Bezeichner für Beschriftung₁ suchen
 - Im Verzeichnis zu "Beschriftung₁" nach Beschriftung₂ suchen

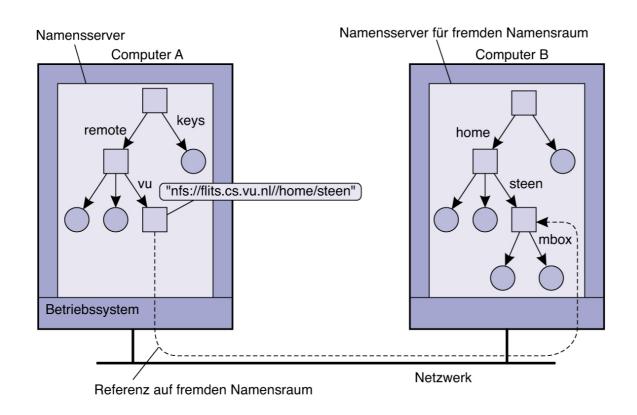
– ...

•Mounting unter Unix/Linux

- Virtuelles Dateisystem (VFS virtual file system)
- Systemaufrufe um Unix- und Network-Dateisysteme anzusprechen
- Idee: Wurzelverzeichnis eines Dateisystems wird an eine Stelle des Virtuellen Dateisystems gemountet
 - z.B. CD-ROM unter /media/cdrom
- Nutzer -> Standard POSIX Systemaufrufe an das VFS
- VFS greift über spezielle Funktionen auf jeweiliges Dateisystem zu

- •Mounting in Verteilten Systemen
 - Network File System (NFS)
 - Notwendigkeit spezieller Infos
 - Name des Zugriffsprotokolls
 - Servername
 - Name des Mountpoints im fremden Namensraum

Beispiel für Mounting in Verteilten Systemen

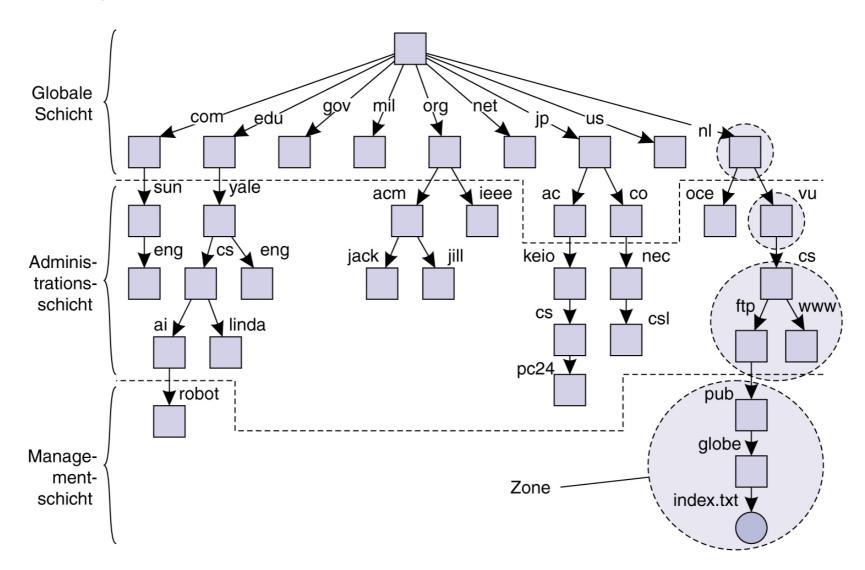


Implementierung eines Namensraumes

- Globale Schicht
 - Knoten der höchsten Ebene, Wurzelknoten und seine Kindknoten
 - Verzeichnistabellen ändern sich selten
 - Abbildung von Organisationen oder Gruppen von Organisationen
- Administratorenschicht
 - Verwaltungsknoten innerhalb einer Organisation
 - Relative Stabilität, aber mehr Änderungen als in globaler Schicht

- Implementierung eines Namensraumes
 - Managementschicht
 - Knoten mit regelmäßigen Änderungen
 - Host im lokalen Netzwerk oder gemeinsam genutzte Dateien

