Fault tolerance in dynamic distributed systems

Pierre Sens

Delys Team
LIP6 (Sorbonne Université/CNRS), Inria Paris

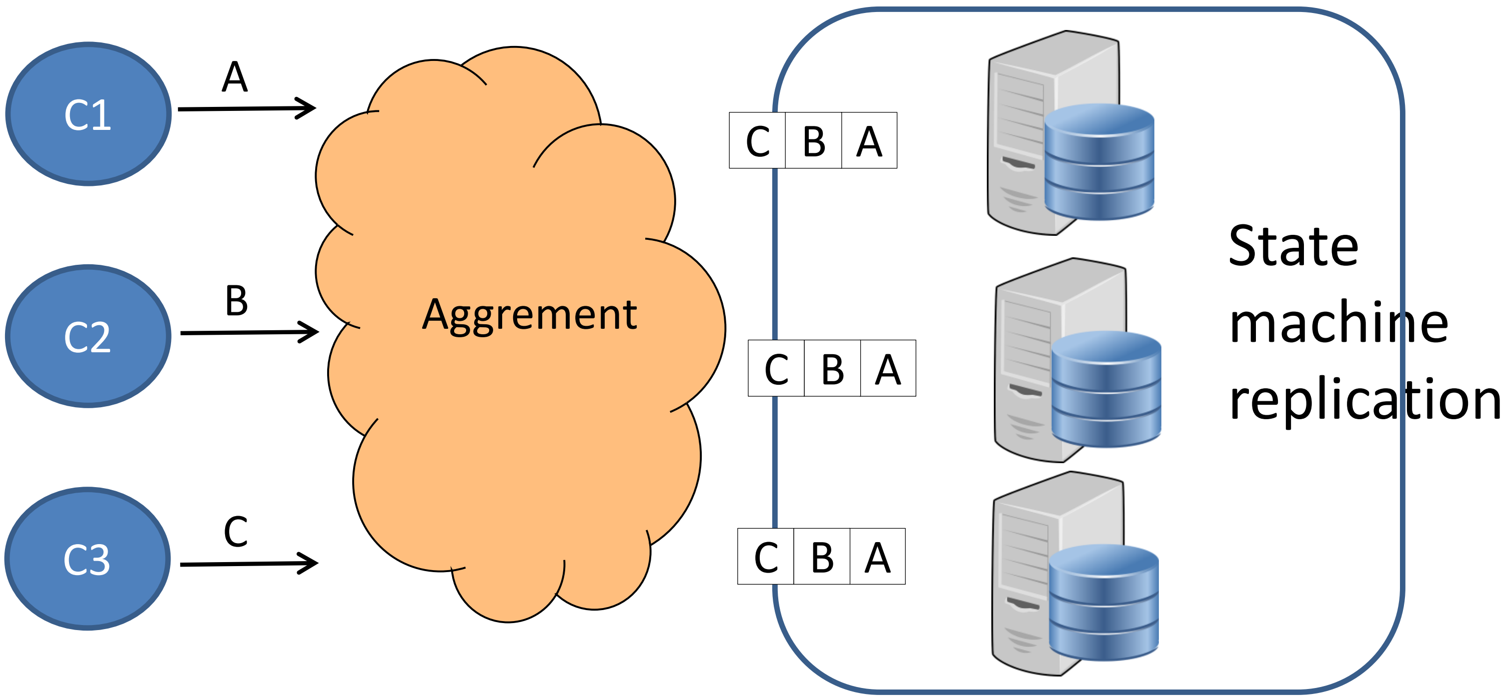
Pierre.Sens@lip6.fr

Outline

- Fundamental abstractions for distributed algorithms
- Modeling dynamic systems
- Leader election in a dynamic systems

Agreement problems

- Fondamental abstraction to build reliable services

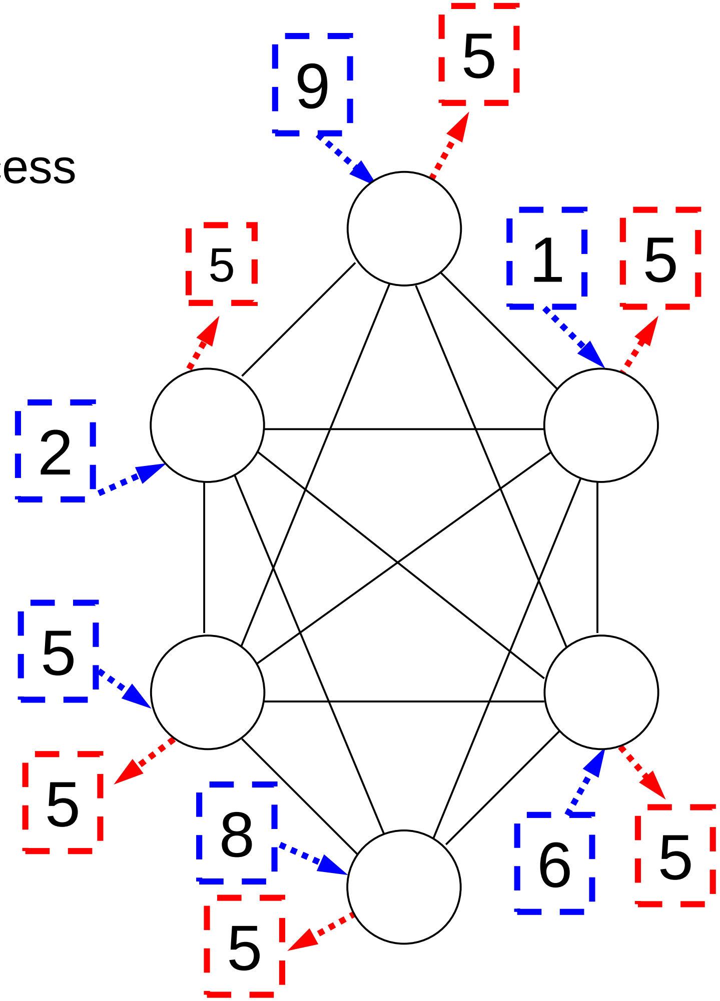


agreement on order of operations

Agreement problems: consensus

Initially 1 value proposed by each process

Eventually Every correct process decided the same proposed value



Validity: Any value decided is a value proposed

Agreement: No two correct processes decide differently

Termination: Every correct process eventually decides

Other agreement problems

all correct processes try to agree on **some set** of proposed values

- k-set agreement
  - **Agreement**: At most k values are decided.
  - **Validity**: Every value decided must have been proposed.
  - **Termination**: Eventually, every correct process decides.

  Generalization of consensus (k=1)

- set agreement: k=n-1

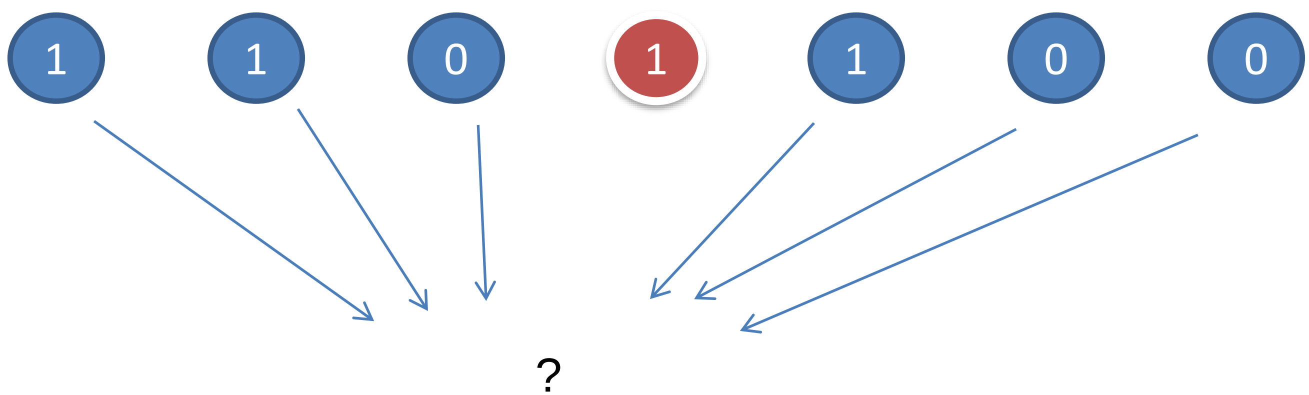
Traditional assumptions

- Connectivity
  - $\pi = \{p1,p2, .., pn\}$ **known processes**
  - n processes strongly connected (**no partition**)
- Time
  - Synchronous (known bound on transmission delays)
  - Asynchronous (no bound)
- Failures
  - Crash, recovery, Byzantine

