Blin Lélia

Etat global consistant

Algorithmique répartie
M1

Etat global

- Définir un état global c'est très difficile
- Qu'il vérifie certaine propriété encore plus difficile
- Dans les systèmes répartis, la complexité de cet exercice est du a:
  - l'hétérogénéité des délais d'acheminement des messages
  - la vitesse de réaction et de calcul des sites

Etat global

- Définir un état global c'est comme vouloir prendre une photo de groupe:
  - Certains personnes sont prêtes
  - D'autres sont pas encore arrivées
  - D'autres encore sont partis

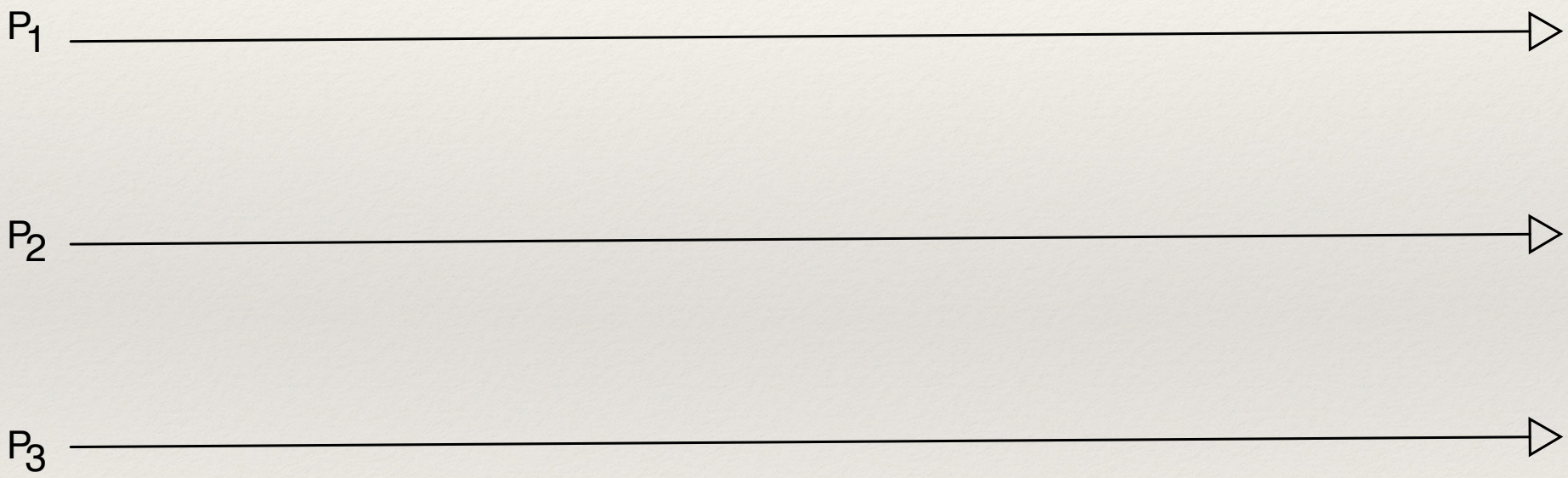
Exemple

- Trois sites partagent un objet **O**
  - Ils maintiennent cet objet.
  - Ils font évoluer cet objet.

