Diplomski studij

Informacijska i komunikacijska tehnologija:

Telekomunikacije i informatika

Računarstvo:

Programsko inženjerstvo i informacijski sustavi

Računarska znanost

Ak.god. 2009./2010.

Raspodijeljeni sustavi

9.

Otpornost na neispravnosti

17.11.2009.

Sadržaj predavanja

- Uvod, definicija pojmova
- Otpornost procesa na ispade
- Sporazum grupe procesa
- Pouzdana komunikacija grupe procesa
- Raspodijeljeno izvršavanje operacije
- Oporavak nakon ispada

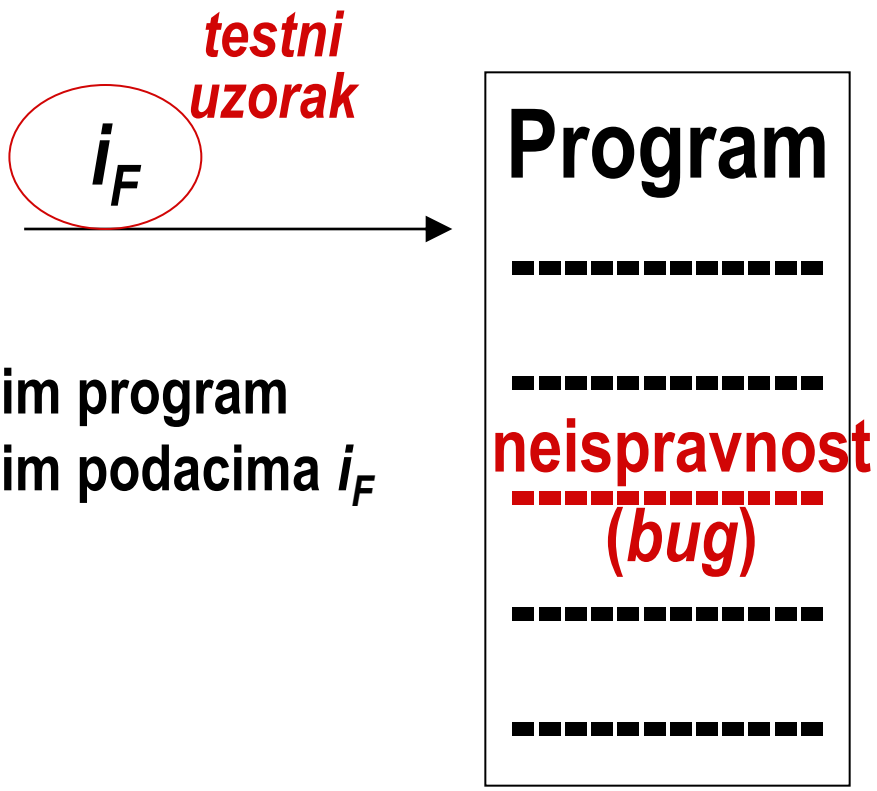
Otpornost na neispravnosti

Fault tolerance (*provide service even in the presence of faults*)

- sposobnost sustava za obavljanje definirane usluge bez obzira na postojeće **neispravnosti** koje izazivaju **ispad** nekih komponenti sustava
- osobina sustava, može prikriti ispad komponenti raspodijeljenog sustava, definiraju se procedure za oporavak sustava
- funkcionalnost sustava može biti ograničena uz negativan utjecaj na njegove performanse

Lamport: "*A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable*."

Podsjetimo se izvođenja programskog koda…



izvodim program ulaznim podacima i_F

detektira se *ispad* programa jer rezultat obrade o_F nije jednak očekivanom izlazu definiranom specifikacijom

$o_F \neq Spec(i_F)$

Osnovni pojmovi

- Ispad sustava (*failure*)
  - stanje sustava koje se detektira kroz nemogućnost korištenja jedne ili više njegovih usluga
  - posljedica neispravnosti, signalizira postojanje neispravnosti u raspodijeljenom sustavu
- Neispravnost (*fault*)
  - npr. dio programskog koda (*bug*), neispravan komunikacijski kanal, pogreške prilikom oblikovanja sustava
  - uzrok ispada sustava, pronalaženje neispravnosti je težak i važan problem
  - prolazne, isprekidane i trajne neispravnosti

Vrste ispada

- Ispad procesa
  - proces ne mijenja stanja (ne izvode se prijelazi) premda se ne nalazi u završnom stanju (*stoping failure*)
  - proces generira proizvoljne izlaze (*Byzantine failure*)
- Ispad kanala
  - proces *p* je poslao poruku procesu *q*, ali *q* poruku ne prima, jer npr. kanal gubi poruke

Vrste ispada procesa

Vrsta ispada	**Opis**
Ispad procesa	Proces neočekivano ulazi u stanje zaustavljanja i ne odgovara na nove zahtjeve.
Pogreška u komunikaciji *pogreška primanja* *pogreška slanja*	Proces ne odgovara na primljeni zahtjev. Proces ne prima zahtjev. Proces ne šalje odgovor.
Vremenska pogreška	Proces šalje odgovor nakon isteka vremenskog roka.
Pogrešan odgovor *sadržaj* *pogrešna promjena stanja poslužitelja*	Generirani odgovor je neispravan. Sadržaj odgovora je neispravan. Poslužitelj ulazi u pogrešno stanje nakon primljenog zahtjeva.
"Bizantska pogreška"	Proces proizvodi proizvoljan odgovor u proizvoljnom trenutku.

Redundancija

Ključna tehnika za prikrivanje neispravnosti raspodijeljenog sustava.

Primjeri redundancije:

- redundancija informacija
  - npr. Hammingov kod
- vremenska redundancija
  - ponavljanje neke operacije u vremenu
- fizička redundancija
  - dodavanje dodatne opreme (npr. vruća rezerva) ili procesa u sustav (repliciranje procesa)

Sadržaj predavanja

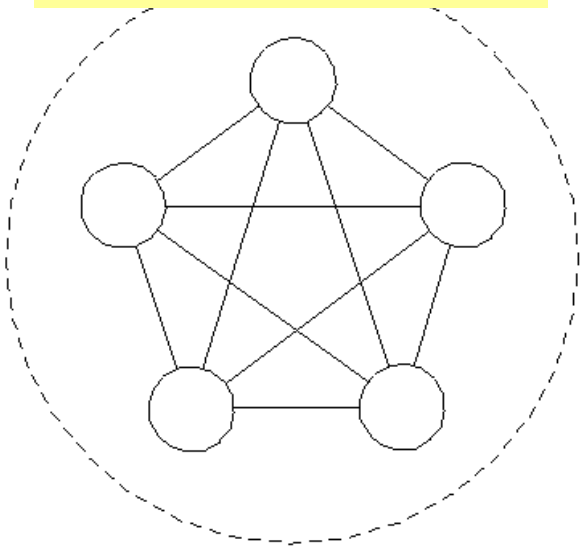
- Uvod, definicija pojmova
- Otpornost procesa na ispade
- Sporazum grupe procesa
- Pouzdana komunikacija grupe procesa
- Raspodijeljeno izvršavanje operacije
- Oporavak nakon ispada

Otpornost procesa na ispade

- Zamjena jednog procesa grupom identičnih procesa
  - replikacija procesa radi prikrivanja ispada jednog ili *k* procesa
- **tolerancija *k* ispada**
  - grupa može "preživjeti" ispad najviše *k* procesa
- dovoljan je ***k* + 1** proces da se osigura tolerancija na *k* ispada, jedan proces može preuzeti poslove grupe
- ako pretpostavimo *k* bizantskih ispada, potrebno je **2*k* + 1** procesa u grupi (*k* + 1 proces će "nadglasati" *k* neispravnih)

