Resümee: Konzepte der Vorlesung

- Prinzipielle Phänomene und Begriffe herausarbeiten
  - Kausalität, Konsistenz, verteilte Berechnung, safety und liveness,...
- Geeignete Modelle und Abstraktionen entwickeln
  - z.B. Zeitdiagramme, Atommodell, Zustandsgitter, Gummibandtransform.
- Problemlösungs-, Analyse- und Verifikationstechniken
  - z.B. Beweis über Invarianten
- Techniken, Einsichten, Zusammenhänge,...
  - Komplexitätsanalyse
  - Transformationen zwischen Problemklassen
  - Problemverständnis von einem höheren Standpunkt

Resümee (1)

- Verteilte Systeme
- Kooperation durch Kommunikation
  - keine globale Sicht
  - keine gemeinsame Zeit
  - parallel
  - nicht-deterministisch
  - unbestimmte Nachrichtenlaufzeit
- Typische Probleme verteilter Systeme / Algorithmen:
  - Beobachtungsproblem (keine Gleichzeitigkeit)
  - Schnappschussproblem (wieviel Geld ist in Umlauf?)
  - Deadlockproblem (Phantomdeadlock?)
  - Kausalitätsproblem (indirekte Wirkung vor Ursache)
  - Terminierungserkennungsproblem
- Problem globaler Prädikate ("relativistischer Effekt")
  - es gibt i.a. mehrere "gleichberechtigte" Beobachter
  - diese stimmen i.a. bzgl. der Gültigkeit des Prädikates nicht überein!
  - gibt es beobachterinvariante Prädikate?
- Verteilter Euklidischer Algorithmus
  - als erstes Beispiel für einen verteilten Algorithmus
  - reaktives Verhalten ("nachrichtengesteuert")
  - Korrektheit der Idee / des konkreten Algorithmus? (Invarianten...)
- Übungen

Resümee (2)

- Folienkopien auf der Homepage der Vorlesung:
  - `www.inf.ethz.ch/department/IS/vs/lectures/WS9900/VA/`

- Verteilter Euklidischer Algorithmus
  - als erstes Beispiel für einen verteilten Algorithmus
  - reaktives Verhalten ("nachrichtengesteuert")
  - Korrektheit der Idee / des konkreten Algorithmus? (Invarianten...)
- Zahlenrätsel
  - Parallele Constraint-Propagation
  - Abwechselnd mit Backtracking-Schritt

allgemeines Schema ("verteilte Approximation")

- Konzeptuelle Hilfsmittel
  - Zeitdiagramme
  - Atomare Aktionen
- Problem der *verteilten Terminierung*
  - Geeignete Definition?
  - Verfahren zur Feststellung?
- Übungen

Resümee (3)

- Flooding-Algorithmus
  - Nachrichtenzahl
  - Problem der Terminierungserkennung
- Echo-Algorithmus (Variante von Flooding)
  - Nachrichtenzahl 2e
  - Explorer- / Echo-Welle
  - Spannbaum
  - Überlagerung; upcall / downcall
  - Formalere Fassung in Pseudo-Code
  - Zwei "disjunkte" Wellen (rot; grün)
  - Verbesserung durch Mitführen von Knotenidentitäten?

Resümee (4)

- Zeitkomplexität
  - Einheitszeitkomplexität
  - Variable Zeitkomplexität

- Besprechung Übungen (1): verteilte ggT-Berechnung
  - Varianten des Algorithmus
  - Verifikationsidee

Resümee (5)

- Broadcasts auf Hypercubes
  - Hypercube: Definition und Eigenschaften
  - Einzelnachrichten: Routingverfahren, mittlere und max. Weglänge
  - Broadcast entsprechend der rekursiven Definition
  - Broadcast durch Fluten in jeweils höhere Dimensionen
  - Optimalität (Nachrichten- und Zeitkomplexität) des Broadcastproblems
  - schneller Broadcast durch paralleles Senden von Teilnachrichte

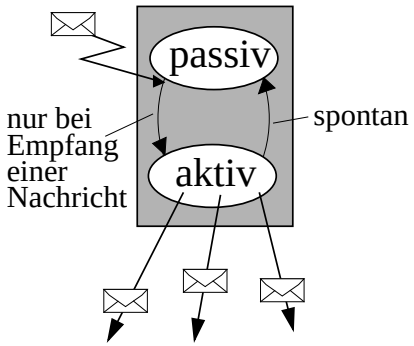
- Berechnung von Routing-Matrizen
  - verteilte Version des Bellmann-Ford-Algorithmus
  - auch wieder das bekannte Schema der verteilten Approximation
  - Anwendung in Rechnernetzen: Count to infinity-Problem, Spann-baum für lokale Netze (Zyklenfreiheit)