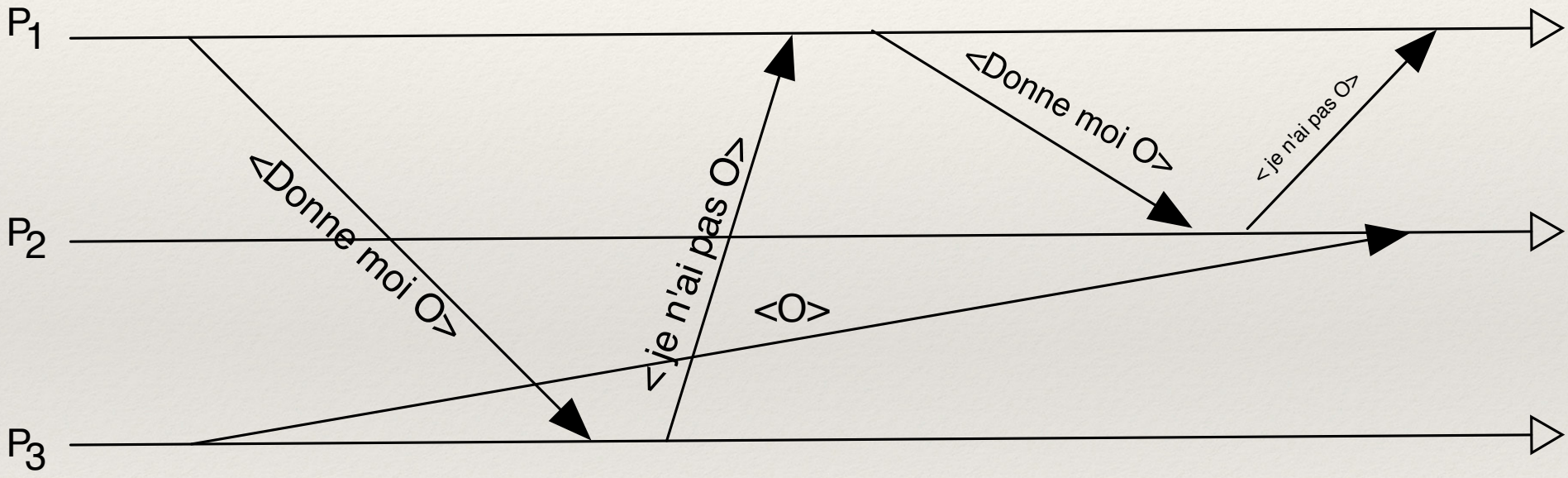
Exemple

- Lorsque un site a besoin de l'objet **O**
  - Il le demande à un autre
  - Si l'autre lui répond qu'il ne l'a pas
  - Il le demande au troisième

Diagramme d'espace-temps



Exemple



Analyse de l'exemple

- Le problème de cette situation provient du fait
  - Que le message contenant l'objet **O** met beaucoup de temps avant d'arriver en P_2 par rapport aux demandes et aux réponses
  - P_1 conclut que **O** est perdu alors qu'il est en transit entre deux sites

Etat global

- L'état local du processus i est l'ensemble des variables locales du processus i.
- LS_i notera l'état local du processus i.
- L'état global d'un système est l'union des états locaux de ses processus.
- GS l'état global
- Formellement: $GS=\{\cup_i LS : \forall i \in V\}$