A fundamental result

- "Impossibility to solve deterministically the **consensus** in a asynchronous networks with only 1 crash failure" **[Fischer-Lynch-Paterson 85]**

- *The idea*: impossible to distinguish faulty hosts from slow ones



Circumvent FLP impossibility

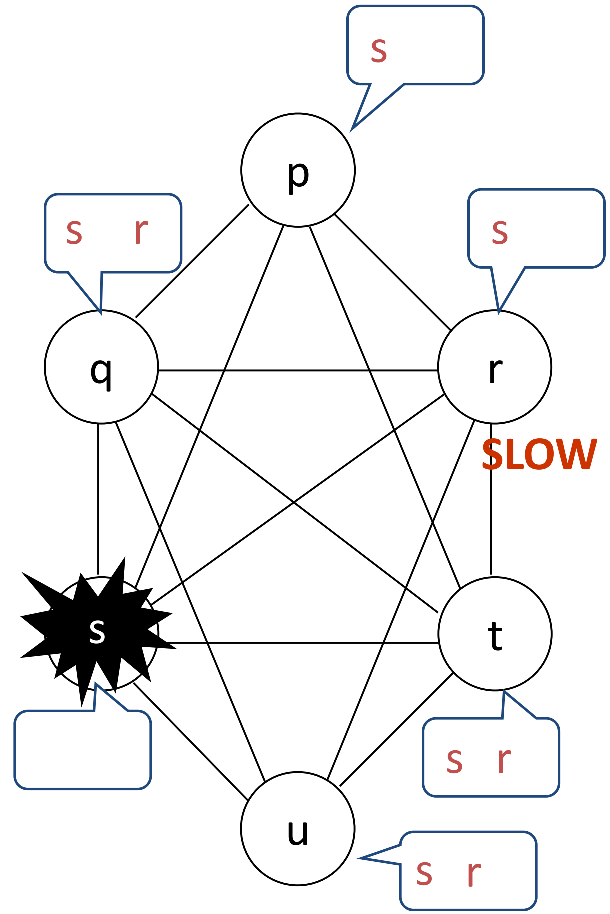
4 approaches:

- Probabilistic (probabilistic consensus, e.g., Ben-Or)
  - Possibly no termination
- k-agreement
  - A relaxed consensus (may output k different values)
- Partial synchrony
  - Add assumptions on the network
  - Eg, There is an unknown bound on the transmission delay
- **Unreliable failure detectors**

Unreliable failure detectors

- Introduced in the beginning of 90's by Chandra and Toueg
- Failure detector = an oracle per node
- Oracles provide lists of hosts *suspected* to have crashed

  => possibly false detections



System model

- *n* processes $\pi=\{p_1, . . . ,p_n\}$
- Processes communicate by message passing
- Fully connected asynchronous network
- Reliable channels
- Processes may crash (processes that do not crash are called correct)
- The system is enhanced with failure detectors

Properties of FD

- **Strong Completeness**:
  - Eventually every process that crashes is permanently suspected by ***every*** correct process
- **Accuracy**:
  - [Eventual] Strong: [There is a time after which] correct processes are not suspected by any correct processes
  - [Eventual] Weak: [There is a time after which] **some** correct processes are not suspected by any correct proc

	Accuracy			
	Strong	Weak	Eventually strong	Eventually Weak
Strong completeness	Perfect P	Strong S	◊ P	◊S

Variantes : Eventual leader

Ω : Output only **one trusted process**, the eventual leader

The leader is eventually the **same correct** process for every correct process

Weakest failure detectors

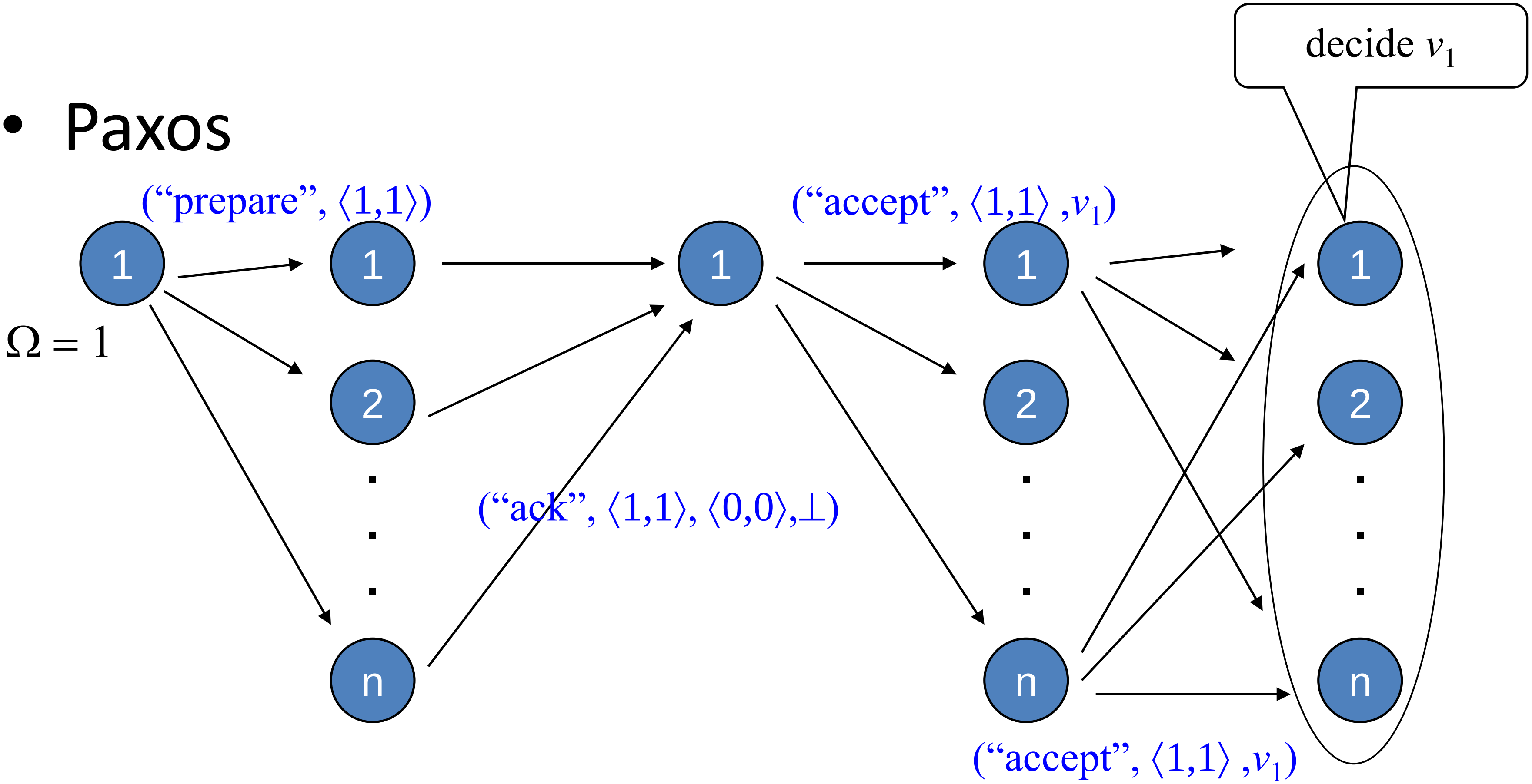
- Introduced by Chandra, Hadzilacos and Toueg
- A weakest failure detector D for a problem P has to be :
  - Sufficient: with D it is possible to solve P
  - Necessary: every other sufficient FD D’ is stronger than D (D’ can emulate D)

