Distributed System Paradigms (5)

3. Group (multipoint, multicast) communication

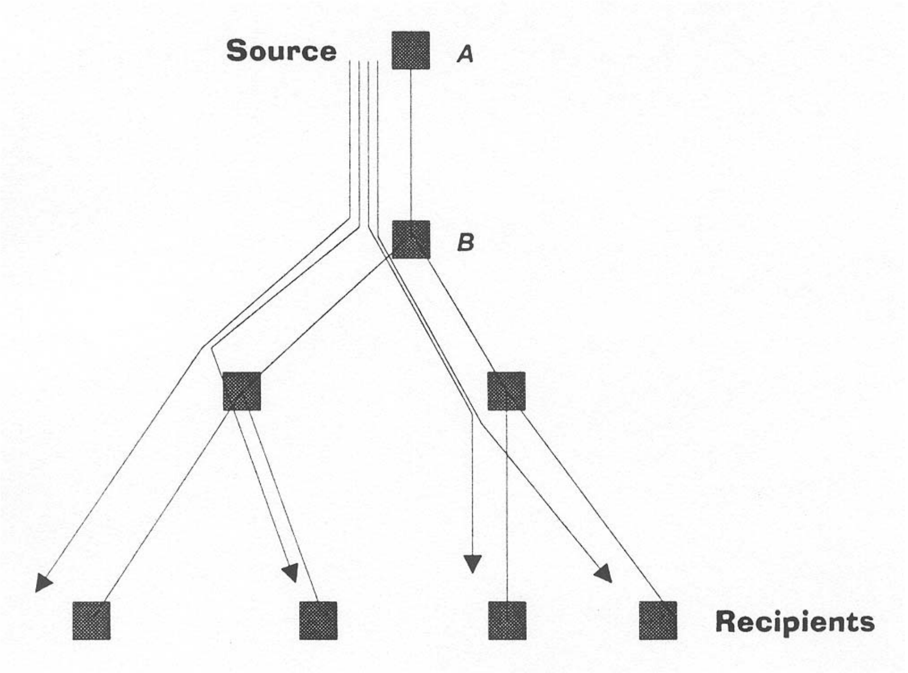
broadcast: all participants in the system are addressed

multicast: a selected *group* of participants is addressed

Efficiency of this addressing method depends on the implementation:

- if hardware multicast is available (e.g., LANs), a message is sent by one channel only once
- if not, significant savings can be obtained by using a tree structure (e.g., multicast-IP routing)

Example of a Multicast Tree



Distributed System Paradigms (6)

visible groups: idea is to support the functionality of the application (cooperative applications)

Examples:

- multicasting the name look-up request in a distributed name server to speedup name resolution
- video and audio diffusion to a group of registered recipients (video/audio conference)

invisible groups: idea is to realize *transparency* properties of the system (e.g. replication *transparency*)

group membership:

A group membership service provides two functions to its participants:

- providing primitives that explicitly allow to join and leave a group
- dynamically providing *group views*, i.e. updated information of the actual mutually reachable members

A group membership service has to deal with

- two fundamental properties
    - *accuracy:* the information provided reflects the real actual state of the group (liveness property)
    - *consistency:* the information provided is consistent to all group members (safety property)
- two main aspects:
    - *reliability aspects:* addressing message delivery guarantees
    - *ordering aspects:* addressing message ordering guarantees

group communication: allows group members to exchange information offering QoS properties

group platform: a protocol suite offering both group membership and group communication services

Distributed System Paradigms (7)

Multicast Protocol

responsible to deliver a message to all group members. Its main components are:

- routing
- omission tolerance
- flow control
- ordering w.r.t. messages as well as w.r.t. group views
- node failure recovery --> view changes

Link, *Network* and *Transport* layer of the network protocol stack*:* addressing the first three components

4. Time and Clocks

Time is a very useful artifact to represent the ordering, sequencing, synchronizing of events in any system. The passage of time is marked by an abstract monotonically increasing continuous function, called *real time*

Along history, time has been represented (measured) in different ways, mainly dependent on how the unit of time,called second, was measured.

timeline: graphical representation of time units as an infinite straight line

The use of time in computer systems addresses two aspects:

- observing and recording the place of events in a timeline (ordering, sequencing)
- enforcing the future positioning of events in the timeline (synchronizing)

Distributed System Paradigms (8)

local physical clock:

implements in hardware the mapping of real time t into a clock time pc(t), which is a monotonically increasing discrete function. They are based typically on oscillators such as quartz. The timeline now becomes a sequence of discrete ticks.