Implementierung des DNS-Namensraumes



Anforderungen an die Nameserver

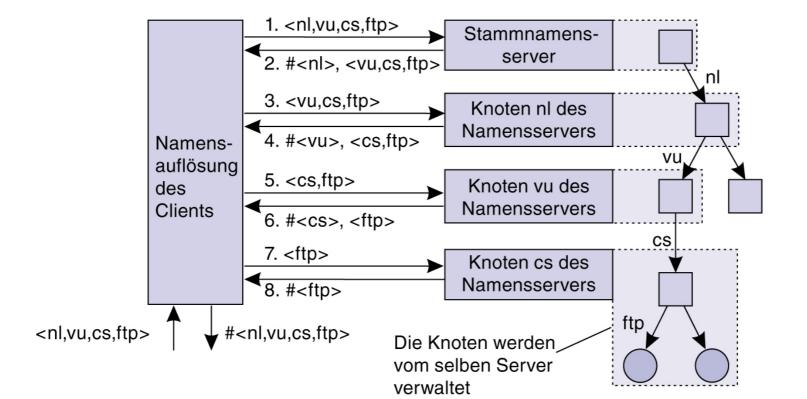
Aspekt	Globale Schicht	Administrations- schicht	Management- schicht
Geografische Ausdehnung des Netzwerkes	Weltweit	Organisation	Abteilung
Knotengesamtzahl	Wenige	Viele	Zahllose
Reaktionsfähigkeit auf Nachschlageanforderungen	Sekunden	Millisekunden	Sofort
Verbreitung von Aktualisierungen	Verzögert	Sofort	Sofort
Anzahl Replikate	Viele	Keine oder wenige	Keine
Caching durch Clients	Ja	Ja	Manchmal

- Anforderungen an die Nameserver
 - Globale Schicht
 - Hohe Verfügbarkeit, ansonsten Nichterreichbarkeit eines Teilnetzes
 - Antwortverhalten weniger zeitkritisch, da meistens Caching (bei den Clients)
 - Verwendung vieler Replikate von Servern
 - Verbreitung von Aktualisierungen kann verzögert sein.
 ("Fehlt ein Eintrag, dann frag den Chef.")

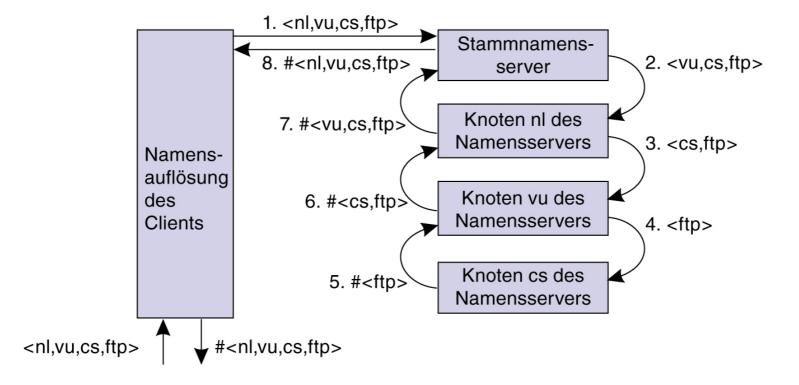
- Anforderungen an die Nameserver
 - Administrationsschicht
 - Verfügbarkeit für Clients der Organisation
 - Schnelle Rückgabe bei Nachschlageergebnisse
 - Verwendung von Caching und Replikaten
 - Managementschicht
 - Verfügbarkeit unkritisch, aber schnelle Antwortzeiten und sofortige Verbreitung von Aktualisierungen

- Implementierung der Namensauflösung
 - Annahme: keine Replikate und kein Caching
 - Ansätze:
 - Iterative Namensauflösung
 - Rekursive Namensauflösung

- Implementierung der Namensauflösung
 - Iterative Namensauflösung Beispiel:
 - Auflösung von ftp://ftp.cs.vu.nl/pub/globe/index.html
 - bzw. root:<nl, vu, cs, ftp, pub, globe, index.html>

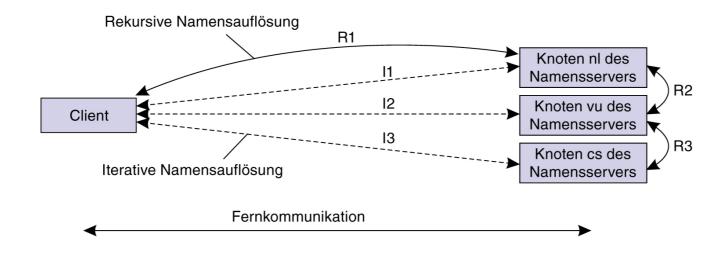


- Implementierung der Namensauflösung
 - Rekursive Namensauflösung Beispiel:
 - Auflösung von ftp://ftp.cs.vu.nl/pub/globe/index.html
 - bzw. root:<nl, vu, cs, ftp, pub, globe, index.html>