Ω and ◊S are the weakest FD to solve consensus with a **majority of correct** processes (eg. Paxos)

=> Ω and ◊S are equivalent

Consensus on weakest FD

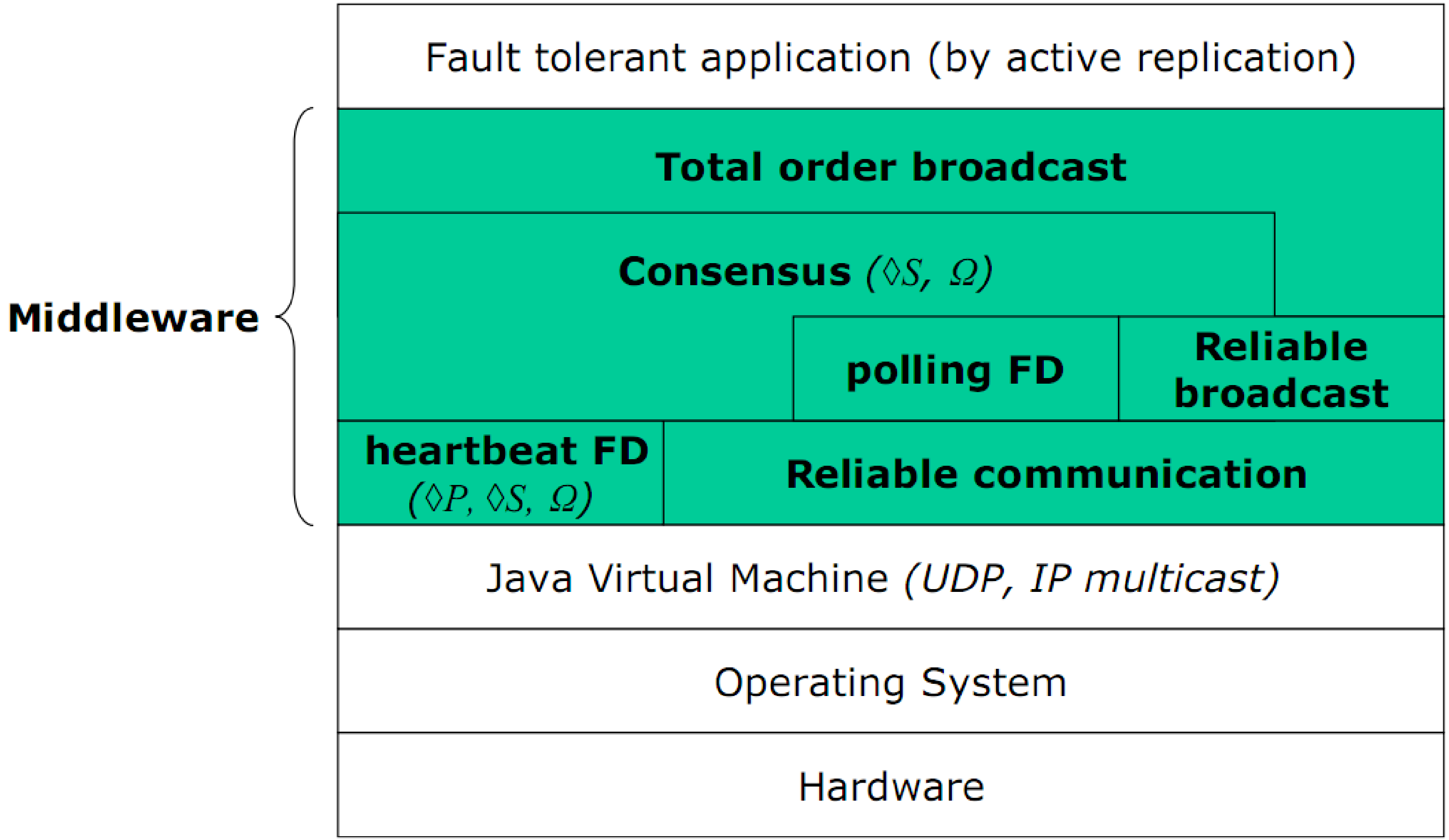
- Paxos



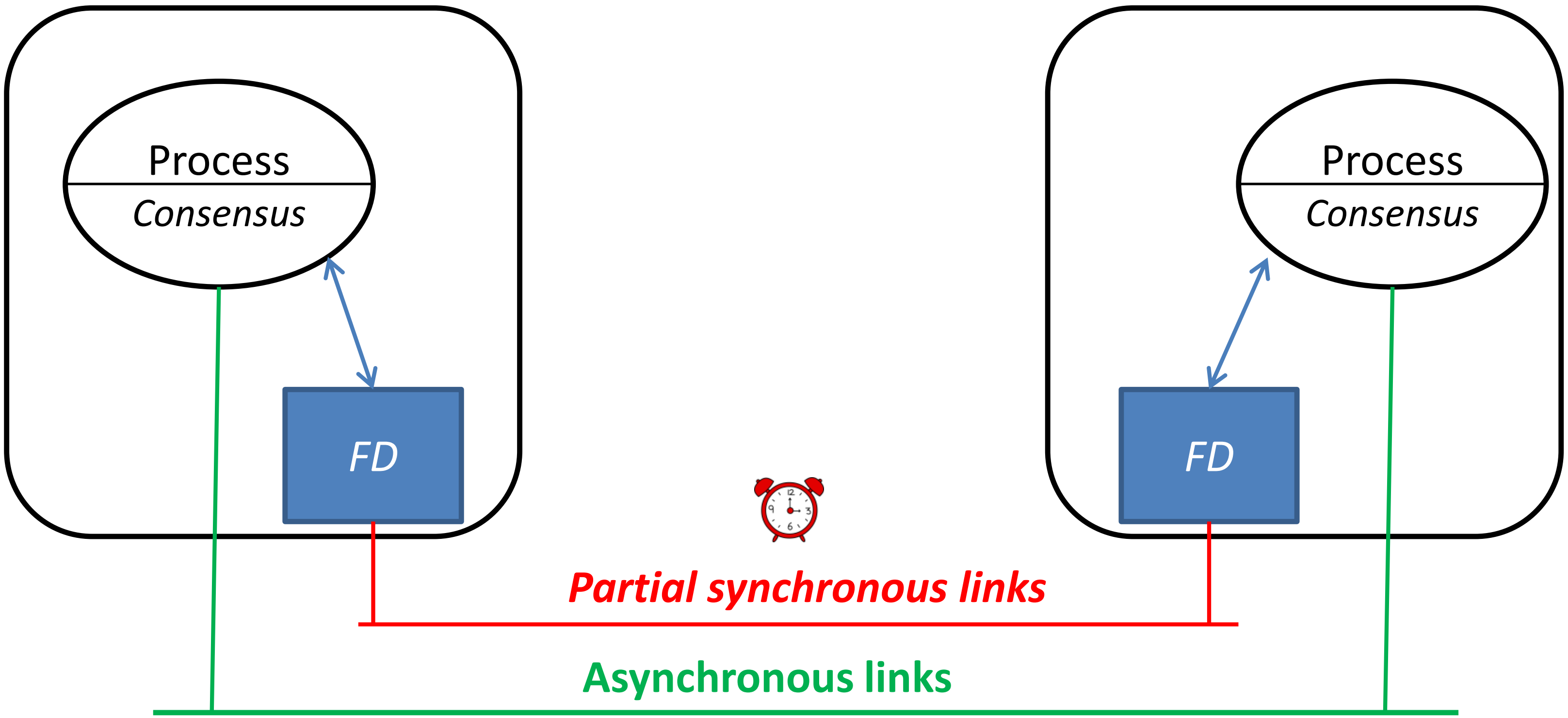
Some weakest FD results

Problems / Models	Consensus	k-set agreement	set agreement	Eventual consistency
Shared memory	Ω [LH94]	k-anti-Ω [GK09]	anti-Ω [Z10]	
Message passing	(Ω,Σ) [DFG10]	?	$\mathcal{L}$ [DFGT08]	Ω [DKGPS15]