Resümee (5)

- Paradigma der verteilten Approximation
  - Verallgemeinerung verschiedener ähnlicher Algorithmen

- Verteilte Terminierung
  - Problemdefinition

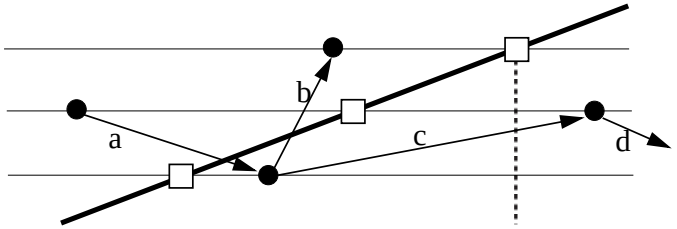
- Atommodell
  - Vereinfacht die Betrachtung des Wesentlichen
  - Terminierungskriterium: "keine Nachricht unterwegs"



- Schiefes Bild beim Beobachten verteilter Berechnungen

  --> Pauschales Zählen von Nachrichten genügt nicht zur Erkennung der verteilten Berechnung

  --> Suche nach den eigentlichen Ursachen für Fehlschlag des Zählkriteriums



- Lösungsansätze zur Terminierungserkennung
  - Durch Vermeidung der "Ursachen" für das schiefe Bild

Resümee (6)

- Terminierungserkennung: *Eindeutige Nachrichtennamen*

- Terminierungserkennung: *Kanalzählerkriterium*
  - Widerspruchsbeweis (es gibt kein frühestes Ereignis nach dem Schnitt)

- Terminierungserkennung: *Doppelzählverfahren*
  - informeller Beweis (Aussage über gedachten senkrechten Schnitt zwischen den beiden Wellen)
  - formaler Beweis

- Safety- und Liveness-Eigenschaften verteilter Algorithmen

- Kontrolltopologien zur Realisierung von Schnitten
  - Ring
  - Spannbaum
  - Echo-Algorithmus als zugrundeliegendes Basisverfahren (Hin- und Rückwelle für die beiden Schnitte des Doppelzählverfahrens!)

Resümee (7)

- Terminierungserkennung: *Zeitzonenverfahren*
  - Prinzip: *Erkenne* "Nachricht aus der Zukunft"
  - binäre "schwarz/weiss"-Zeit genügt

- Terminierungserkennung: Vermeiden inkonsistenter Schnitte durch geeignetes Vorziehen der Schnittlinie

- Synchrones / asynchrones Senden
  - synchron: senkrechte Nachrichtenpfeile sind gerechtfertigt
  - nicht alles geht synchron (Deadlock-Beispiele und andere Zyklen)

Resümee (8)

- Besprechung Teile von Übung 1
  - Formalisierung Raum-Zeit-Diagramm, Ereignis, Kausalitätsrelation
  - kausaltreue Beobachtungen als lineare Erweiterungen ("Einbettung") der halbgeordneten Kausalitätsrelation
  - Zyklenfreiheit, Schnitt aller Beobachtungen,...

- Charakterisierung synchroner Kommunikation
  - alle Nachrichtenpfeile können senkrecht gezeichnet werden; Kommunikationskanäle sind immer leer
  - es gibt eine lineare Erweiterung der Kausalitätsrelation, so dass ein Empfangsereignis immer direkt nach seinem Sendeereignis kommt
  - Senden und Empfangen bilden "atomare Einheit"
  - Nachrichten-Scheduling-Relation (m < n gdw. send(m) < receive(n)) ist zyklenfrei
  - kein Zyklus im Raum-Zeit-Diagramm, auch wenn Nachrichtenpfeile rückwärts durchlaufen werden können
  - Zyklenfreiheit der "synchronen Kausalitätsrelation <<" ("common past" / " common future"); dadurch Identifizierung von send und receive

zusammengehörige send/receive-Ereignisse sind "in gewissem Sinne" atomar

- Fragen...
  - sind die Charakterisierungen alle äquivalent?
  - kann man nun Nachrichtenlaufzeiten immer vernachlässigen?
  - funktioniert ein Algorithmus, der unter der Voraussetzung synchroner Kommunikation gemacht wurde, auch bei asynchroner Kommunikation?
  - und umgekehrt?
  - Terminierungserkennung bei synchroner Kommunikation? (das Atommodell ist dann offenbar nicht mehr adäquat, oder?)