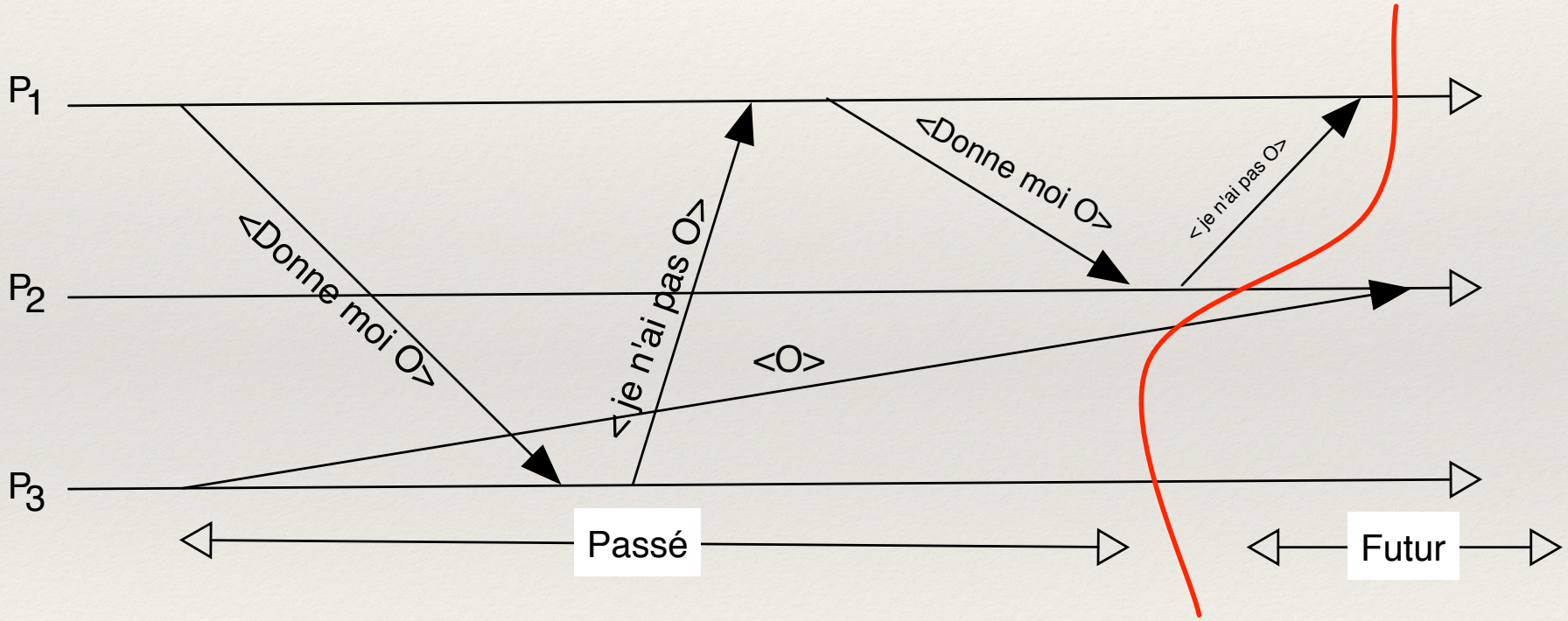
Coupe de l'espace-temps

- Une coupe dans le diagramme d'espace-temps est une ligne joignant de façon arbitraire un point de chaque processus.
- Toutes les informations avant cette ligne représente le passé.
- Toutes les informations après cette ligne représente le futur.

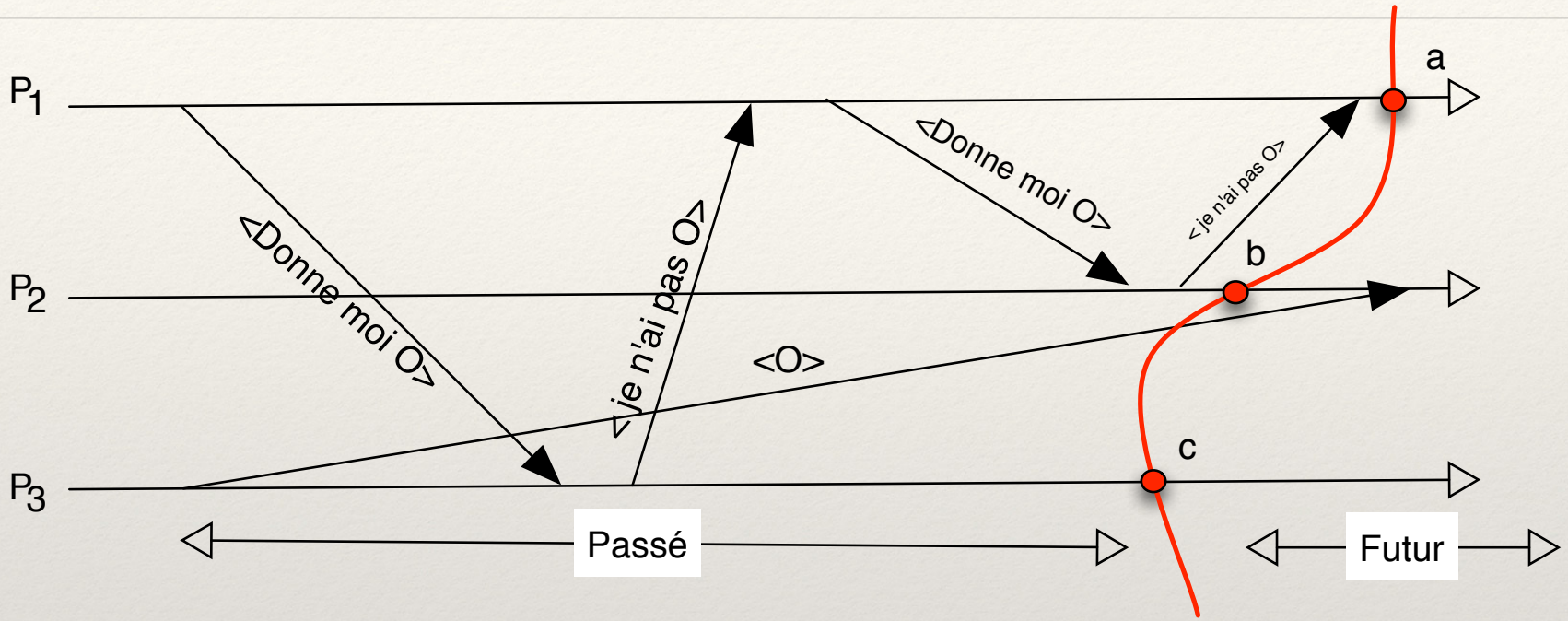
Coupe et état global

- Si on prend les états locaux au moment on la coupe rencontre chaque processus on obtient un état global

Retour sur l'exemple



Analyse



- Si un programmeur prend des traces post mortem
  - par exemple en (a,b,c)
- il constate que l'objet a bien été envoyé par P3
  - Mais qu'il n'a jamais été reçu par P2
- Il pourrait conclure hâtivement que O est perdu