Implementation : Fault-tolerant Architecture



Implementation of FDs



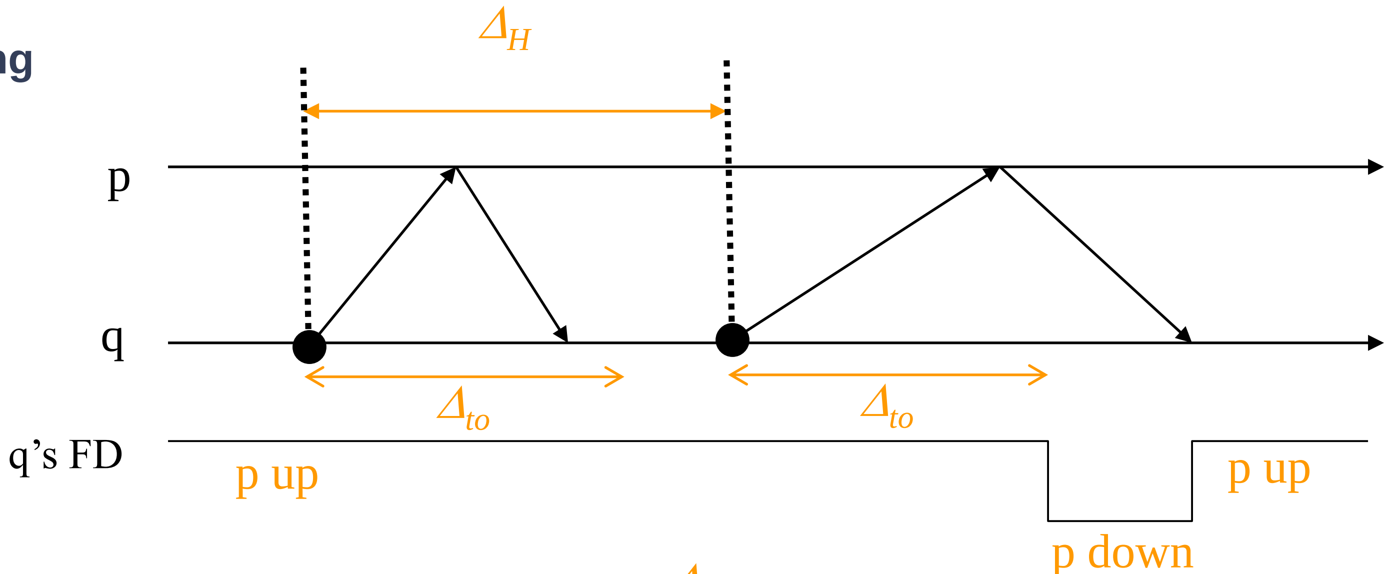
Additional assumptions

- Assumptions on transmission delay Δ and relative process speed δ
- Partial synchrony [DLS88]   *timer approach*
  1. Either Δ (δ) is known but holds only eventually, or
  2. Δ (δ) exists but is not known.
- Relative speed [MMR03]   *timer-free approach*
  – Constraints on the message pattern (message delivery order)

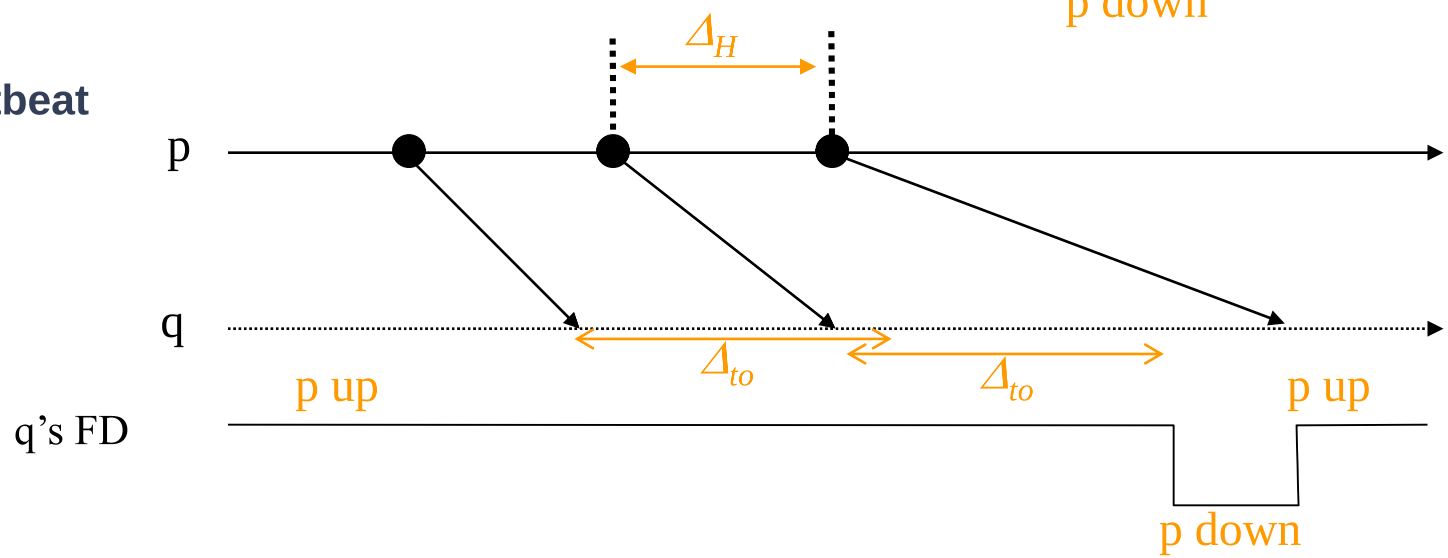
  e.g., some processes always response among the first ones