Organizacija grupe procesa (1/2)

Grupa s ravnopravnim procesima

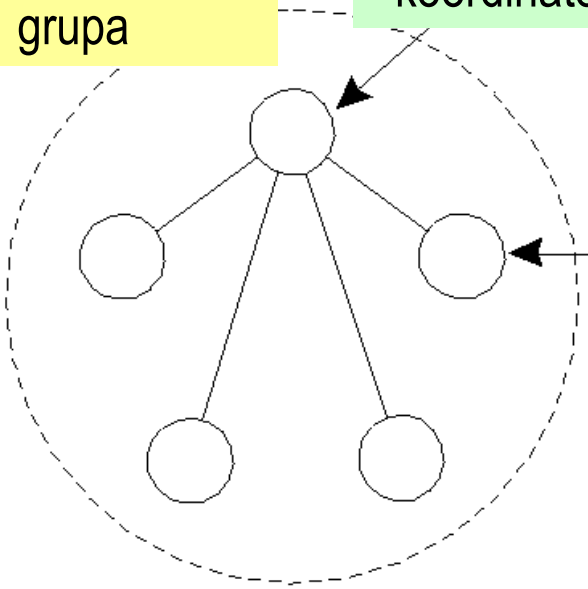


(a)

Hijerarhijska grupa

koordinator

“radnik”



(b)

Organizacija grupe procesa (2/2)

- Grupe procesa su dinamične (slično sustavima P2P)
- Jedan proces može biti član više grupa
- Usluga za komunikaciju grupe procesa
  - osigurava isporuku jedne poruke iz okoline svim članovima grupe
- Administracija grupe procesa
  - usluga koja omogućuje kreiranje grupe, dodavanje procesa u grupu i izlazak procesa iz grupe

Otkrivanje ispada procesa iz grupe

- svaki proces periodički šalje upit ostalim procesima i provjerava njihovo stanje (“are u alive?”)
- proces pasivno čeka i prati poruke koje prima od ostalih članova grupe

Sadržaj predavanja

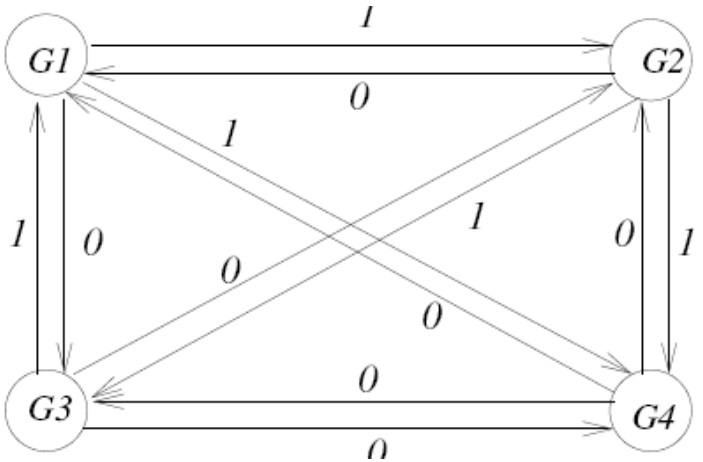
- Uvod, definicija pojmova
- Otpornost procesa na ispade
- Sporazum grupe procesa
- Pouzdana komunikacija grupe procesa
- Raspodijeljeno izvršavanje operacije
- Oporavak

Sporazum grupe procesa

- Procesi trebaju postići sporazum o npr. vrijednosti neke varijable ili o tome hoće li ili neće izvršiti transakciju

Primjer

- 4 generala trebaju započeti napad koordinirano u isto vrijeme
- za komunikaciju koriste glasnike
- komunikacija je nepouzdana i može trajati prilično dugo
- general izdajnik bi mogao slati pogrešne informacije ostalima



Mogu li generali postići sporazum o vremenu napada?

Problem bizantskog sporazuma (1/2)

Byzantine agreement problem [Lamport et al. 1982]

- **Problem**: svaki proces definira inicijalnu vrijednost, a grupa procesa treba postići sporazum o nizu vrijednosti za svake procese iz grupe
- **Sporazum**: Svi ispravni procesi prihvaćaju isti niz vrijednosti [v_1, v_2, … v_n].
- **Ispravnost**: Ako je proces *i* ispravan i definira vrijednost v_i, svi ostali ispravni procesi prihvaćaju vrijednost v_i kao i-ti element niza. Ako je proces neispravan, ostali ispravni procesi mogu prihvatiti bilo koju vrijednost za taj proces.
- **Završetak**: Svaki ispravan proces će u konačnici prihvatiti vrijednosti niza.

Problem bizantskog sporazuma (2/2)

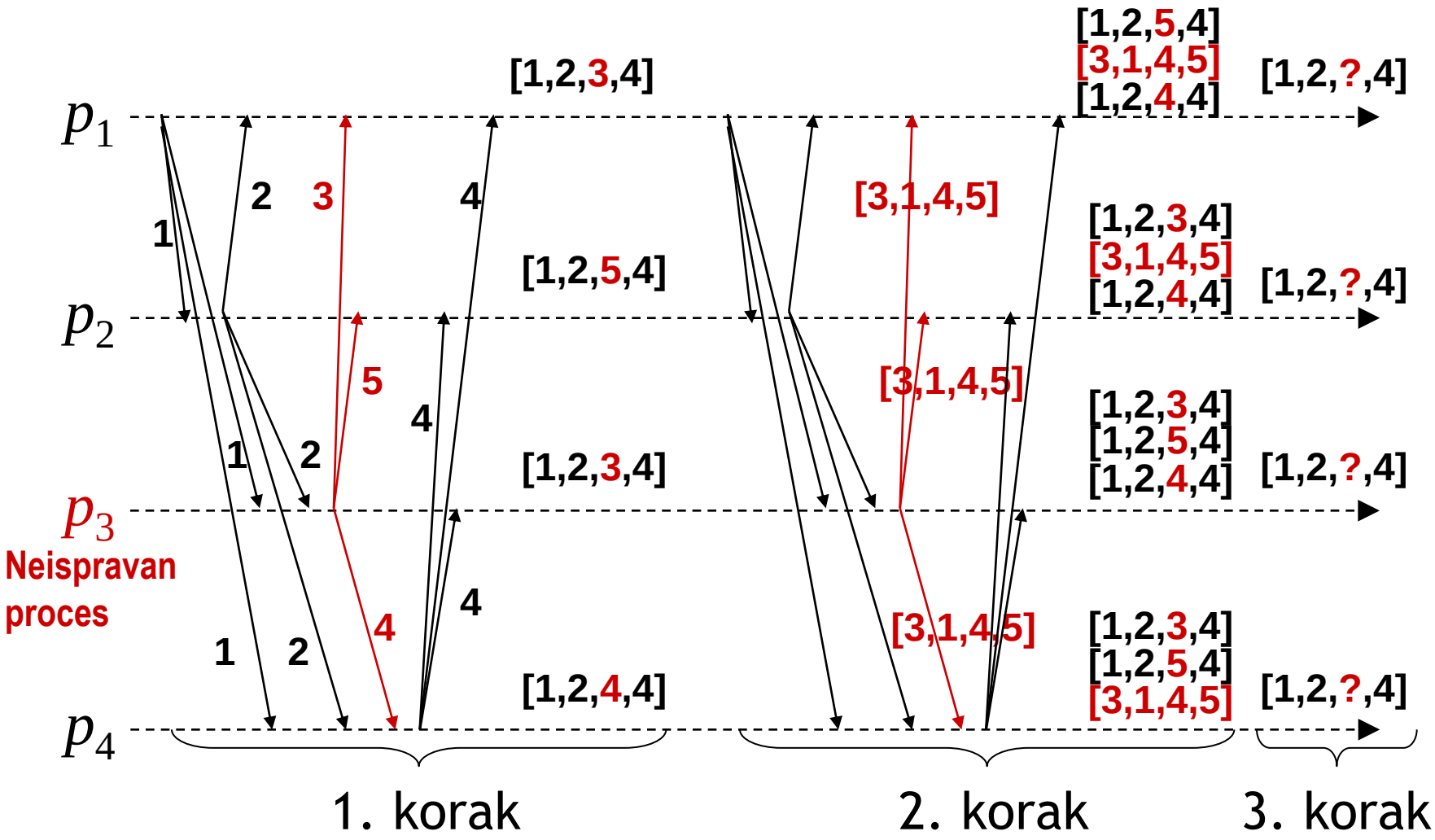
- Pretpostavke:
  - *n* procesa
  - *k* neispravnih procesa (**bizantski ispad**)
  - proces *i* šalje vrijednost v_i ostalim procesima u grupi
  - svaki proces treba kreirati niz **V** duljine *n* takav da vrijedi **V**[i] = v_i