Mehr dazu: Charron-Bost, Mattern, Tel:
Synchronous, Asynchronous and Causally Ordered Communication.
Distributed Computing, Vol. 9 No. 4, pp. 173 - 191, 1996
http://www.informatik.tu-darmstadt.de/VS/Publikationen/papers/syn_asy.ps

Resümee (9)

- Def. verteilte Terminierung bei synchroner Kommunikation

X_p: {$state_p$ = *aktiv*}
    $state_q$:= *aktiv* (* "atomares" aktivieren *)

I_p: $state_p$:= *passiv*

- Verhaltensmodelle verteilter Anwendungen

  - Transaktionsmodell
  - Atommodell
  - Synchronmodell

  } gegenseitige Simulation bzw. Transformation der Modelle

- Modelle in der Informatik
  - nicht nur zum Erkenntnisgewinn, zur Simulation etc., sondern auch Implementierung von "ausgedachten, idealisierten Wirklichkeiten"

Resümee (10)

- Terminierungserkennung bei synchroner Kommunikation

- Algorithmus von Dijkstra et al. ("DFG")
  - schwarz / weiss-Färbung; Token auf einem Kontrollring
  - Beschreibung durch Menge von Verhaltensregeln
  - Überlegungen zu Korrektheit, Varianten, Nachrichtenkomplexität

- Sticky-Flags-Methode zur Terminierungserkennung
  - für Safety sehr einfacher Algorithmus angebbar
  - Liveness erfordert kleinen "Zusatz"
  - entspricht Empfangsflag (statt Sendeflag bei DFG-Verfahren)

Resümee (11)

- Parallele Berechnungsschemata
  - Bsp.: Integration mittels Trapezmethode
  - Lastausgleich durch Migration von Arbeitseinheiten
  - Gesamtlast = 0 <==> Terminierung
- Terminierungserkennung mit der Kreditmethode
  (Halbieren von Tickets; Einsammeln von "Krümeln")
- Safety: "Gesamtkredit" ist invariant
- Realisierung in verschiedenen Varianten möglich:
  - geeignete Darstellung der Krümel (negativer Zweierlogarithmus)
  - geeignete Realisierung des Einsammelns (Liveness!)
  - geeignete Verwaltung der Krümel bei den Prozessen
  - geeignete Informationsverwaltung im Urprozess
- Nachrichtenkomplexität: Worst-case-optimal
- Variante: direktes Nachlaufen
  - Analogie zum Echo-Algorithmus!

Resümee (12)

- Besprechung von Teilen von Übung 3
  - falscher Terminierungserkennungsalgorithmus
  - es genügt nicht, nur über den Zustand seiner Nachbarn informiert zu sein
- Wechselseitiger Ausschluss
  - safety
  - liveness
  - fairness
  - nachrichtenbasierte Lösung ("requests" etc.): zentraler Manager
  - andere Lösung: Token-Ring
- Maekawa's $O(\sqrt{n})$-Algorithmus
  - Prinzip: Gitteranordnung

Resümee (13)

- Wechselseitiger Ausschluss

- Maekawa's O($\sqrt{n}$)-Algorithmus (request-basiert)
  - Deadlockproblematik
  - Dreiecksform (--> Grösse der Request-granting-Menge: ca. $\sqrt{2}$ $\sqrt{n}$)
  - projektive Ebene --> $\sqrt{n}$

- *Token-basierte* Algorithmen
  - Umdrehen durchlaufener Kanten ("path reversal")
  - Spannbaum ("Lift-Algorithmus") --> O(log n) bei "guten" Bäumen
  - Verallgemeinerung auf beliebige (gerichtete azykl.) Graphen
  - Invarianten: Zyklenfreiheit; alle Pfade führen zum Tokenbesitzer
  - Request holt Token stets ein

Resümee (14)