Timer approach

- **Pinging**

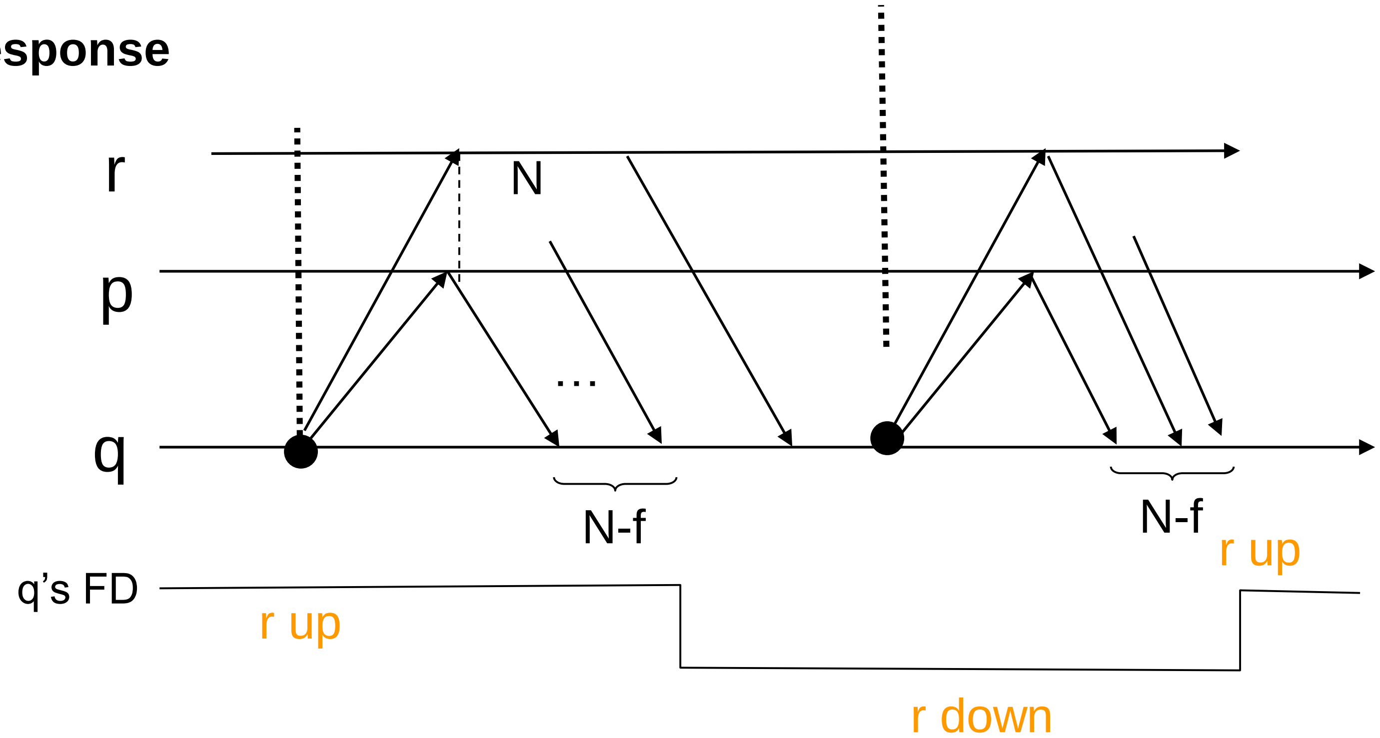


- **Heartbeat**



Timer-free approach

- **Query-response**



Example: Eventual Perfect FD (◊P)

Eventually, no false detection (accuracy)

Initialization: $suspected_i = \{\}$; $\forall\ j \neq i \in \{1, .., n\}\ \Delta_{i,j} = \Delta_0$

Task 1: repeat every Δ

send HEARTBEAT to all – $\{p_i\}$

Task 2 : when did not receive HEARTBEAT during last $\Delta_{i,j}$ from p_j

$suspected_i = suspected_i \cup \{ p_j\}$

Task 3: when received HEARTBEAT from p_j and $p_j \in suspected_i$

$suspected_i = suspected_i - \{ p_j\}$

$\Delta_{i,j} = \Delta_{i,j} + 1$

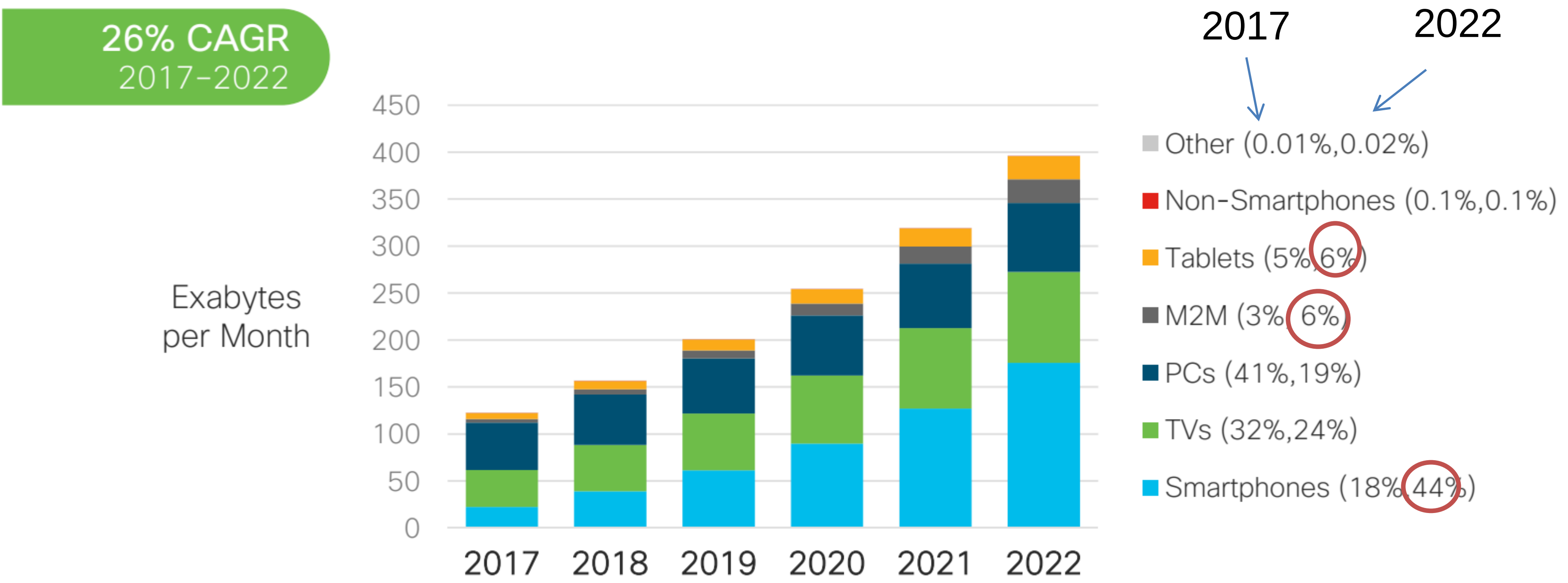
Limits of current implementations

Many implementations of FD target

- **static** systems
  - Membership (set of nodes) is initially set (no arrival)
- **known** topology
  - No change in the topology (no movement)

Distributed systems are more and more dynamic

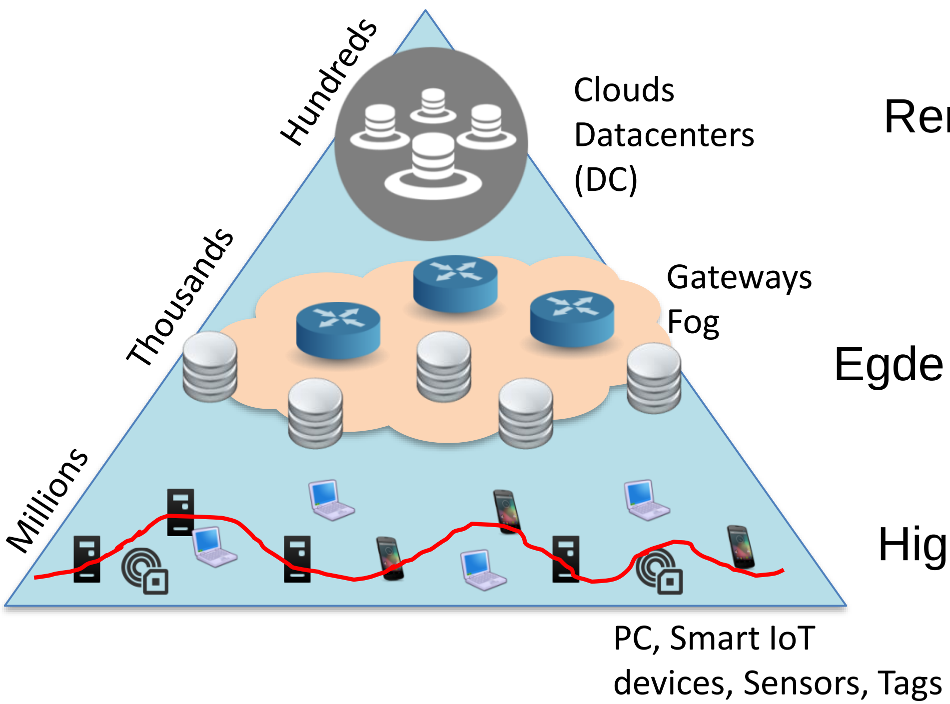
- In 2022, mobile devices will account for a half of global internet traffic