- On peut dire que la «photo» {a,b,c} ne reflète pas toute la réalité
- Pourtant cette photo est une **photo globale** du système
- Dans ce cours nous allons voir quelques critères et méthode pour obtenir des **états globaux ayant certaines propriétés**

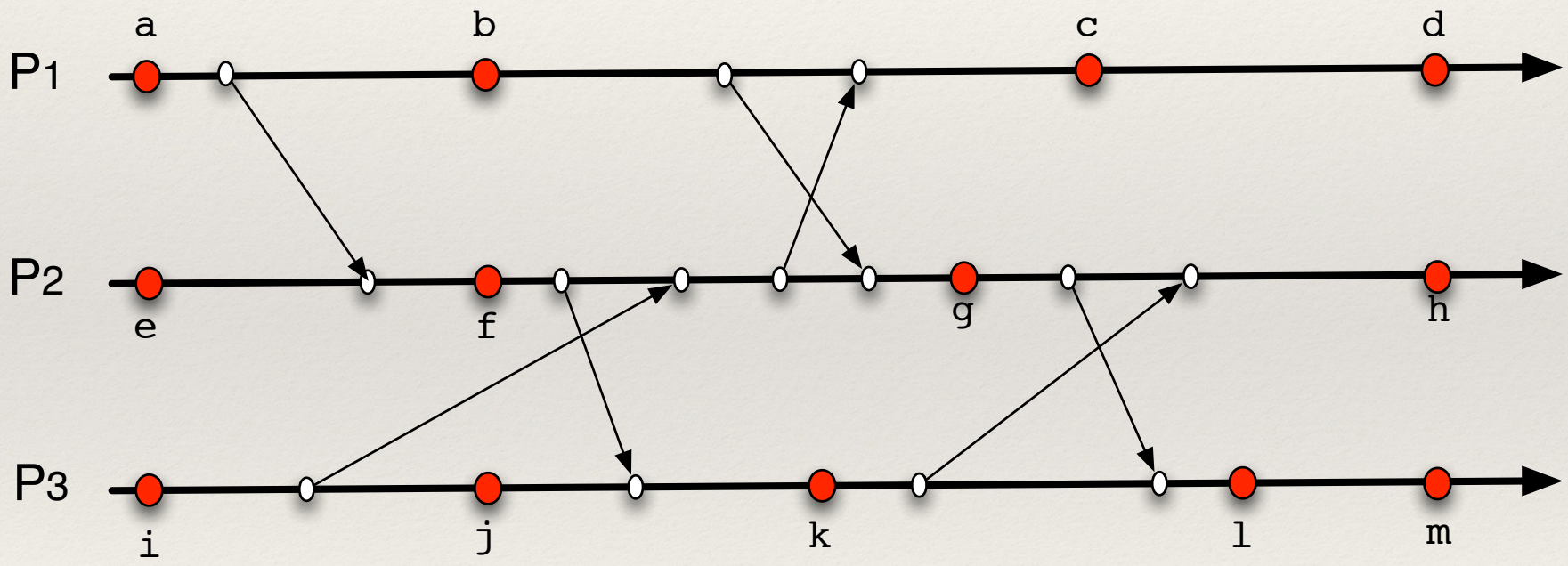
Analyse de traces d'exécution

- En vue de faire l'analyse de traces d'une exécution d'un programme réparti il convient pour chaque site de faire de temps en temps des sauvegardes d'états intermédiaires sur un support stable (disque dur…)
- On appelle ces enregistrements
  - Des points d'enregistrements d'états notés PEE

PEE

- C'est PEE sont réalisés sur chaque site
- Ils contiennent les derniers événements depuis le dernier PEE
  - Notamment les messages envoyés et reçus par le site
- On suppose:
  - Qu'en début d'exécution du processus un PEE initial est créé
  - Et que juste avant la fin du processus un PEE final est créé