- *Wechselseitiger Ausschluss*: Token mit "path reversal"
  - Variante: Nachbarn informieren, dass Token "jetzt" hier ist (Quittungen?)
  - spezielle Topologien (Ring; Stern; lineare Kette)
  - Nachrichtenkomplexität bei starker Last (≈4)

- Vergleich von Algorithmen für den wechselseitigen Ausschluss (quantitative und qualitative Kriterien)

- *Election-Problem*: Symmetriebrechung
  - Auswahl genau eines Prozesses aus mehreren (bis auf die eindeutige Identität) gleichartigen

- Election-Algorithmus mit dem Message-extinction-Prinzip
  - funktioniert auf allgemeinen (zusammenhängenden) Graphen

- Chang/Roberts-Algorithmus auf unidirektionalem Ring
  - beim Ring kein Terminierungserkennungsproblem
  - nur grösste Identität schafft Ringumlauf --> ist damit "gewählt"
  - Worst-case-Nachrichtenkomplexität $O(n^2)$
  - mittlere Nachrichtenkomplexität?

Resümee (15)

- Chang/Roberts-Algo.: Mittlere Nachrichtenkomplexität
  - Summation abhängiger aber unkorrelierter Fälle
  - Anzahl der Rekorde, Abnahme der Rekordhäufigkeit
  - Wartezeit bis zum ersten Rekord: "heuristische" Ansätze
  - Wahrscheinlichkeit für einen Rekord an Position i

Resümee (16)

- Chang/Roberts-Algo.: Mittlere Nachrichtenkomplexität
  - Wahrscheinlichkeit, genau i Positionen weit zu kommen
  - Erwartungswert für die Länge der Nachrichtenkette = H_n
  - mittlere Nachrichtenkomplexität = nH_n (= ca. n ln n)

- Qualität des Chang/Roberts-Algorithmus
  - worst-case; average-case

- Fehlertoleranz verteilter Systeme / Algorithmen
  - am Beispiel des Chang/Roberts-Algorithmus

- Bidirektionale Varianten des Chang/Roberts-Algorithmus
  - probabilistisch / deterministisch (z.B. bei Kenntnis der Nachbaridentitäten)
  - mittlere Nachrichtenkomplexität

- Algorithmus von Hirschberg und Sinclair (bidir. Ring)
  - sukzessive grössere Gebiete erobern
  - worst-case Nachrichtenkomplexität $< 8\ n \log_2 n$

Resümee (17)

- Petersons Election-Algorithmus (bidir. Ring)
  - solange sukzessive Identität in beide Richtungen senden, bis man von einem grösseren Nachbarn erfährt
  - worst-case Nachrichtenkomplexität $< 2\ n \log_2 n$
  - mittlere Nachrichtenkomplexität ca. $2\ n \log_3 n$
  - Simulation ("kostenneutral"!) auf einem unidirektionalen Ring
  - Variante mit abwechselnden Richtungen
  - worst-case Nachrichtenkomplexität (ca. $1.44\ n \log_2 n + c$) mittels Fibonacci-Folge abgeschätzt

Resümee (18)

- Election auf Bäumen
  - Explosionswellen vereinigen sich
  - Explosionswelle wird an den Blättern reflektiert
  - Kontraktionsphase endet in zwei Zentrumsknoten
  - Nachrichtenkomplexität O(n)
- Echo-Election auf allgemeinen Graphen
  - Idee wie Chang/Roberts, aber Echo-Algorithmus statt Ringumlauf
- Nachrichtenkomplexität des Election-Problems
  - mindestens e Nachrichten

- Ubiquitous Computing
- Interesse an Diplomarbeit?
  - Election / Spannbaum in Funknetzen

Resümee (19)

- Verteilte Spannbaumkonstruktion
  - Zusammenhang zum Election-Problem ("gleich schwierig")

- Anonyme Netze
  - De-Anonymisierung

- Election in anonymen Netzen
  - kein stets terminierender (deterministischer) Algorithmus möglich

- Probabilistische Algorithmen
  - Las Vegas (terminiert nicht immer, Ergebnis ist aber korrekt)
  - Monte Carlo (terminiert, aber ggf. mit falschem Ergebnis)

- Probabilistische Election-Algorithmen

  1) Itai / Rodeh: Chang/Roberts-Verfahren mit Zufallsidentität
  2) Ringshift eines einzigen Zufallsbits pro Prozess

- Interesse an Diplomarbeit?
  - Election / Spannbaum in Funknetzen

Resümee (20)