* Figures (n) refer to 2017, 2022 traffic share

Source: Cisco VNI Global IP Traffic Forecast, 2017–2022

New distributed architectures



Remote datacenters

Egde and local datacenters

Highly dynamic networks

Features of large and dynamic distributed systems

- **Huge** number of resources
  - >1M nodes

- **Dynamicity**
  - Churn: Permanent arrival and leave of nodes
  - Mobility: Devices, virtual machines … can move or migrate
  - High failure rate, failure = common event

- “Chaotic” systems with no global state

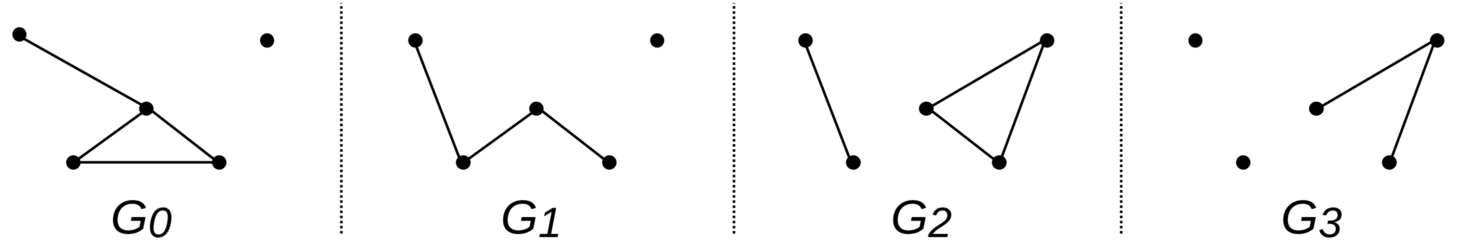
Models for dynamic systems

- Toward more dynamics : Infinite arrival models [M. Merritt and G. Taubenfeld 00]
  - Processes can be up or down
  - The number of up processes in any interval of time is upperly bounded by a **known constant C**

- Dynamic networks : dynamic graphs

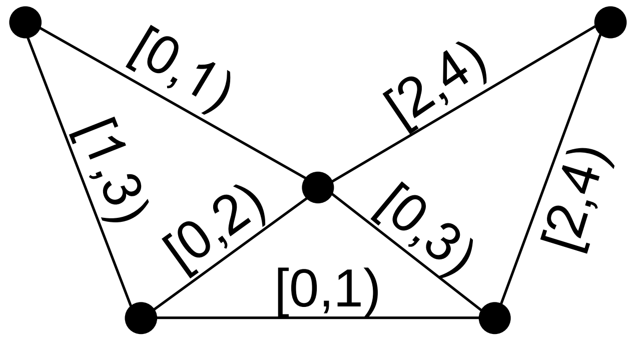
Graph Representation

- Sequence Based [B. Bui-Xuan, A. Ferreira, A. Jarry, JFCS 2003]

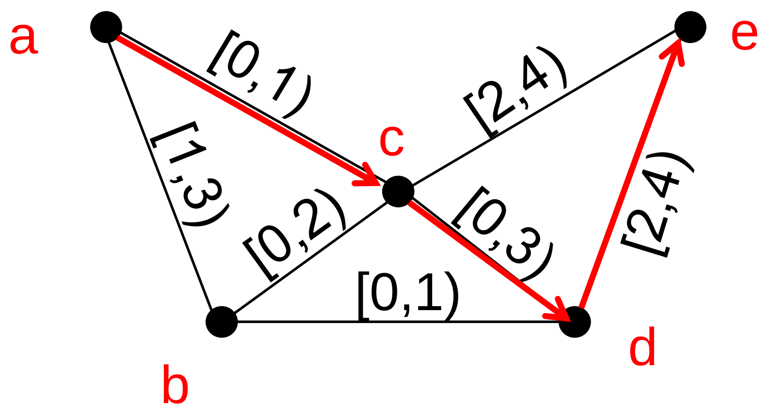


$$\mathbb{G} = G_0, G_1, G_2, G_3, \ldots G_i, \ldots, \quad i \in \mathbb{N}$$

- Time varying graphs (TVG) [A. Casteigts, P. Flocchini, W. Quattrociocchi, N. Santoro, 2012]



TVG: Basic Properties



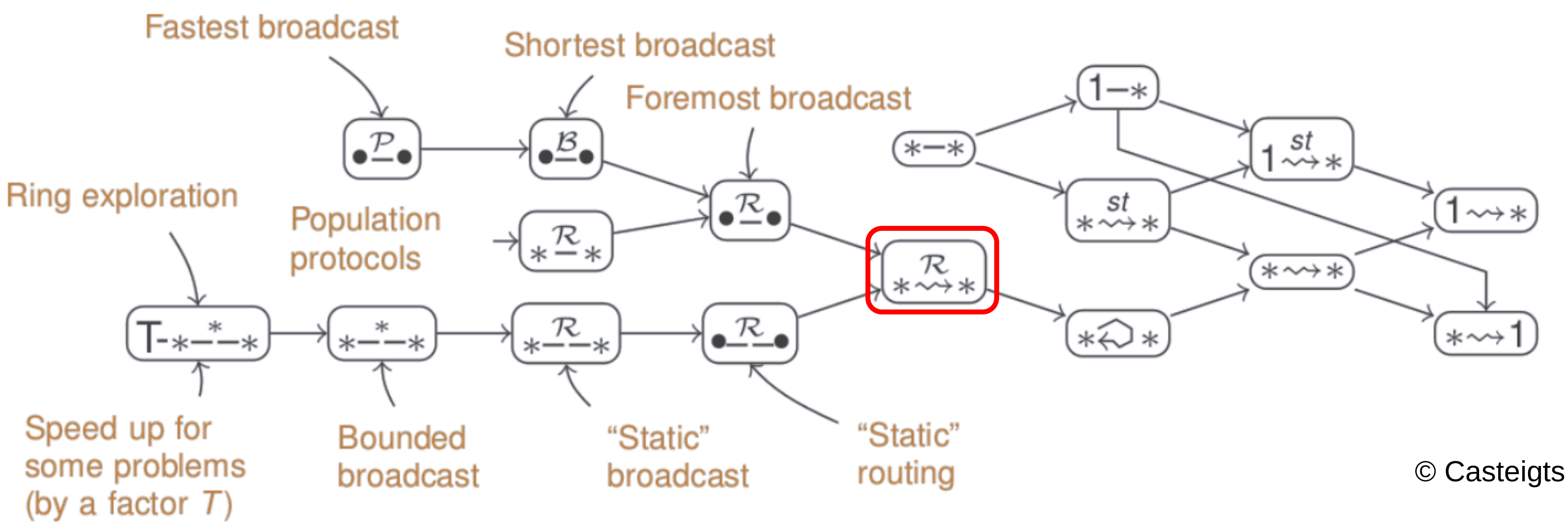
- *Temporal path (a.k.a Journey), e.g., a ↝ e*

  a ↝ *, b ↝ *, c ↝ *, d ↝ *, except e!