Skica algoritma

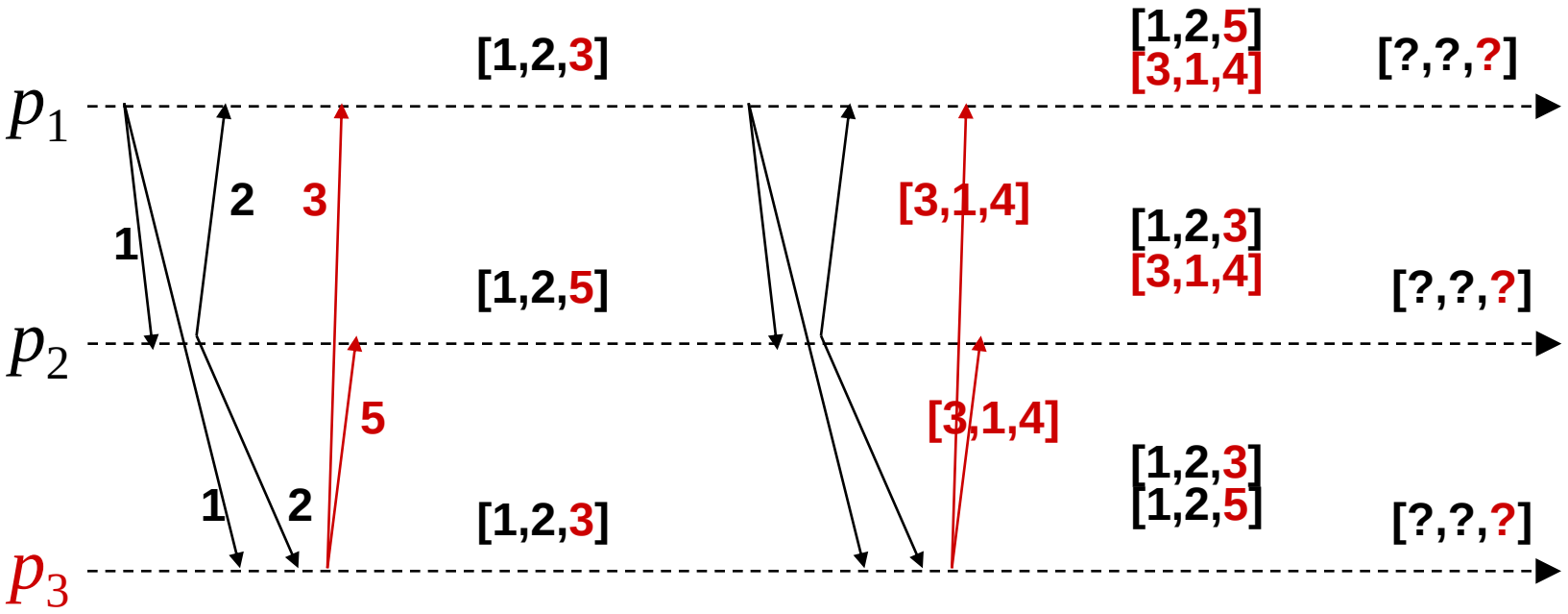
- **1. korak:**
  Ispravni procesi šalju svoj v_i ostalim procesima, dok neispravni procesi šalju proizvoljne vrijednosti
- **2. korak:**
  Svi procesi prikupljaju vrijednosti i spremaju ih u **V**.
- **3. korak:**
  Svaki proces šalje ostalima svoj **V**.
  Konačno svaki proces uspoređuje sve primljene nizove i odlučuje se za većinske vrijednosti.

- Procesi mogu donijeti odluku samo o vrijednostima za ispravne procese!

Primjer (n=4, k=1)



Primjer (n=3, k=1)



Mogu li procesi donijeti odluku za prikazani slučaj?

Zaključak

- Rješenje za sporazum u slučaju *n* procesa i *k* neispravnih procesa u bizantskom ispadu može se riješiti samo u slučaju **sinkronog sustava** i kada vrijedi

$$k \leq \left\lfloor \frac{n-1}{3} \right\rfloor$$

- Dokazano je da u slučaju asinkronog sustava procesi ne mogu postići sporazum niti za *k*=1

Za sporazum u slučaju *k* neispravnih procesa, potrebno je (2*k* + 1) ispravnih, tj. ukupno *n* = 3*k* + 1 procesa!