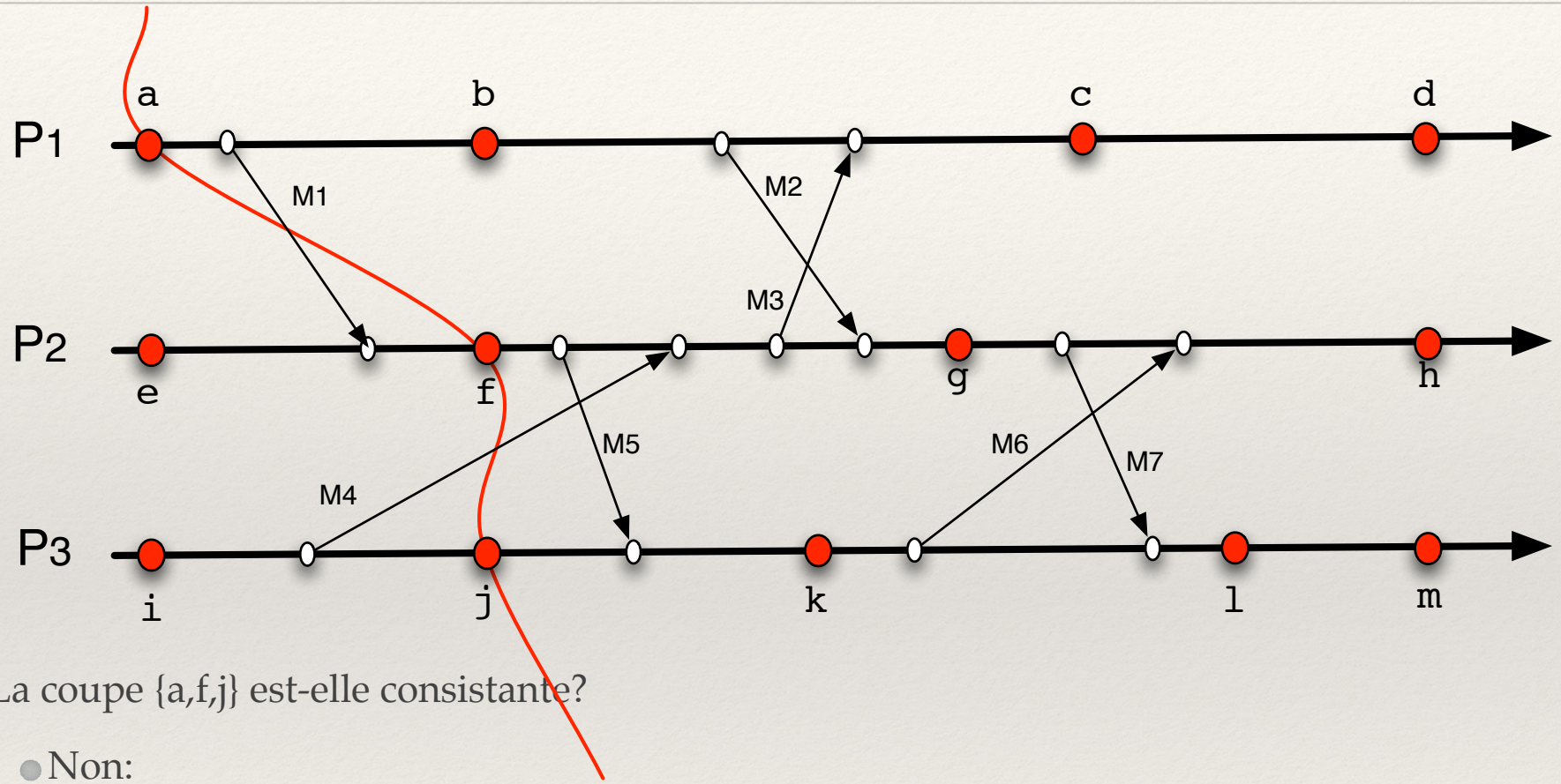
Exemple



Etat global consistant

- Un état global consistant correspond à une coupe dans laquelle:
  - Chaque message reçu dans le passé de la coupe a été envoyé dans le passé de la coupe.