- Kausaltreue Beobachtungen
  - Beispiel: Aussterben aller Exemplare eines "Typs" (--> Terminierung)
  - Analogie: konsistente Referenzzähler (--> Garbage-Collection!)
  - Lösungen: Synchrone Kommunikation oder getrennte FIFO-Puffer pro Prozess für das Empfangen und Senden von Nachrichten

---

- Garbage-Collection
  - Objekte und Zeiger; Wurzelobjekte
  - nicht mehr von der Wurzel erreichbar --> Garbage
  - rekursives Freigeben (Zyklen bleiben übrig!)
  - *Mutator* (new, copy, delete: Manipulation von Zeigern)
  - *Collector* soll Garbage-Objekte identifizieren
  - Paradigmen: "stop the world" / on the fly (= "parallel")
  - "Mark and sweep"-Verfahren
  - bei paralleler Variante: Problem mit "behind the back copy"
    ==> Mutator / Collector müssen sich koordinieren!
    (sonst bekäme der Collector ggf. ein "schiefes Bild")

- Verteiltes Garbage-Collection (= GC in verteilten Systemen)
  - Referenzen u.U. "in transit"
  - copy nicht mehr atomar ("send/receive copy")
  - increment / decrement per Nachricht (z.B. an den Ort des Referenzzählers)
  - inc bzw. dec daher nicht "gleichzeitig" mit copy bzw. delete
  - Unterschied zwischen lokalen und "remote" Referenzen
  - lokales und globales GC (dezentral, echt parallel, typw. hierarchisch)

- Formalisierung des GC-Problems: Operationen C_p, R_p, D_p

Resümee (20b)

- Referenzzähler-Verfahren
  - Problem: "zyklischer Garbage" wird nicht entdeckt
  - bei verteilter Variante: Problem bei decrement *vor* increment

zeitlich?
kausal?

- Lösungen für verteiltes Reference-Counting:
  - prinzipiell: Causal Order garantieren (d.h. indirekte Überholungen vermeiden)
  - "naiv": auf Bestätigung jeder Increment-Nachricht warten
  - Varianten von Lermen/Maurer und Rudalics (zwei bis vier Nachrichten pro copy-Operation)

- Weighted Reference Counting (WRC)
  - Kopieren ohne Zusatznachricht: Splitten des Reference Weight

- Analogie (verteiltes) GC <==> verteilte Terminierung

- Transformation GC-Algorithmus --> Algorithmus zur Erkennung der verteilten Terminierung
  - Umformung des Terminierungsproblems in ein GC-Problem
  - darauf gegebenen GC-Algorithmus ansetzen

- Interesse an Diplomarbeit?
  - Election / Spannbaum in Funknetzen

Resümee (21)

- Zum Patent der Referenzgewichtsmethode (WRC)

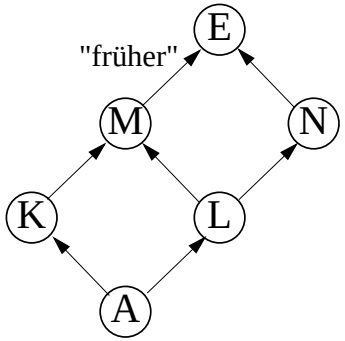
- Patentieren von Algorithmen

- Local Reference Counting (LRC)
  - jede Maschine besitzt für *jedes* Objekt einen (lokalen) Zähler

Resümee (22)

- Local Reference Counting (LRC)
  - jede Maschine besitzt für *jedes* Objekt einen (lokalen) Zähler
  - logische Baumstruktur ("Verantwortlichkeit")
  - viele interessante Eigenschaften
  - Migration von Objekten "leicht" zu unterstützen
  - lokal u.U. ein anderes GC-Verfahren (nur global LRC verwenden)
  - IRT / ORT-Tabellen ("Proxy-Objekte"; Bündelung von Referenzen)
  - Aufgabe: Transformation in den Terminierungserkennungsalgorithmus von Dijkstra und Scholten (Literatur nachlesen)

- *Verteilte Berechnungen*: Formale Definition (Modellierung!)
  - Partition von Ereignissen, Sende/Empfangsereignisse, Kausalrelation
  - Zeitdiagramme von verteilten Berechnungen; Gummibandtransformation
  - Globale Zustände als Endzustände von Präfixberechnungen (Präfixberechnungen sind linksabgeschlossen bzgl. der Kausalrelation)
  - Menge der Zustände (bzw. Präfixberechnungen) bilden Verband



Berechnung läuft entlang eines "unbestimmten" Weges vom Anfangszustand A zum Endzustand E.