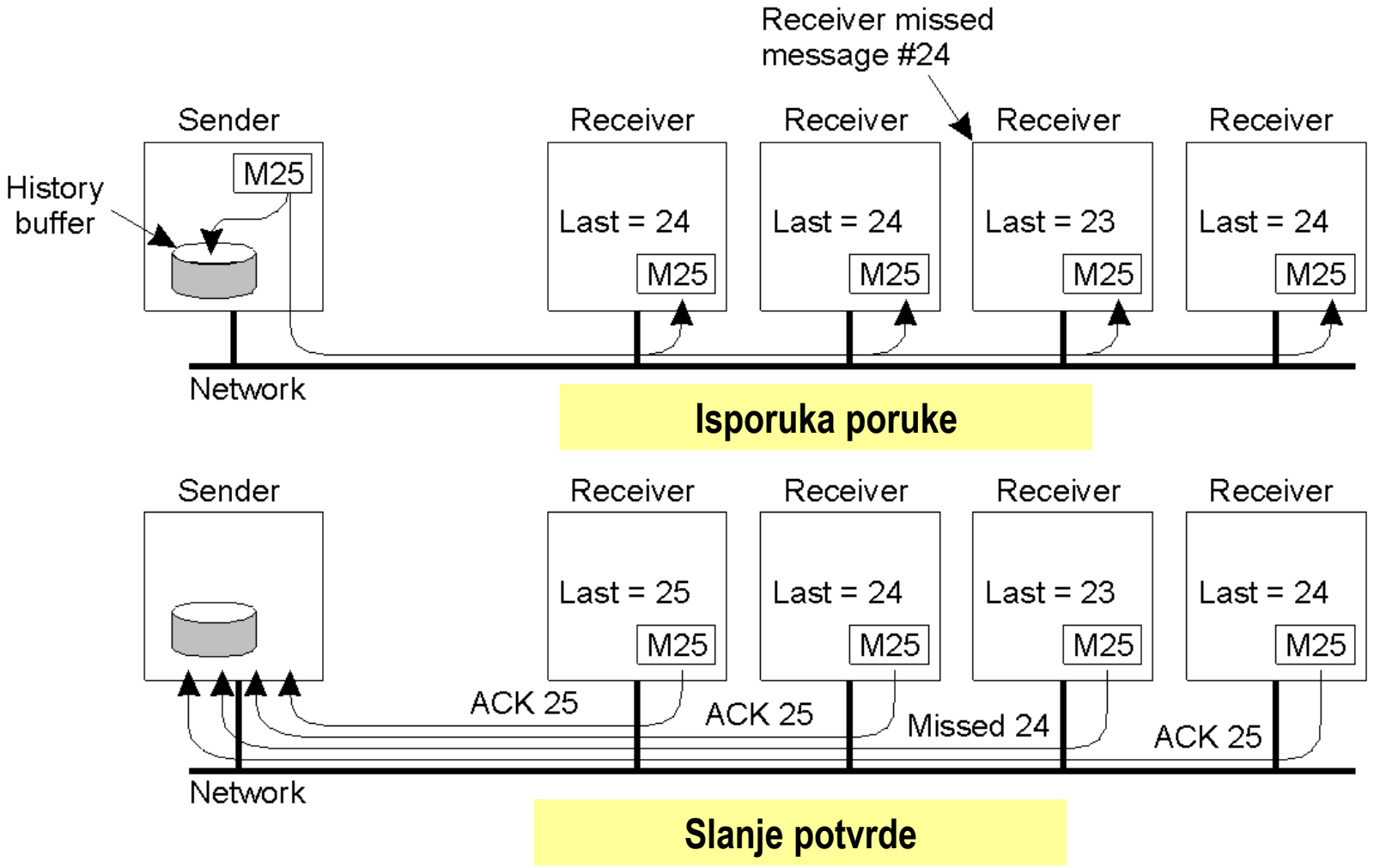
Sadržaj predavanja

- Uvod, definicija pojmova
- Otpornost procesa na ispade
- Sporazum grupe procesa
- Pouzdana komunikacija grupe procesa
- Raspodijeljeno izvršavanje operacije
- Oporavak nakon ispada

Pouzdana komunikacija grupe procesa

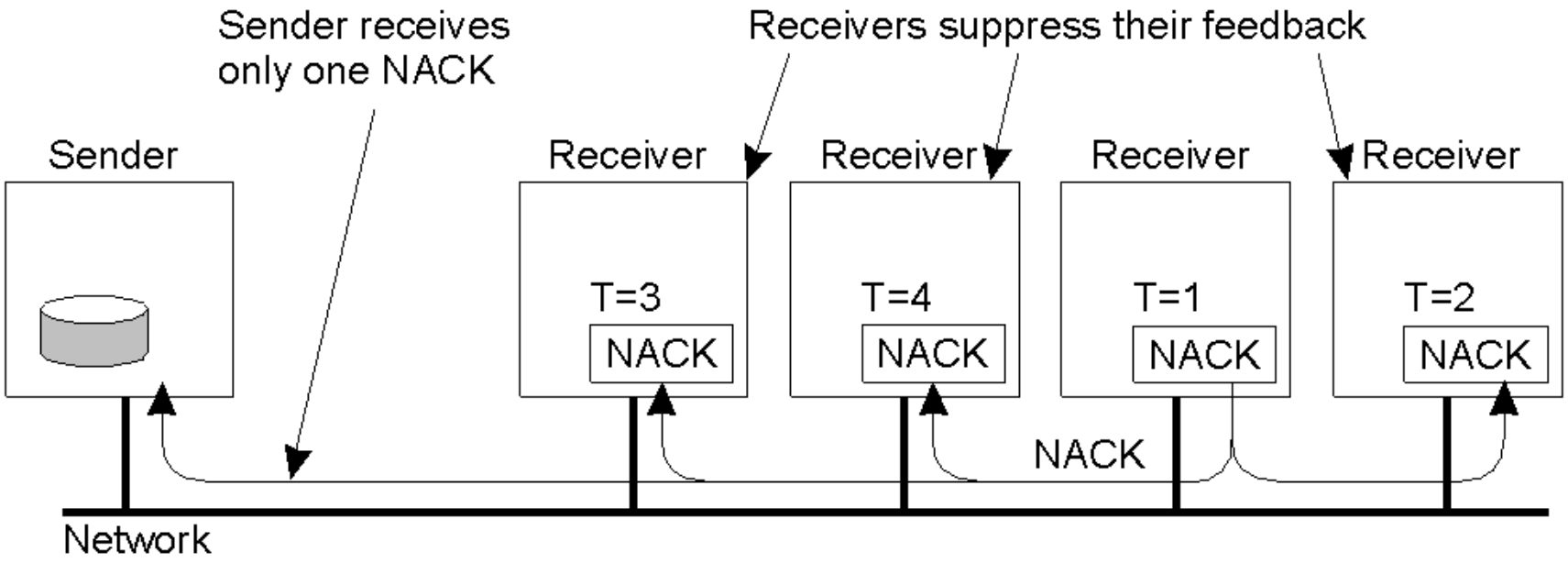
- garantira isporuku poruke svim procesima u grupi
- problemi
  - Koji procesi čine grupu u trenutku slanja poruke?
  - Što se događa ako novi proces ulazi u grupu procesa dok je isporuka poruke grupi u tijeku?
  - Što se događa ako dođe do ispada pošiljatelja poruke tijekom isporuke poruke ostalim procesima?
  - Što se događa ako jedan od primatelja ispadne tijekom isporuke poruke?
- najjednostavnija praktična implementacija
  - pouzdana komunikacija od točke do točke (npr. TCP) između svakog para procesa iz grupe

Pouzdana komunikacija bez mogućih ispada procesa (1/2)



Pouzdana komunikacija bez mogućih ispada procesa (2/2)

Potisnuta potvrda



Pouzdana komunikacija s ispadima procesa

- garantira isporuku poruke svim ispravnim i dostupnim procesima u grupi ili niti jednom
- potrebno je osigurati i isporuku poruka u određenom slijedu

Notacija

- p - proces
- G – grupa, skup procesa (*group view*)
- m – generirana poruka
- vc – poruka koja prenosi informaciju o dolasku ili odlasku procesa iz grupe (*view change*)