Implementierung der Namensauflösung

- Nachtteil der rekursiven Namensauflösung
 - Größere Anforderungen an Leistung jedes Nameservers
 - Resultat: Server der globalen Schicht meist nur iterativen Ansatz
- Vorteile der rekursiven Namensauflösung
 - Leistungssteigerung durch Caching
 - Senkung der Kommunikationskosten
- Optimierung bei iterativen Namensauflösung
 - Nutzung eines lokalen zwischengeschalteten Nameserver



Beispiel hinsichtlich der Kommunikationskosten:

Themenüberblick

Benennung und Namenssysteme

- Begrifflichkeiten
- Lineare Benennung
- Hierarchische Benennung
- (Attributbasierte Benennung)
- Synchronisierung

Attributbasierte Benennung Konzept

- •Kernziele der linearen und hierarchischen Benennung:
 - Standortunabhängigkeit und Benutzerfreundlichkeit

- •Attributbasierte Benennung:
 - Suche einer Entität aufgrund Beschreibung
 - Hinterlegung von (Attribut, Wert)-Paaren je Entität
 - Aufgabe
 - Anfrage: Beschreibung eines Benutzers
 - Rückgabe: eine oder mehrere passende Entitäten

Themenüberblick

- Benennung und Namenssysteme
- .Synchronisierung
 - Uhrensynchronisierung
 - Logische Uhren

•Fragen:

- Warum ist Uhrensynchronisierung wichtig in verteilten Systemen?
- Wie findet eigentlich Zeitmessung heute statt?
- Welche Schwierigkeiten könnte es bei der Uhrensynchronisierung in verteilten Systemen geben?
- Welche Möglichkeiten / Ansätze bestehen Uhren mehrerer Rechner zu synchronisieren?

- Wichtigkeit exakter Zeitmessung
 - Aktienhandel
 - Sicherheitsprüfungen
 - Messungen

- ...

- Wichtigkeit der Uhrensynchronisierung
 - Reihenfolge von Ereignissen, z.B.
 - Einzahlung von 100€, danach Zinsaufschlag von 10%
 - Zinsaufschlag von 10%, danach Einzahlung der 100€

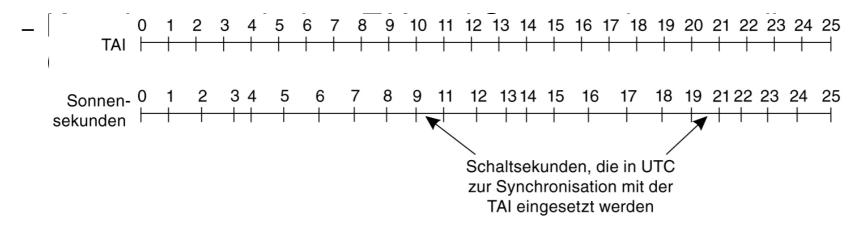
Problematik der Zeitmessung

- Mit der Einführung mechanischer Uhren im 17ten Jahrhundert
 → astronomische Zeitmessung
- Meriadiandurchgang der Sonne = scheinbar höchste Punkt der Sonne am Himmel
- Zeitliche Distanz zw. 2 Meridiandurchgängen = 1 Sonnentag
- => 1 Sonnensekunde = 1/86400 Sonnentag

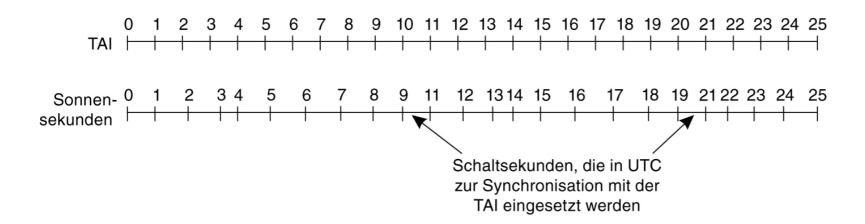
- Problematik der Zeitmessung
 - Achtung: Der Sonnentag ist nicht konstant!!
 - Vor 300 Mio Jahren -> 1 Jahr entspricht ungefähr 400 Tage
 - Grund: Gezeitenreibung
 - Kurzfriste Schwankungen ebenso möglich
 - Grund: Turbulenzen im Inneren des Erdkerns