- 1 ↝ * ∃u ∈ V, ∀v ∈ V, u ↝ v
- * ↝ 1 ∀u ∈ V, ∃v ∈ V, u ↝ v
- * ↝ * ∀u, v ∈ V, u ↝ v

TVG: Classes

- $u \overset{\mathcal{P}}{\rightsquigarrow} v$ - Periodic journey
- $u \overset{\mathcal{B}}{\rightsquigarrow} v$ - Bounded journey
- $u \overset{\mathcal{R}}{\rightsquigarrow} v$ - Recurrent journey



- What assumption for what problems

Eventual Leader Election in Dynamic Environments

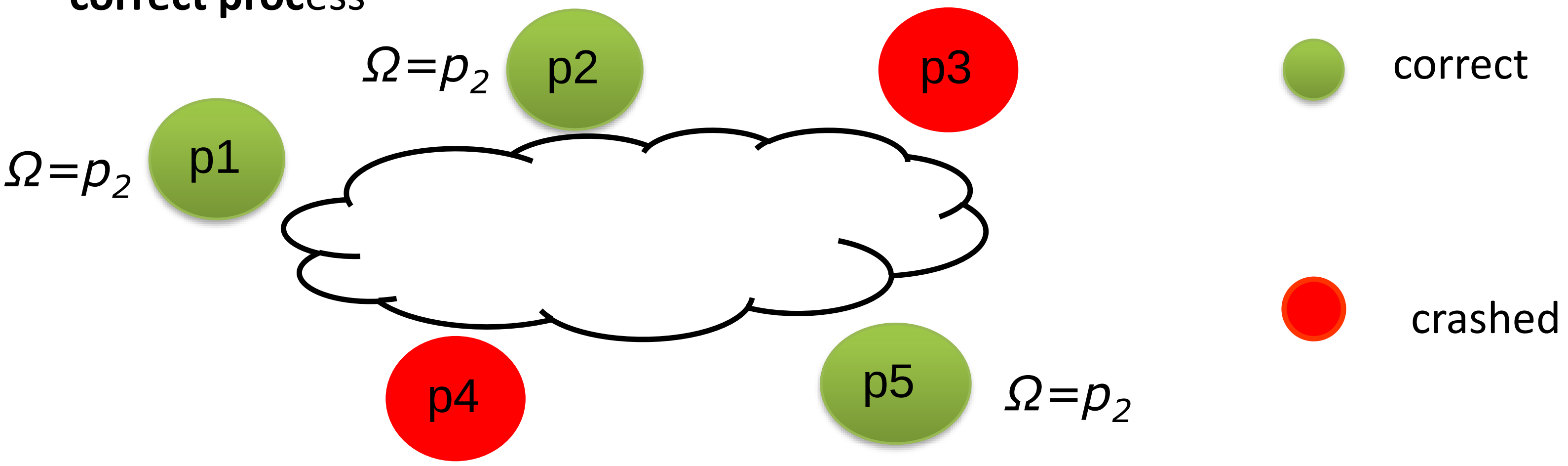
Luciana Arantes[1], Fabiola Greve[2], Véronique Simon[1], and Pierre Sens[1]

LIP6, Inria, France[1]

Federal University of Bahia (UFBA), Brazil [2]

Eventual leader election
(Ω : omega failure detector)

- The Ω failure detector satisfies ("eventual leader election"):
  - there is a time after which **every correct process** always trusts **the same correct process**



Context

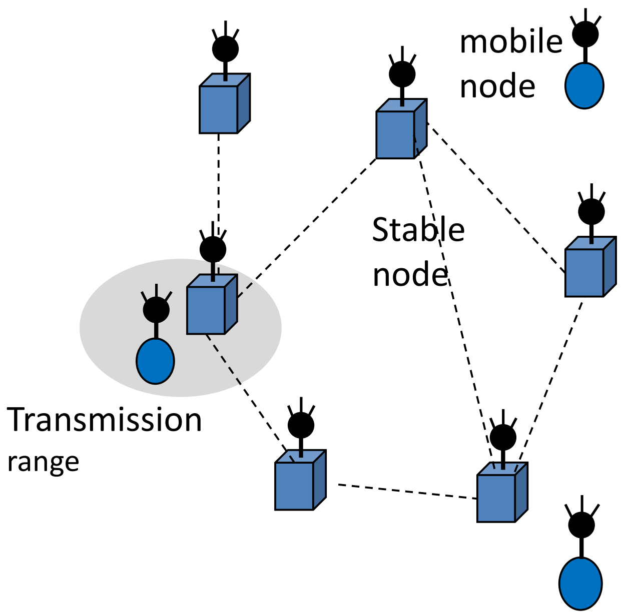
- Dynamic self-organized systems
  - Multi-hop networks (e.g. wireless ad-hoc networks)
    - broadcast /receive messages to/from neighbors within transmission range
- Communication
  - Channels are **fair-lossy**
  - there is no message duplication, modification or creation
- The system is **asynchronous**
  - There are no assumptions on the relative speed of processes nor on message transfer delays.
- Failure model : **crashes**
- The membership is **unknown**
  - A node is not aware about the set of nodes nor the number of them.
- Nodes have partial view of the network

Dynamics of the network

- Dynamic changing topology
  - join/leave of nodes,
  - mobility of nodes, failure of nodes (crash)
  - Finite arrival model
    - The network is dynamically composed of infinite mobile nodes, but each run consist of a finite set of *n* nodes.

Processes status and network connectivity

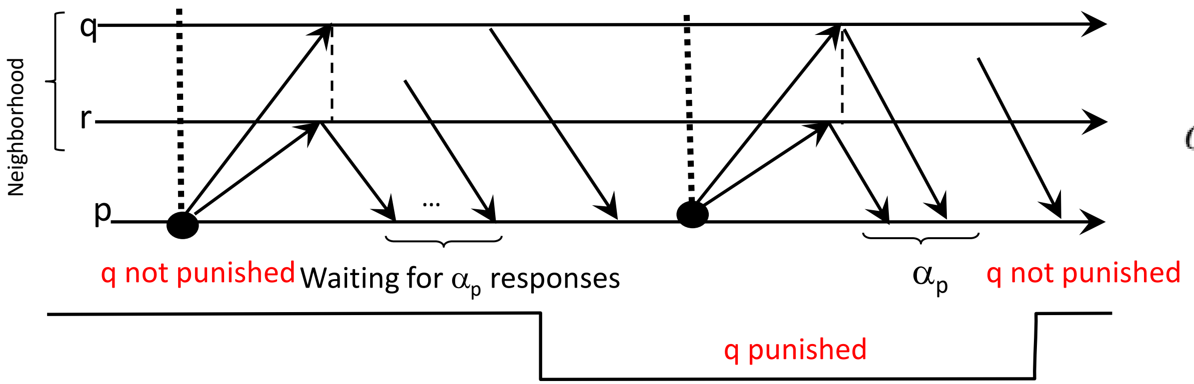
- Two sets of nodes:
  - STABLE (correct): nodes eventually and permanently correct
  - FAULTY: nodes which crash



- Network connectivity
  - Eventually, the *TVG* is connected over the time
    - There exists a journey between all stable nodes at any time
    - Network recurrent connectivity (**class** $* \overset{\mathcal{R}}{\rightsquigarrow} *$ **)**

An Eventual Leader Election Algorithm

- Principle
  - Election of a leader process based on **punishment**
    - Round counter to control the freshness of the information
  - Periodic local **query-response** exchange
    - Wait for α responses
      - If *q* is **locally known** by *p*, **has not moved**, and **does not respond** to a query of *p* among α_p first responses, *q* is punished by *p*.



$$\alpha_i = |N_i^t| - f_i + 1$$

Implementation of Ω on dynamic networks

- Each node maintains 3 sets:
  - local_known: the current knowledge about its neighborhood
  - global_known: the current knowledge about the membership of the system

  => set of tuples **<round, node id>**

  - punish: a set of tuples <punish counter, node id>

  leader: the process with the smallest counter in punish set

- Diffusion of information over the network by p:
  - *p*'s current round counter
  - set of processes punished by *p*
  - current knowledge of *p* about the membership of the system