Virtualna sinkronost

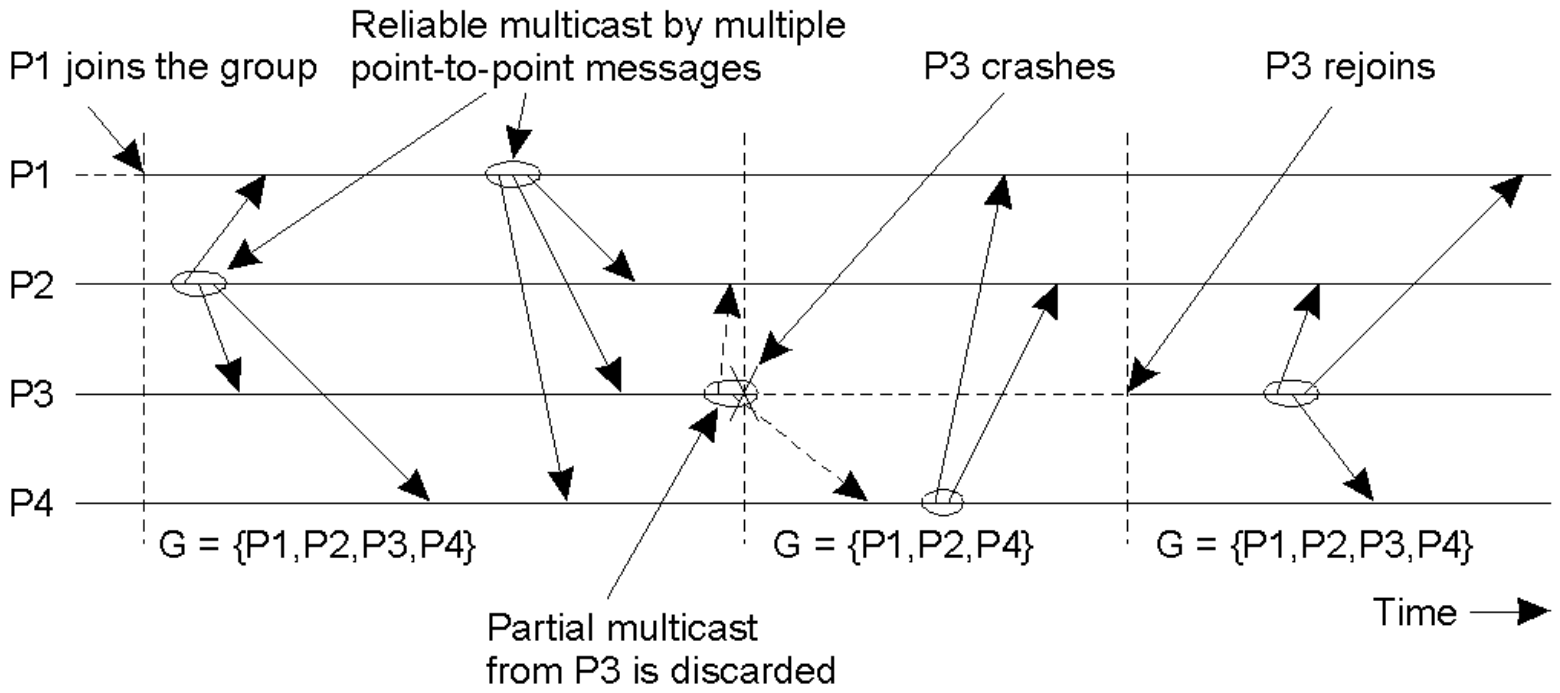
Scenarij 1

- Proces p šalje poruku m, u tome trenutku postoji grupa procesa G
- Tijekom isporuke m novi proces se uključuje u grupu i generira se poruka vc (view change) koja se opet šalje svim G
- Posljedica: p i vc su istovremeno u tranzitu
- 2 moguća rješenja
  - m isporučen svim članovima G prije isporuke vc
  - m nije isporučen niti jednom procesu iz G

Scenarij 2

- Ako p pošalje m i ispadne prije isporuke m članovima iz G, ostali procesi ignoriraju m (ne treba osigurati isporuku ostalim ispravnim procesima, kao da je p ispao prije slanja m)

Primjer virtualne sinkronosti



Poruka se može isporučiti članovima iz G samo ako ne postoji poruka vc koja je istovremeno u tranzitu.

Implementacija nije trivijalna.

Slijed poruka

- slijed kojim procesi primaju poruke od velike je važnosti jer utječe na promjene stanja tih procesa
- slijed primljenih poruka može biti:
  - neuređen (*unordered multicast*)
  - FIFO (FIFO-*ordered multicast*)
  - potpuno uređen (*totally-ordered multicast*)

Neuređeni slijed poruka

Process P1	Process P2	Process P3
sends m1	receives m1	receives m2
sends m2	receives m2	receives m1

Nije važan redoslijed kojim procesi primaju poruke.

Pouzdani neuređeni multicast – to je pouzdani multicast koji je istovremeno i **virtualno sinkron**.

Slijed poruka FIFO

Process P1	Process P2	Process P3	Process P4
sends m1	receives m1	receives m3	sends m3
sends m2	receives m3	receives m1	sends m4
	receives m2	receives m2	
	receives m4	receives m4	

Poruke koje dolaze od istog procesa moraju se isporučiti u redoslijedu kojim su poslane, npr. kako P1 šalje prvo m1 a zatim m2, m1 mora biti isporučen prije m2. Isto vrijedi za m3 i m4. Procesi P2 i P3 primaju poruke u različitom slijedu, važno je poštivati redoslijed isporuke poruka koje dolaze od istog procesa.

Pouzdani FIFO multicast

Potpuno uređen slijed poruka

Process P1	Process P2	Process P3	Process P4
sends m1	receives m1	receives m1	sends m3
sends m2	receives m3	receives m3	sends m4
	receives m2	receives m2	
	receives m4	receives m4	

Poruke se isporučuju u istom redoslijedu svim procesima u grupi (treba uzeti u obzir i FIFO redoslijed poruka koje dolaze od istog procesa, što znači da npr. m2 ne može biti isporučen prije m1).

Atomic multicast – pouzdani virtualno sinkroni multicast s potpuno uređenim slijedom poruka

Sadržaj predavanja

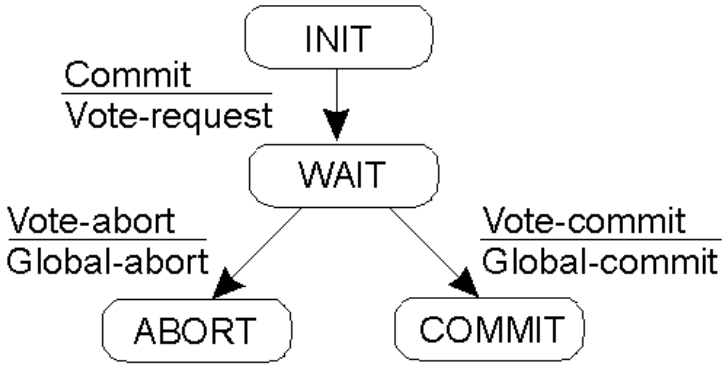
- Uvod, definicija pojmova
- Otpornost procesa na ispade
- Sporazum grupe procesa
- Pouzdana komunikacija grupe procesa
- Raspodijeljeno izvršavanje operacije
- Oporavak nakon ispada

Raspodijeljeno izvršavanje operacije

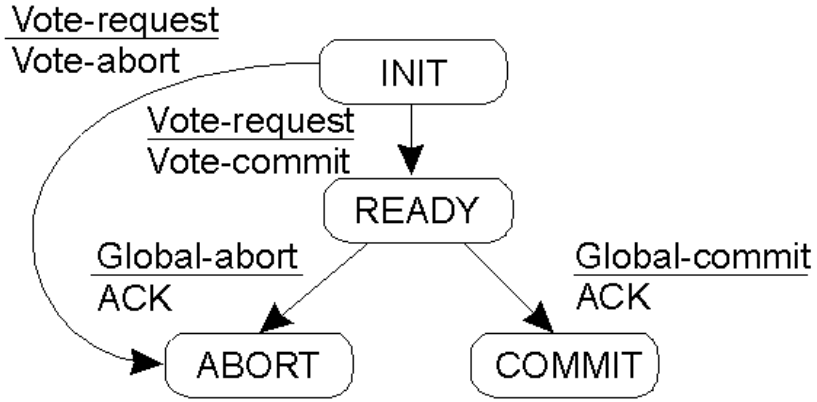
Distributed commit

- operaciju izvodi svaki proces u grupi ili niti jedan
- ako je operacija isporuka poruke, riječ je o atomic multicast
- česta primjena kod raspodijeljenih transakcija
- ideja rješenja (*one-phase commit*)
  - postoji koordinator u grupi procesa
  - koordinator šalje zahtjev ostalim procesima za (lokalno) izvršavanje operacije
  - nedostaci: procesi ne mogu obavijestiti koordinatora u slučaju nemogućnosti izvršavanja operacije, ispad koordinatora

Protokol *two-phase commit* (1/3)



Automat stanja koordinatora



Automat stanja procesa

Protokol *two-phase commit* (2/3)

Problemi

- Postoje stanja blokiranja na strani koordinatora i procesa
  - Proces je blokiran u stanju INIT kada čeka VOTE_REQUEST – ako ne primi poruku nakon određenog vremena, proces može lokalno odustati od izvršavanja operacije
  - Koordinator je blokiran u stanju WAIT kada čeka odgovore svih procesa – ako nakon nekog perioda ne primi odgovor od svih procesa, koordinator može zaključiti da treba odustati od izvršavanja operacije
  - Proces je blokiran u stanju READY čekajući konačnu odluku koordinatora – proces treba saznati koju je poruku koordinator poslao!