Zeitmessung

- Atomuhr seit 1948
 - 9.192.631.770 Übergänge des Caesium-Atoms = mittlere Sonnensekunde im Einführungsjahr
- Einführung der Internationalen Atomzeit TAI (Temps Atomique International)
- Problematik: TAI ist konstant, Sonnensekunden nicht



48



(Für Interessierte bzgl. Schaltsekunden: https://www.timeanddate.de/zeitzonen/schaltsekunden)

 Verfügbarkeit der UTC in Deutschland Über Langwellensender der Physikalisch-Technischen Bundesanstalt in Braunschweig (PTB-Zeit, siehe https://www.ptb.de/)

•Problemstellung 1:

- 1 Rechner besitzt Empfänger für UTC-Zeit
- Lokale Uhren auf Rechner haben Drift (aufgrund Ungenauigkeiten)
- Aufgabe: Aller anderen Rechner sollen dazu synchron gehalten werden.

Verfügbarkeit der UTC in Deutschland Über Langwellensender der Physikalisch-Technischen Bundesanstalt in Braunschweig (PTB-Zeit, siehe https://www.ptb.de/)

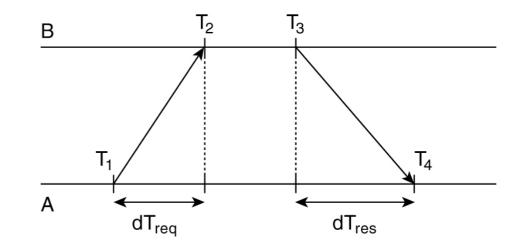
•Problemstellung 2:

- Kein Rechner besitzt Empfänger für UTC-Zeit
- Lokale Uhren auf Rechner haben Drift (aufgrund Ungenauigkeiten)
- Aufgabe: Abweichungen kompensieren, Zeitdrift entgegenwirken

.Network Time Protocol (NTP)

- A sendet Zeit an B
- B antwortet und übermittelt
- Zeiten
- Übertragungszeiten

$$\delta = (dT_{req} + dT_{res}) / 2$$



- Abweichungsberechnung: $\theta = T_3 + \delta T_4$
- Messung: 8 Wertepaare (θ,δ)
 - kleinster Wert f
 ür δ als Sch
 ätzung f
 ür Verz
 ögerung
 - Zugehöriges θ als zuverlässigste Schätzung der Abweichung

.Network Time Protocol (NTP) — Teil 2

- Probleme 1: θ < 0 (Uhr von A geht zu schnell)
 - Zurücksetzung der Zeit nicht erlaubt (Konsistenzprobleme auf A)
 - "Verlangsamen" der Uhr auf A (z.B. statt je Timeinterrupt 10ms hinzuzufügen, nun nur +9ms)
 - Hinweis: Analoges Vorgehen, wenn A zu langsam geht

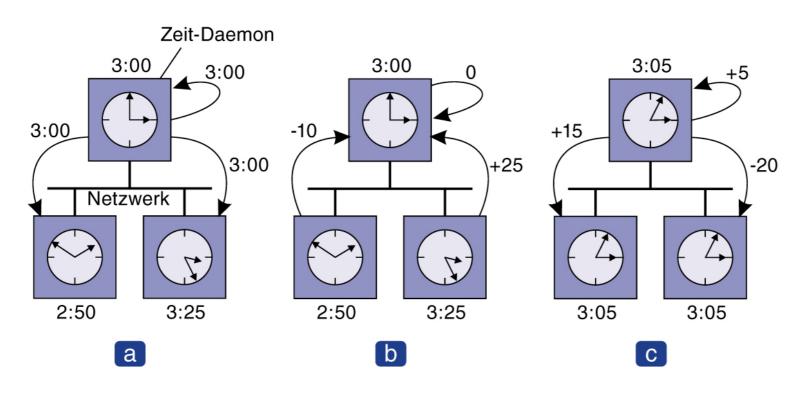
.Network Time Protocol (NTP) — Teil 2

- Probleme 2: Welche Uhr ist genauer?
 - Einteilung der Server in Ebenen (Strata)
 - Stratum 0 ist Refernzuhr (z.B. PTB-Zeit)
 - Stratum 1 ist Empfänger der UTC-Zeit
 - Server, der mit Stratum-n-Server synchronisiert, wird Stratum-(n+1)-Server