Exemple

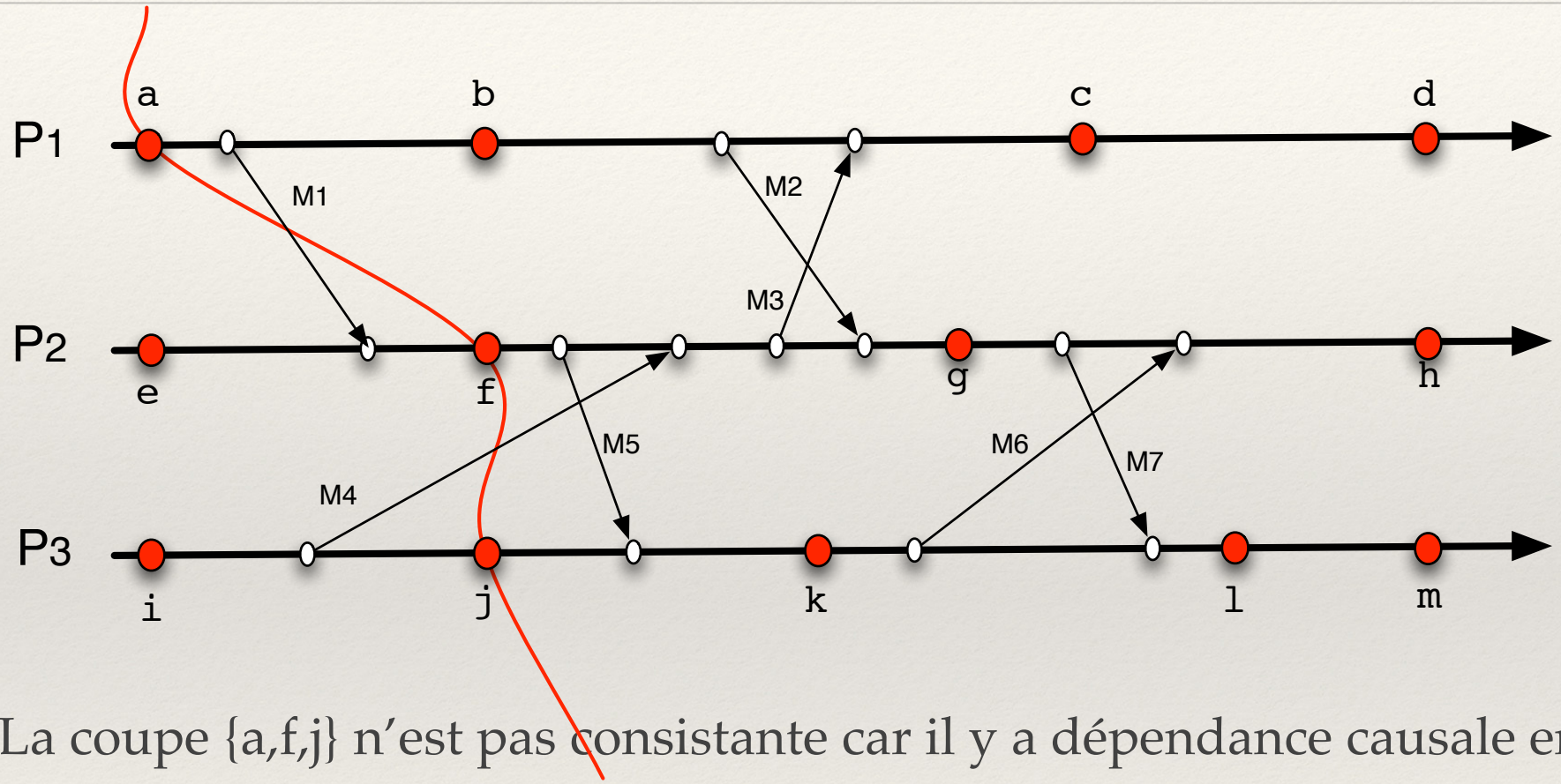


- La coupe {a,f,j} est-elle consistante?
  - Non:
    - Le message M1 reçu dans le passé a été envoyé dans le futur
    - Le message M4 envoyé dans le passé et reçu dans le futur

A quoi sert un état global consistant

- La principale motivation est la reprise sur erreur en cas de panne.
- On parle de reprise sur erreur par recouvrement arrière
- Lorsque survient une erreur dans l'exécution d'un programme réparti, on peut sous certaines conditions reprendre son exécution générale sans repartir du début