Protokol *two-phase commit* (3/3)

Stanje procesa Q	Akcije procesa P
COMMIT	Obavi prijelaz u COMMIT
ABORT	Obavi prijelaz u ABORT
INIT	Obavi prijelaz u ABORT
READY	Kontaktiraj drugi proces

Akcije procesa P kada se nalazi blokiran u stanju READY i kada kontaktira drugi proces Q iz G

Algoritam za koordinatora


```
while START _2PC to local log;
multicast VOTE_REQUEST to all participants;
while not all votes have been collected {
    wait for any incoming vote;
    if timeout {
        while GLOBAL_ABORT to local log;
        multicast  GLOBAL_ABORT to all participants;
        exit;
    }
    record vote;
}
if all participants sent VOTE_COMMIT and coordinator votes COMMIT{
    write GLOBAL_COMMIT to local log;
    multicast GLOBAL_COMMIT to all participants;
} else {
    write GLOBAL_ABORT  to local log;
    multicast GLOBAL_ABORT to all participants;
}
```

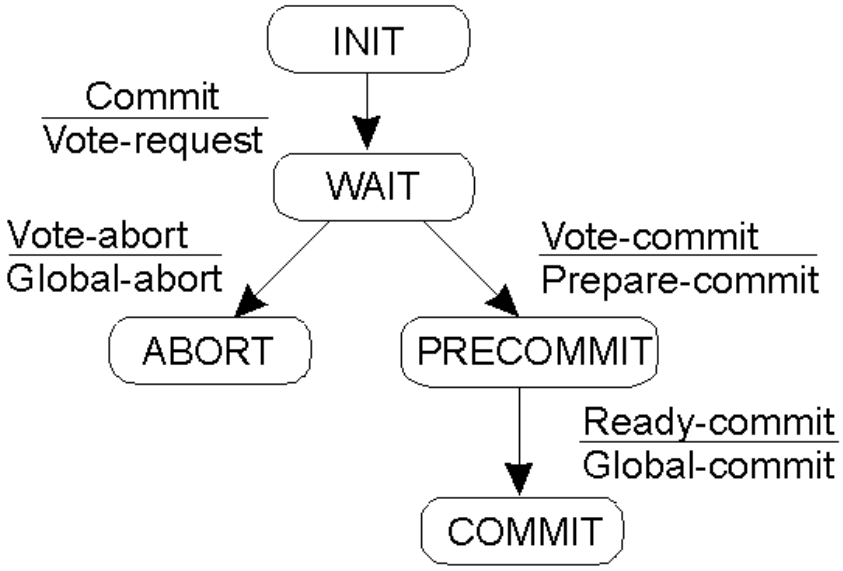
Algoritam za proces


```
write INIT to local log;
wait for VOTE_REQUEST from coordinator;
if timeout {
    write VOTE_ABORT to local log;
    exit;
}
if participant votes COMMIT {
    write VOTE_COMMIT to local log;
    send VOTE_COMMIT to coordinator;
    wait for DECISION from coordinator;
    if timeout {
        multicast DECISION_REQUEST to other participants;
        wait until DECISION is received; /* remain blocked */
        write DECISION to local log;
    }
    if DECISION == GLOBAL_COMMIT
        write GLOBAL_COMMIT to local log;
    else if DECISION == GLOBAL_ABORT
        write GLOBAL_ABORT to local log;
} else {
    write VOTE_ABORT to local log;
    send  VOTE ABORT to coordinator;
}
```

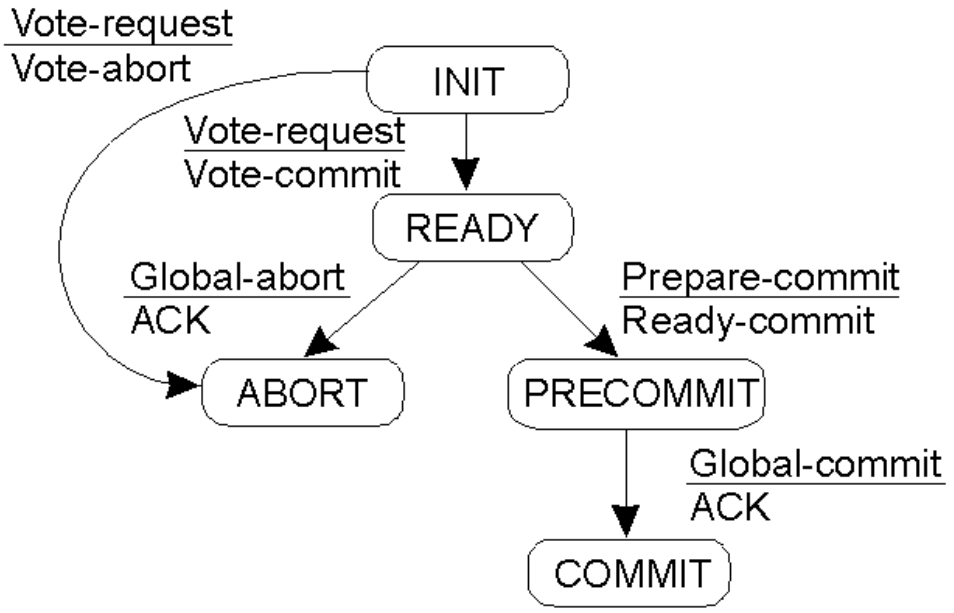
Protokol *three-phase commit* (1/3)

- rješava problema blokiranja procesa u slučaju ispada koordinatora
- relativno se slabo koristi u praksi jer se situacija koja dovodi do stanja blokiranja prethodnog protokola događa iznimno rijetko

Protokol *three-phase commit* (2/3)



Automat stanja koordinatora



Automat stanja procesa

Protokol *three-phase commit* (3/3)

- Koordinator može biti blokiran u stanju PRECOMMIT zbog ispada jednog procesa, ali može ostalim procesima poslati GOBAL_COMMIT
- Proces može biti blokiran u stanjima READY i PRECOMMIT
  - nakon isteka vremenske kontrole zaključuje da je došlo do ispada koordinatora i kontaktira ostale procese
  - ako je Q u stanju COMMIT i P prelazi u COMMIT
  - ako je Q u stanju ABORT i P prelazi u ABORT
  - ako su svi procesi (ili većina) u stanju PRECOMMIT, mogu svi preći u COMMIT
  - ako je Q u INIT, P prelazi u ABORT
  - ako su svi procesi u stanju READY, mogu svi preći u ABORT