.Berkley-Algorithmus

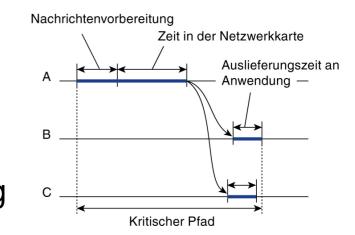
- Aktiver Zeitserver (Zeit-Daemon) ohne UTC-Empfang
- Anfrage vom Zeitserver an alle Rechner nach deren Systemzeit
- Aus Antworten: Berechnung einer durchschnittlichen Zeit
- Mitteilung an jeden Rechner hinsichtlich Verlangsamung oder Beschleunigung zwecks Synchronisierung
- Ziel: Einigung aller Rechner auf selbe Zeit

.Berkley-Algorithmus – Teil 2



- (a) Der Zeit-Daemon fragt alle anderen Rechner nach ihren Uhrzeiten.
- (b) Die Rechner antworten.
- (c) Der Zeit-Daemon teilt allen mit, wie sie ihre Uhren einzustellen haben.

- Uhrzeitsynchronisierung in drahtlosen Netzwerken
 - Bsp.: Sensornetzwerk
 - Problematik: Verwendung von Konkurrenzprotokollen
 - Zeitzusammensetzung zwecks Austausch
 - Nachrichtenvorbereitung
 - Zeit in der Netzwerkkarte
 - Übertragungszeit
 - Auslieferungszeit an die Anwendung



- Hinweise:
 - Übertragungszeit annähernd konstant (wenn keine Hops)
 - Zeit in Netzwerkkarte mit größten Schwankungen

Nachrichtenvorbereitung

A Auslieferungszeit an Anwendung —

Kritischer Pfad

- RBS (Reference Broadcast: Synchronization)
 - Anwendung für Synchronisierung
 - in drahtlosen Netzwerken
 - Besonderheit: Sender wird nicht synchronisiert
 - Nachricht an Empfänger, wobei Zeitmessung beim Verlassen der Netzwerkkarte beginnt
 - Unsicherheitsfaktoren seitens des Senders beseitigt

Nachrichtenvorbereitung

A Auslieferungszeit an Anwendung —

RBS (Reference Broadcast: Synchronization)

- Idee:

- Aussenden von m-Referenznachrichten von Knoten pimit Zeit T_{D.m}
- . Empfang am Konten q und Vergleich der Zeiten $T_{p,k}$ - $T_{q,k}$
- Mittelung der Abweichung → Abspeichern der Abweichung anstatt der Anpassung der Uhrengeschwindigkeit
- Verbesserung unter Berücksichtigung der Drift:

$$Offset_{p,q}(t) = a t + b$$

(lineare Approximation der Drift)

Themenüberblick

- Benennung und Namenssysteme
- .Synchronisierung
 - Uhrensynchronisierung
 - Logische Uhren

Beobachtung

- Uhrzeitsynchronisierung zwar möglich, aber kein absolutes Muss
- Bei nicht-wechselwirkenden Prozessen: fehlende Synchronisierung nicht beobachtbar und kein Resultat von Fehlern
- Aber wichtig: Einigung auf eine Reihenfolge von Ereignissen

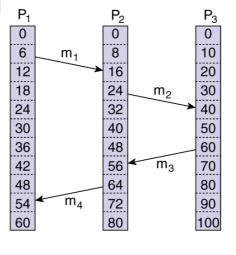
•Frage: Wie kann man die Reihenfolge von Ereignissen in verteilten Systemen sicherstellen?