Leader Election: Sending of Query

```
Task T1:  [Punishment]
Repeat forever
  Wait until |recvfrom_i| ≥ α_i
  If ∀p_j : ⟨−, p_j⟩ ∈ local_known_i ∧ p_j ∉ recvfrom_i ∧ MaxKnown(p_j) then *
    If ⟨0, p_j⟩ ∈ punish_i then
      c_min ← min c : ⟨c, −⟩ ∈ punish_i
      replace in punish_i ⟨0, p_j⟩ by ⟨c_min + 1, p_j⟩
    Else
      replace in punish_i ⟨v, p_j⟩ by ⟨v + 1, p_j⟩
  recvfrom_i ← ∅
  mid_i ← mid_i + 1
  broadcast QUERY(mid_i, punish_i, global_known_i)
End repeat
```

punishment

* - p_j is a neighbor of p_i,
- p_j does not answer to p_i,
- p_j is not suspected to have moved

Reception of Query and Response; Invocation of the Leader

Task T2: [Response]
upon reception of RESPONSE $(mid_j, punish_j, global\_known_j)$ from p_j

> $UpdateState(mid_j, punish_j, global\_known_j, p_j)$*
> $recvfrom_i \leftarrow recvfrom_i \cup \{p_j\}$

Task T3 [Query]
upon reception of QUERY $(mid_j, punish_j, global\_known_j)$ from p_j

> $UpdateState(mid_j, punish_j, global\_known_j, p_j)$*
> send RESPONSE $(mid_i, punish_i, global\_known_i)$ to p_j

Task T4 [Leader Election]
upon the invocation of *leader()*

> return l such that $\langle c, l \rangle = Min(punish_i)$*

***update of p_i's state about punishment, membership, and p_i's neighborhood with more recent information : keeps the tuples with the greatest counter.**

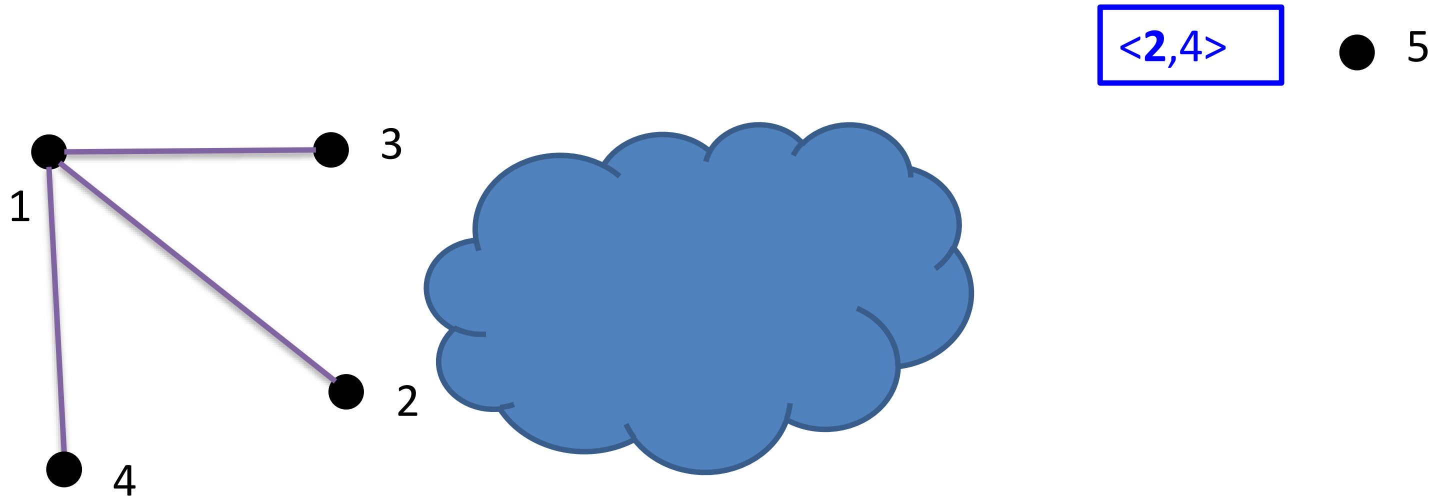
***process with the smallest counter**

Exemple: Mobility of nodes

global_known$_1$ <1,2>,<1,3>,**<2,4>**

punished$_1$ <0,1>,<0,2>,<0,3>,<**3**,4>

local_known$_1$ <1,1>,<1,2>,<1,3>,<1,4>



x:<x,4> in local_known$_1$ < y:<y,4> in global_known$_1$

➡ **1 stops punishing 4**

Additional properties

- *Stable Termination Property (SatP)*:
  - Each *QUERY* must be received by at least one stable and known node

    Necessary for the diffusion of the information

- *Stabilized Responsiveness Property (SRP):*
  - There exists a time *t* after which all nodes of *p* 's neighborhood receive, to every of their queries, a response from *p* which is always among the first responses

    ***SRP* should hold for at least one *stable* known node (the eventual leader)**

Concluding remarks

Distributed systems are **dynamic**

Failure detection a key component to build reliable application

Unreliable FDs

- A clear extension of asynchronous model
- A tool to build services in asynchronous network

Open issues : models

- Minimal condition in terms of time / connectivity / dynamicity to solve agreement problems

- Adversary models (omission, byzantin failures)

Open issues: distributed algorithms

- Non deterministic algorithms

- Probabilistic algorithms / Indulgent algorithms

  - Ensure safety properties (eg. agreement)
  - Relax liveness properties (termination)

References

[AGSV13] L. Arantes, F. Greve, P. Sens, V. Simon. Eventual Leader Election in Evolving Mobile Networks. 17th International Conference On Principles Of DIstributed Systems, 2013 (OPODIS'13)

[DFGT08] C. Delporte-Gallet, H. Fauconnier, R. Guerraoui, A. Tielmann: The Weakest Failure Detector for Message Passing Set-Agreement. DISC 2008: 109-120

[DFG10] C. Delporte-Gallet, H. Fauconnier, R. Guerraoui: Tight failure detection bounds on atomic object implementations. J. ACM 57(4): 22:1-22:32 (2010)

[DKGPS15] Swan Dubois, Rachid Guerraoui, Petr Kuznetsov, Franck Petit, Pierre Sens: The Weakest Failure Detector for Eventual Consistency. PODC 2015: 375-384

[DLS88] Dwork, C., Lynch, N., and Stockmeyer, L. Consensus in the presence of partial synchrony. Journal of the ACM 35, 2 (Apr.), 288–323, 1988

[GCG01] I. Gupta, T. D. Chandra, and G. S. Goldszmidt. On scalable and ecient distributed failure detectors. In Proceedings of the twentieth annual ACM symposium on Principles of distributed computing, pages 170179. ACM Press, 2001.

[GK09]E. Gafni, P. Kuznetsov: The weakest failure detector for solving k-set agreement. PODC 2009: 83-91

[LH94] W-K. Lo, Vassos Hadzilacos: Using Failure Detectors to Solve Consensus in Asynchronous Sharde-Memory Systems (Extended Abstract). WDAG 1994: 280-295

[MT00]. Computing with Infinitely Many Processes. In Proceedings of the 14th International Conference on Distributed Computing (DISC00), pages 164178, 2000.

[MMR03] A. Mostefaoui, E. Mourgaya, M. Raynal, "Asynchronous Implementation of Failure Detectors", Proc. Int'l IEEE Conf. Dependable Systems and Networks (DSN '03), pp. 351-360, 2003.

[WLL07] J. Wieland, M. Larrea, A. Lafuente, An evaluation of ring-based algorithms for the Eventually Perfect failure detector class. PDP 2007: 163-170

[Z10] P. Zielinski: Anti-Omega: the weakest failure detector for set agreement. Distributed Computing 22(5-6): 335-348 (2010)

Thank you !