The “quality” (metric for imperfectness) of hardware clocks is mainly characterized by the parameters

- *granularity g:* time difference between two consecutive ticks t(i) and t(i+1): g:= pc(t(i+1)) - pc(t(i))
- *drift rate* ρ: positive constant denoting the drift of a physical clock from real time

  $\rho \approx 10^{-5}$, i.e. several microseconds per second, e.g. ca. 36 ms per hour, almost 1 s per day
- *clock rate*: : $1-\rho \leq (pc(t(i+1) - pc(t(i)))/g \leq 1+\rho$

local clocks can be used to

- represent a timer to set *timeouts*
- timestamp local events
- measure local durations

They cannot be used for timing analysis regarding global events in a distributed systems because of ρ

––> need to synchronize all local clocks by means of a *clock synchronization algorithm*

global clocks

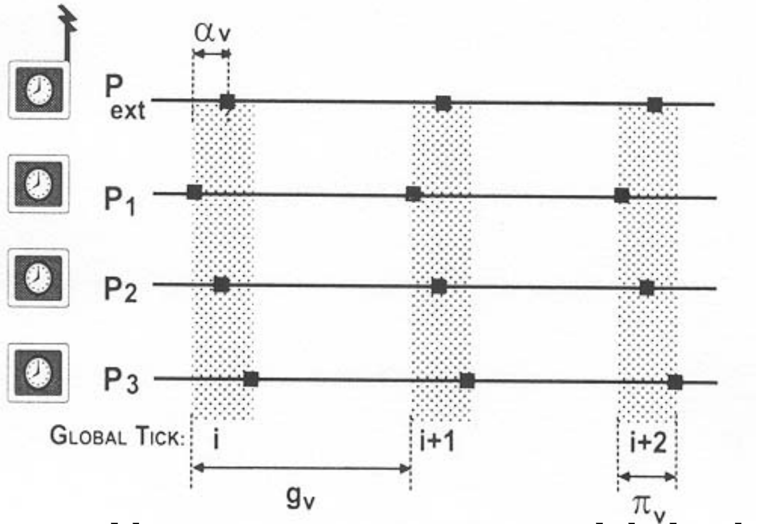
A global clock in a distributed system is built by synchronizing in periodic rounds all local clocks as close as possible to the same initial value.

Distributed System Paradigms (9)

virtual clock: the time vc(t) delivered by a synchronized local physical clock

The set of virtual clocks under the control of the synch. algorithm. constitutes the global clock of the system

Properties of a Global Clock:



granularity g_v denotes the time interval between two consecutive global ticks

precision π_v denotes the maximum deviation between two corresponding ticks of any two virtual clocks, as seen by an outside observer, measured by the external reference clock representing the real time.

$$\pi_v := \max\{\text{for all } i,k,l : |vc_k(t(i)) - vc_l(t(i))|\}$$

accuracy α_v denotes the maximum deviation between a tick of any of the virtual clocks and the corresponding tick of the external reference clock P_{ext}.

$$\alpha_v := \max\{\text{for all } i,k : |vc_k(t(i)) - P_{ext}(t(i))|\}$$

convergence δ_v denotes the maximum deviation between any two ticks of the virtual clocks immediately after the termination of a synchronization round.

$$\delta_v := \max\{\text{for all } k,l : |vc_k(t(0)) - vc_l(t(0))|\}$$

Distributed System Paradigms (10)