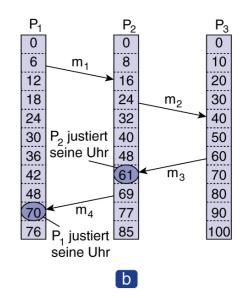
.Lamport-Zeit / Logische Uhr von Lamport

- Bezeichnung: a→b
- Bedeutung: "a passiert vor b" ("happens-before")
- Beobachtung:
 - Im selben Prozess passiert a vor b, dann a→b
 - a entspricht Senden einer Nachricht, b entspricht Empfang einer Nachricht, dann a→b
- Hinweis: Happens-before-Relation ist transitiv
- Wenn a→b und b→c, dann gilt auch a→c
- Einführung eines Zeitwertes C(a),
- wenn a→b, dann C(a) < C(b)

•Lamport-Zeit / Logische Uhr von Lamport – Teil 2

- Bedingungen: Ist a ein Ereignis in Prozess P_i und b ein Ereignis in Prozess P_i, dann gilt a→b wenn
- (i) $C_i(a) < C_j(b)$ oder
- (ii) $C_i(a) = C_j(b)$ und $P_i < P_j$
- (Berücksichtigung der
- Prozessnummern)
- Prinzip: Anpassung der
- jeweiligen Uhren



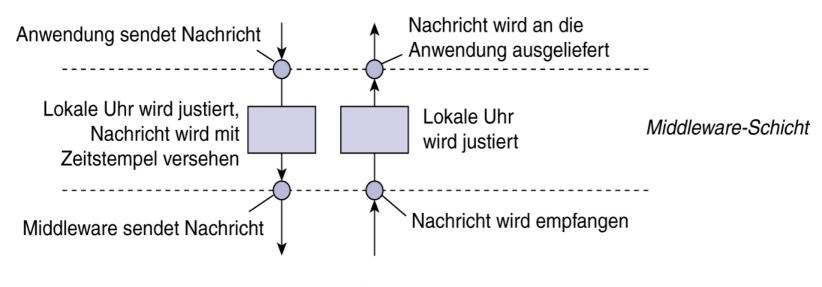


(a) Drei Prozesse, von denen jeder eine eigene Uhr hat. Die Uhren laufen mit unterschiedlichen Geschwindigkeiten.63 (b) Der Algorithmus von Lamport korrigiert diese Uhren.

•Lamport-Zeit / Logische Uhr von Lamport – Teil 3

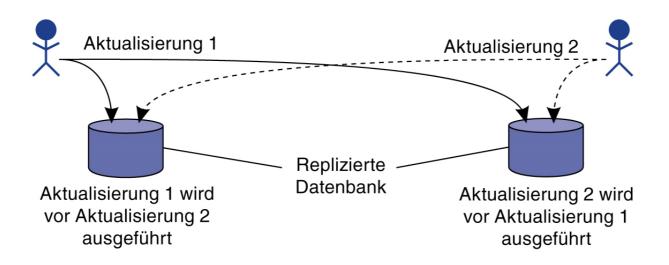
 Lokalisierung der logischen Uhren von Lamport in der Middleware-Schicht

Anwendungsschicht



Netzwerkschicht

- Anwendung: Vollständig geordnetes Multicasting
- •Beispiel: Kontoguthaben 1000€
 - Aktualisierung 1: Einzahlung von 100€
 - Aktualisierung 2: Gutschrift von Zinsen zu 1%
- Problem: Reihenfolge muss auf allen Datenbanken gleich sein



- Anwendung: Vollständig geordnetes Multicasting
- •Annahmen:
 - Keine Nachrichten gehen verloren
 - Eintreffende Nachrichten kommen auch in der Sendereihenfolge eines Sender an

Anwendung: Vollständig geordnetes Multicasting

•Lösung:

- Jede Nachricht mit aktuellem (logischen) Zeitstempel des Senders versehen an alle Knoten (inklusive sich selber)
- Empfänger
 - Einsortierung in Warteschlange gemäß dem Zeitstempel
 - Senden einer Bestätigung via Multicast
- Resultat: Alle Prozesse haben gleiche Kopie der lokalen Warteschlange.

.Vektoruhren

- Bei logischer Uhr von Lamport → Vollständige Ordnung (gemäß der Bedingungen)
- Nachteil bei Lamport: Keine Aussage über Kausalität
- Ziel:
 - Kausalzuordnung möglich
 - Nebenläufigkeit von Ereignissen ermittelbar

.Vektoruhren – Teil 2

- Idee: Vektor mit Zeit für jeden Prozess wird geführt
- Aktualisierung der eigenen Zeit
- Versenden des gesamten Vektors
- Anwendbarkeit:
- Monitor- und
- Debugsysteme