Dépendance causale

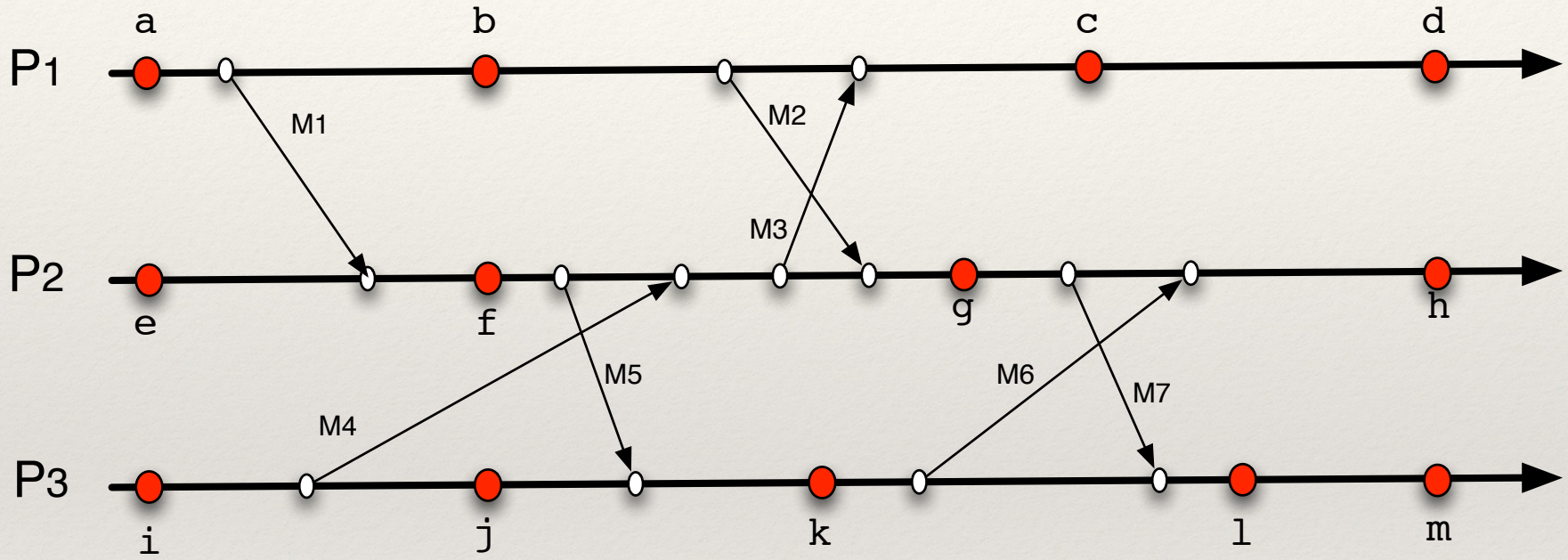


- La coupe {a,f,j} n'est pas consistante car il y a dépendance causale entre:
  - le futur et le passé: a et f
  - le passé et le futur: i et j.

Etat global consistant

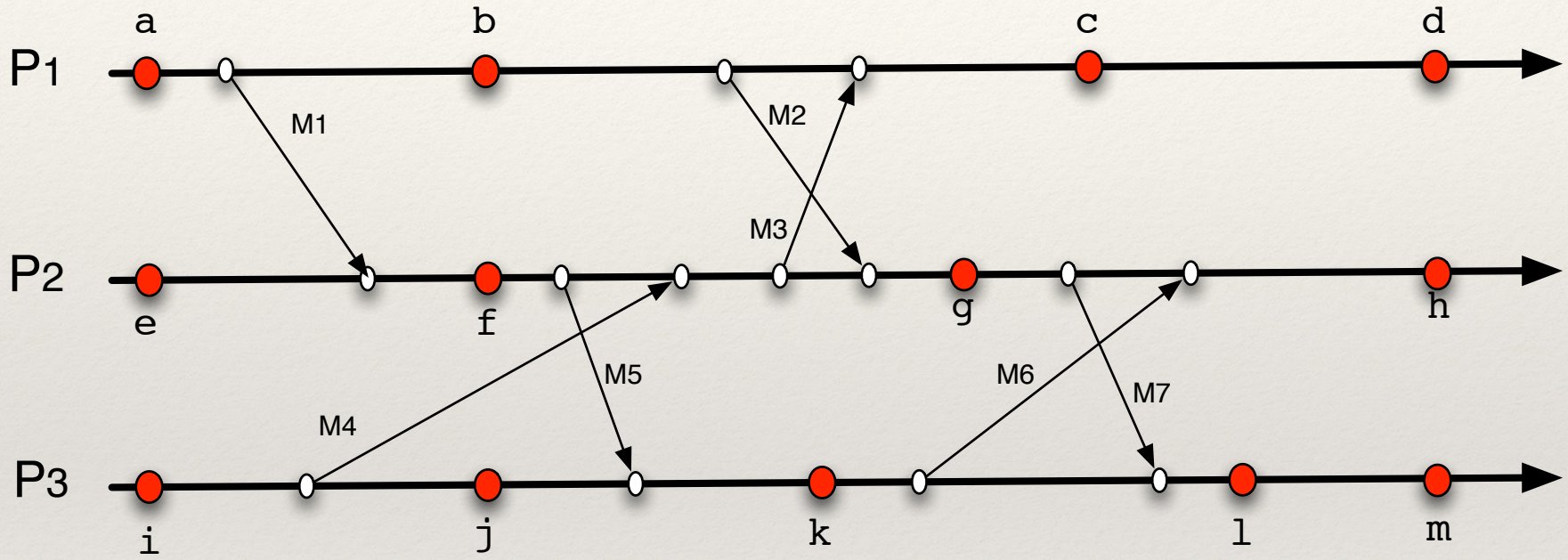
- Dans un système réparti à n sites $P_1,...,P_n$
- un ensemble $E=\{c_1,c_2,...,c_m\}$ de m PEE forme un état global consistant si
  - c_i est un PEE du site P_i pour tout i=1,…,n toute paire de PEE de E est causalement indépendante:
  - $\forall c_i \in E, \forall c_j \in E, i \neq j, c_i \nrightarrow c_j$ et $c_j \nrightarrow c_i$