Resümee (23)

- Wellenalgorithmen
  - Information verteilen / einsammeln; Phasen trennen; Ereignisse triggern...
  - Formale Def: ... init < $visit_i$ < conclude ...
  - Visit-Ereignisse bilden einen *Schnitt* (wann senkrechte Schnittlinie möglich?)
  - Bsp.: Echo-Algorithmus, Ring, Stern...
  - min. n-1 Nachrichten, min e Nachrichten bei unbekannten Nachbarn
  - Spannbaum = jeweils erste empfangene Nachricht eines Knotens
  - Halbwellen (Verzicht auf init bzw. conclude); z.B. flooding

- Virtuell gleichzeitiges Markieren mittels flooding
  - Voraussetzung: FIFO-Kanäle

- "Konsistente" Schnittlinien lassen sich senkrecht zeichnen
  - konsistent: keine Nachricht läuft "rückwärts" über die Schnittlinie

- Sequentielle Traversierungsverfahren
  - spezielle Wellenalgorithmen: visit-Ereignisse linear geordnet

- Algorithmus von Tarry (Labyrinth-Problem)

Resümee (24)

- Algorithmus von Tarry (Labyrinth-Problem)
  - Beweisskizze, dass Tarry-Algorithmus ein Traversierungsverfahren ist
  - Depth-First-Search ist Spezialfall des Tarry-Algorithmus
- Traversierungsverfahren von Awerbuch / Cidon
- Phasenalgorithmus
  - Wissen über maximale Phasenunterschiede --> Welleneigenschaft
- Algorithmus von Finn ("Mengen A = B")
  - Hüllenbildung - Kennenlernen des gesamten Graphen
  - Beweis der Welleneigenschaft über Invariante
  - geeignet für gerichtete Graphen
- Überblick / Zusammenfassung Wellenalgorithmen

Resümee (25)

- Globale konsistente Schnitte / Zustände
- Schnappschussproblem und -algorithmen

  (1) Färben von Prozessen / Nachrichten; Vermeiden von "Tachyonen"; In-Transit-Nachrichten durch Abgleich von Sende-/Empfangspuffern oder durch Weiterleiten von Kopien an den Initiator

  (2) Chandy/Lamport-Algorithmus: Flooding; FIFO-Kanäle ("flushing"); Problem (?): einige Kanäle sind scheinbar immer leer

Resümee (26)

- Beobachten verteilter Berechnungen
  - Wunsch: lückenlos konsistente Schnappschüsse anzeigen
  - rekonstruiertes Bild des Beobachters
  - ideale und kausaltreue Beobachter
- Kausaltreues Beobachten
  - Beispiele für kausal inkonsistente Beobachtungen
  - Def. kausaltreuer Beobachter
  - Pfade im n-dimensionalen Zustandsgitter ("Hyperwürfel")
- Entdecken globaler Prädikate durch Beobachtung
  - Abhängigkeit von konkreten Beobachtungen ("possible worlds")
  - Wirkung von Handshake- und Barrier-Synchronisation
- Stabile Prädikate
- Logische Zeit
  - Zeitstempel von Ereignissen
  - Uhrenbedingung. kausale Unabhängigkeit
  - Lamport’s logische Uhren
    - Definition
    - Realisierung
    - Eigenschaften (kausaltreu, Uhrenbedingung nicht umkehrbar)
- Schnitte und Vektorzeit
  - Später- / Früher-Relation auf Schnitten
  - Definition konsistenter Schnitte als linksabgeschlossene Ereignismengen
  - Zeitstempel eines Ereignisses als Menge seiner kausalen Vorgänger (Repräsentation durch lokal letztes Ereignis --> Vektorzeit)

Resümee (27)

- Vektorzeit
  - Interpretation: repräsentiert gesamte kausale Vergangenheit
  - Zeitstempelarithmetik
  - Implementierung (Supremum beim Empfang)
  - Isomorphie der Zeit- und Kausalstruktur
- Anwendung der Vektoruhren
  - kausaltreue Beobachtungen; kausaler Broadcast
- Relativistische Struktur der Vektorzeit