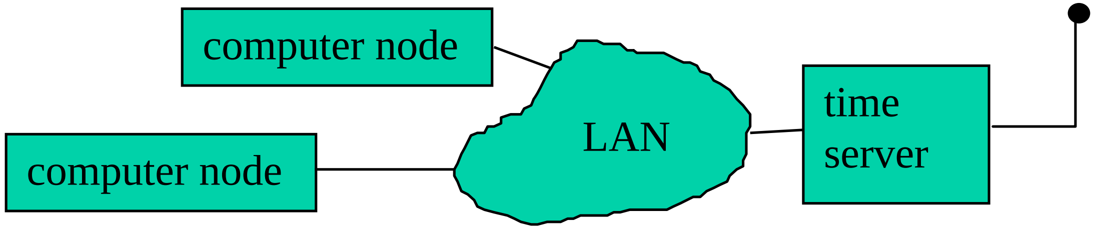
convergence δ is a measure for the quality of the internal clock synch. algorithm (internal synchronization)
accuracy α is a measure for the external synchronization, e.g. with GPS receiver time as reference time

The definitions above imply the following relations: $\pi \geq \delta$, $\pi \leq 2\alpha$, $g > \pi$
(precision cannot be better than convergence, it is at least twice the accuracy, it is senseless to select a granularity finer than the precision)

→ any globally visible event e is timestamped t(e) by different virtual clocks of the system with at most one tick difference
→ let d := $|t(e_1) - t(e_2)|$; if d < 2 --> no physical order of the events e_1 and e_2 can be deduced
→ granularity determines the resolution of the global time grid

Required components to define a global time basis:

- an external reference time, e.g. UTC-based
- local physical clocks
- a synchronization algorithm



Distributed System Paradigms (11)

Zeitzonen: Gebiete für die dieselbe Zeit festgelegt ist. 1884 wird die Welt in 24 Zeitzonen aufgeteilt. Die Zeitzonen unterscheiden sich von UT (GMT) ganzzahlig um jeweils 1 Stunde

.

UT (AT, 1833) Universal Time (UT) Mittlere Sonnenzeit, gemessen am Greenwich 0-Meridian (GMT). Basiert auf der mittleren Länge eines Sonnentags, d.h. auf der Erdrotation

ET (AT, 1955) Ephimeridenzeit (ET), basiert auf der Umlaufzeit der Erde um die Sonne. Harold Spencer Jones stellte 1939 fest, daß die Rotation der Erde variiert, die Umlaufzeit um die Sonne nicht. 1 Sekunde der ET wird festgelegt als der 1/31.566.925,9747 Teil des tropische Jahres, das am Mittag des 1. Januars 1900 begann. (Tropisches Jahr: Periode zwischen zwei aufeinanderfolgenden Umläufen der Sonne durch den Himmelsäquator in derselben Richtung.)

UT2 (AT, 1960) Zeit, basierend auf und gemittelt über den lokalen Beobachtungen verschiedener über die Erde verteilter Observatorien und anschließend nochmals auf empirischer Basis korrigiert

TAI (PT, 1961) Temps Atomique International (TAI) basiert auf mehreren koordinierten Cäsium-Uhren. Fortlaufende Zeitzählung, beginnend mit dem 1.Januar 1958 0 Uhr UT2-Zeit (daher konsistent mit UT2). 1 Sekunde der TAI ist 9 192 631 770 mal die Periode der Strahlung des Atoms Cäsium 133.

UTC (PT, 1972) Universal Time Coordinated (UTC) basiert auf TAI, wird aber ständig an UT2 angepaßt. Immer wenn UTC und UT2 mehr als 800 ms auseinander gedrifted sind, wird eine “Schaltsekunde” eingefügt. UTC beginnt am 1. Januar 1972. Seit dieser Zeit sind (bis 1992) 15 Schaltsekunden eingefügt worden. UTC ist damit eine an AT angepaßte physikalische Zeit. $\rho \approx 10^{-14}$, d.h. Abweichung ca. 1 Sek / 300000 Jahre

Distributed System Paradigms (11a)

GPS (Global Positioning System):

- network of 21 satellites covering earth surface
- equipped with cesium atomic clocks
- GPS-receiver clocks mostly provide UTC with an accuracy of $\alpha_g \leq 100$ns
- GPS receiver antenna must be placed externally (being under the light cone of the satellites)