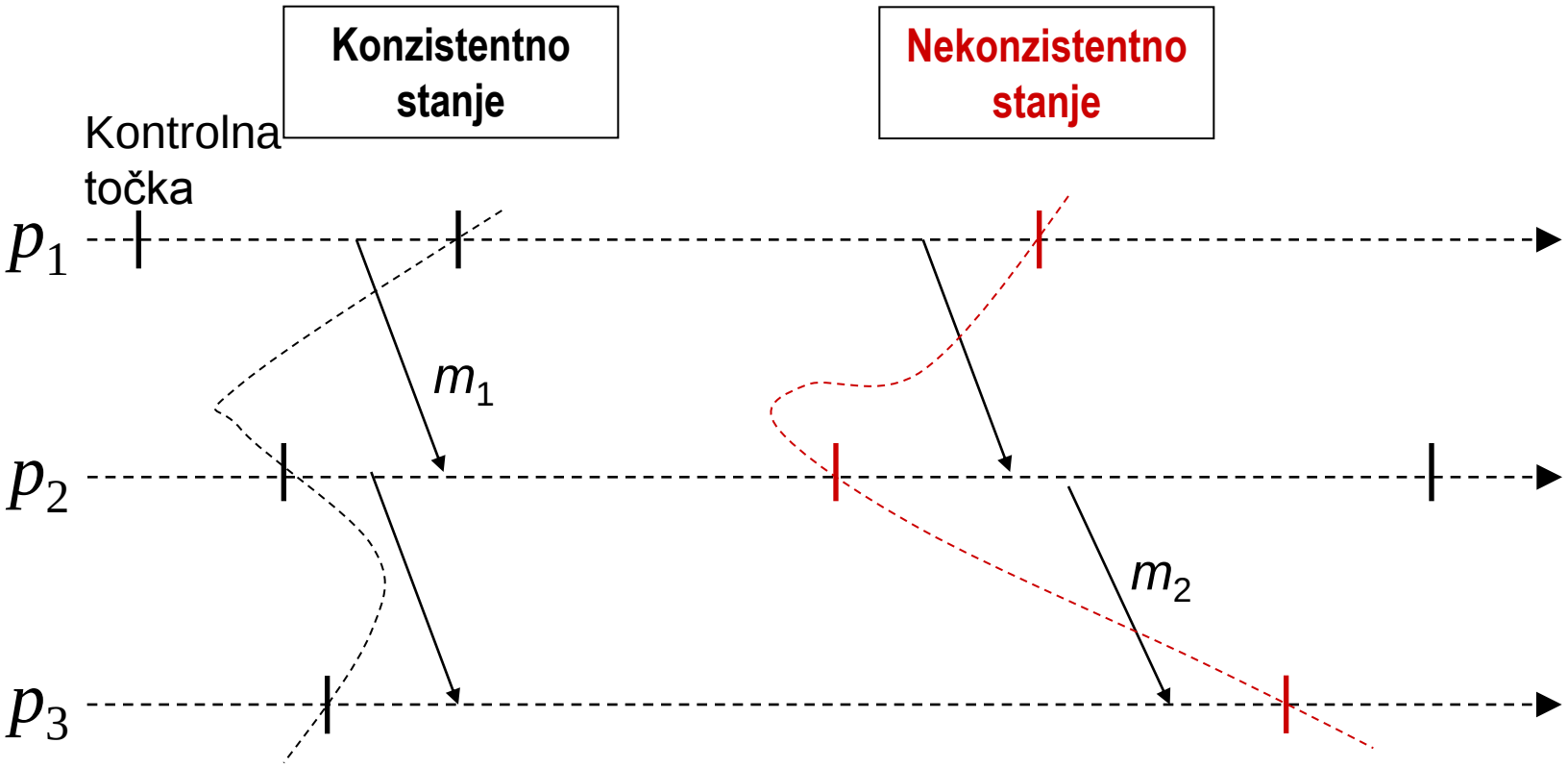
Sadržaj predavanja

- Uvod, definicija pojmova
- Otpornost procesa na ispade
- Sporazum grupe procesa
- Pouzdana komunikacija grupe procesa
- Raspodijeljeno izvršavanje operacije
- Oporavak nakon ispada

Oporavak nakon ispada

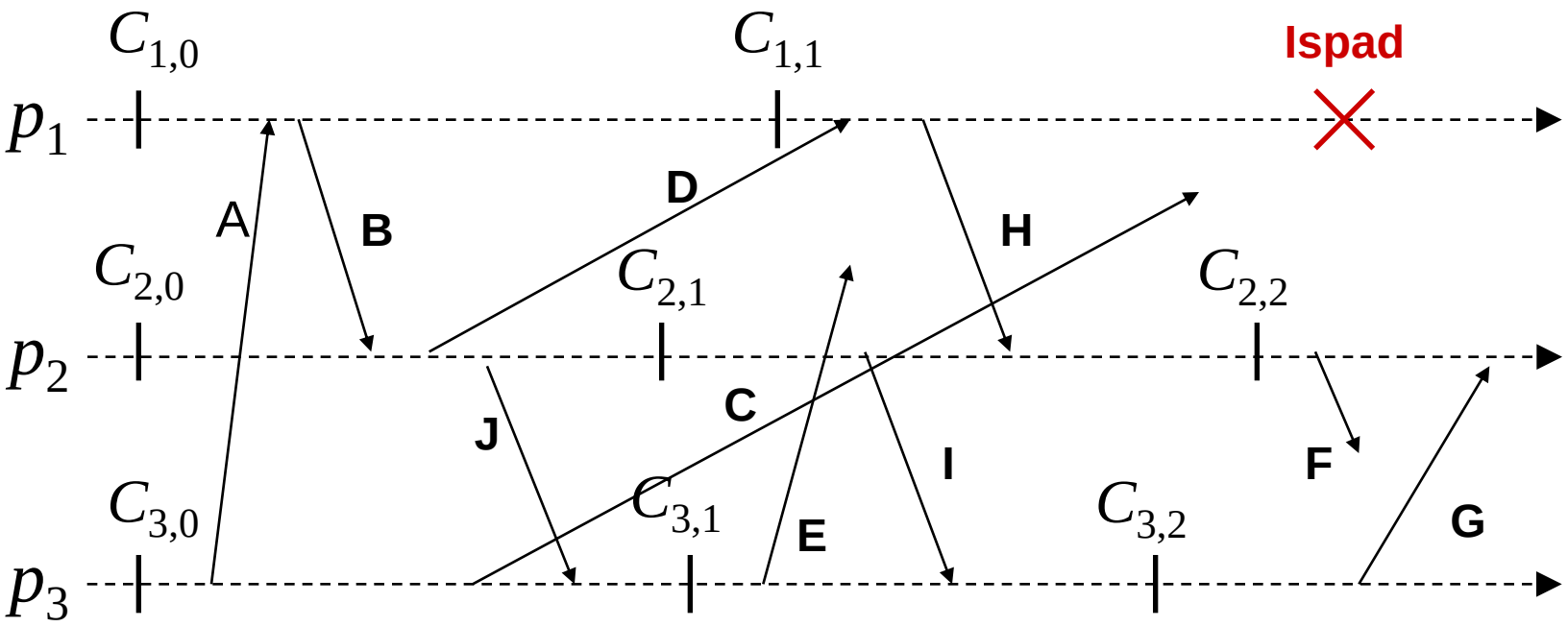
- Nakon ispada procesa nužan je njegov oporavak i povratak u ispravno stanje
- Oporavak unazad
  - vratiti sustav u ispravno stanje u prošlosti
  - potrebno je s vremena na vrijeme pohraniti stanje sustava (kontrolne točke, *checkpoint*)
  - zapisuju se poslane i primljene poruke u dnevnički zapis (*log*)
- Oporavak korištenjem dnevničkog zapisa
  - proces u ispadu vraća se u prethodno ispravno stanje nakon čega izvodi akcije iz dnevničkog zapisa

Konzistentna i nekonzistentna stanja



Procesi bilježe kontrolne točke neovisno jedan o drugom

Primjer oporavka povratkom unazad



Konzistentno stanje sustava:

$$C_{1,1}, C_{2,1}, C_{3,1}$$

A, B, J: normalne poruke

G, H, I: izgubljene poruke

C: odgođena poruka

D: izgubljena poruka

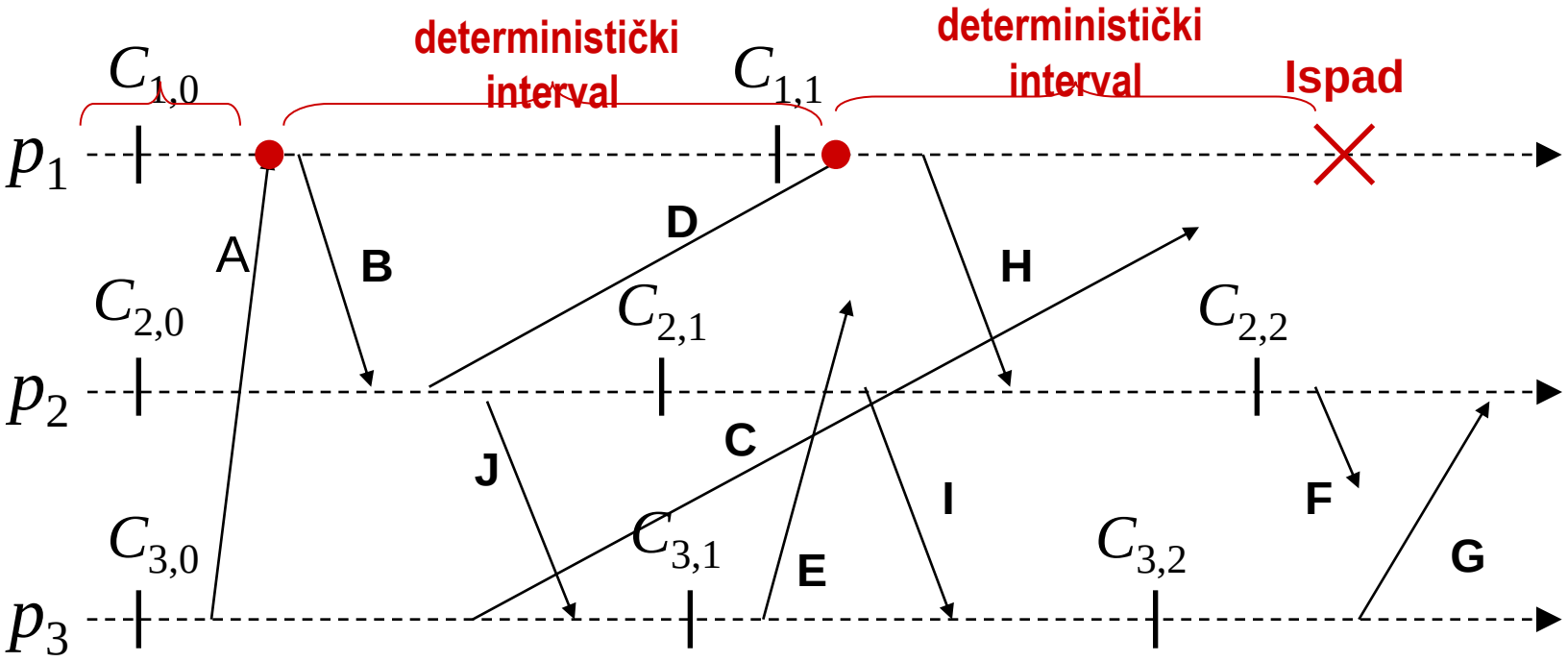
E i F: odgođene suvišne poruke

Bilježenje kontrolnih točaka

- Procesi mogu kontrolne točke bilježiti neovisno ili koordinirano
- **neovisno** zahtjeva bilježenje ovisnosti među procesima koji razmjenjuju poruku (šalju i primaju istu poruku) radi povratka u konzistentna stanja
- **koordinirano** koristi koordinatora i jednostavnije je za implementaciju, koordinator šalje naredbu procesima da pohrane stanje sustava gotovo istovremeno (prije toga moraju primiti poruke u tranzitu i privremeno zaustaviti pripremljena slanja poruka)

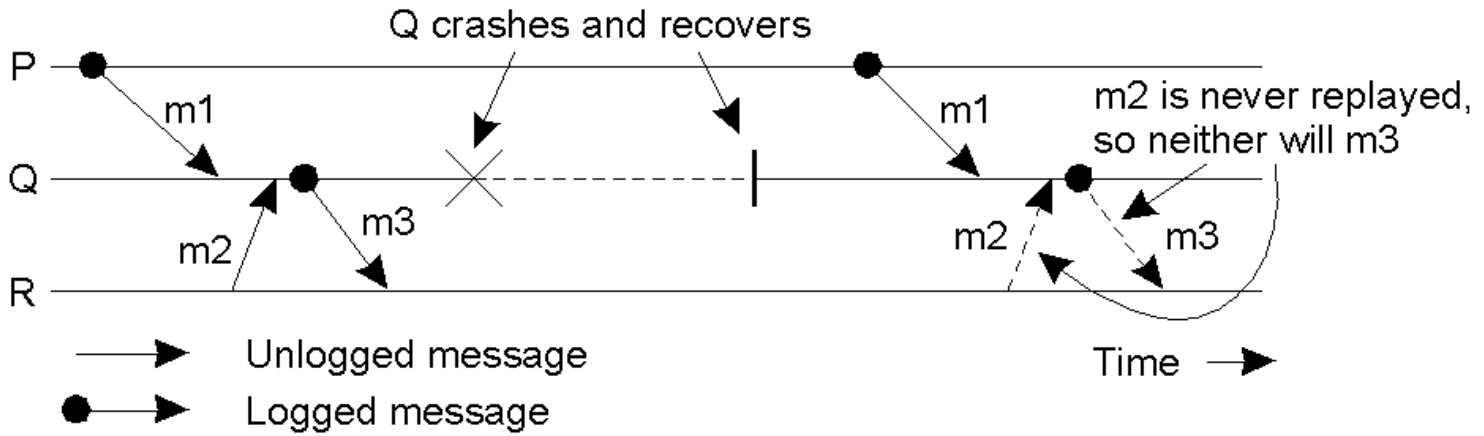
Oporavak korištenjem dnevničkog zapisa

- Koristi kombinirano kontrolne točke i dnevnički zapis
- Deterministički interval: počinje nedeterminističkim događajem koji se zapisuje je dnevnički zapis



Zapisivanje poruka u dnevnički zapis

- Pravilo: Sve nedeterminističke događaje tj. primitak poruke s pripadnim informacijama (pošiljatelj, primatelj, redni broj poruke) treba bilježiti u dnevnik



Primjer neispravnog zapisa u dnevnički zapis

Literatura

- A. S. Tanenbaum, M. Van Steen: Distributed Systems: Principles and Paradigms, Second Edition, Prentice Hall, 2007. (poglavlje 8, *Fault tolerance*)
- Ajay D. Kshemkalyani, Mukesh Singhal, Distributed Computing: Principles, Algorithms, and Systems, Cambridge University Press, 2008.
- Rachid Guerraoui, André Schiper, "Software-Based Replication for Fault Tolerance," Computer, vol. 30, no. 4, 68-74, Apr., 1997.
- Lamport, L., Shostak, R., and Pease, M. “The Byzantine Generals Problem,” *ACM Trans. Program. Lang. Syst.,* vol. 4, no. 3, 382-401, Jul. 1982.
- Défago, X., Schiper, A., and Urbán, P. “Total order broadcast and multicast algorithms: Taxonomy and survey,” *ACM Comput. Surv.* vol. 36, no. 4, 372-421, Dec. 2004.