Exemple



- La coupe {b,f,k} n'est pas un état global consistant car
  - f↦k
  - M5 est envoyé après f et reçu avant k

Exemple



- La coupe {b,f,k} n'est pas un état global consistant car f↦k
- La coupe {b,e,k} n'est pas un état global consistant car e↦k
- {c,g,k} est un état global consistant

Modèles de communication

- Trois modèles de communications:
  - Modèle non-FIFO: les messages envoyés par un site i par l'intermédiaire d'un lien de communication {i,j}, sont reçu par le site j dans un ordre arbitraire.
  - Modèle FIFO: les messages envoyés par un site i par l'intermédiaire d'un lien de communication {i,j}, sont reçu dans l'ordre d'envoie.
  - Modèle causal: Les messages sont délivrés en respectant l'ordre causal.

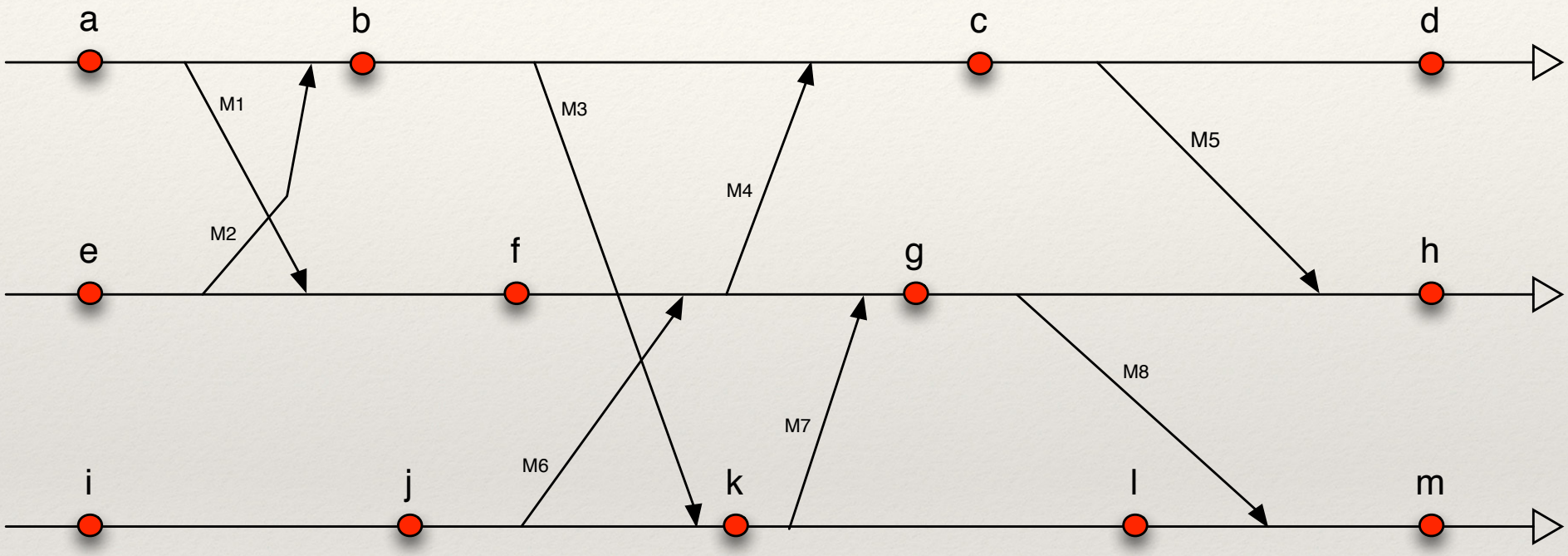
Chemin zigzag

- Chemin zigzag (Netzer and Xu):
  - Il existe un zigzag du PEE $C_{p,i}$ vers le PEE $C_{q,j}$ si et seulement si il existe un ensemble de message $M_1,...,M_m$ ($m \geq 1$) tel que
  - M_1 et envoyé par P après $C_{p,i}$
  - si M_k ($1 \leq k \leq m$) est reçu par le site r alors M_{k+1} est envoyé par r dans le même intervalle de PEE ou plus tard
  - M_m est reçu par q avant $C_{q,j}$
- Un PEE est dans un cycle zigzag s'il est dans un zigzag de lui vers lui même
- Remarque tout chemin causal est un zigzag mais le contraire n'est pas vrai

Conditions nécessaires et suffisantes

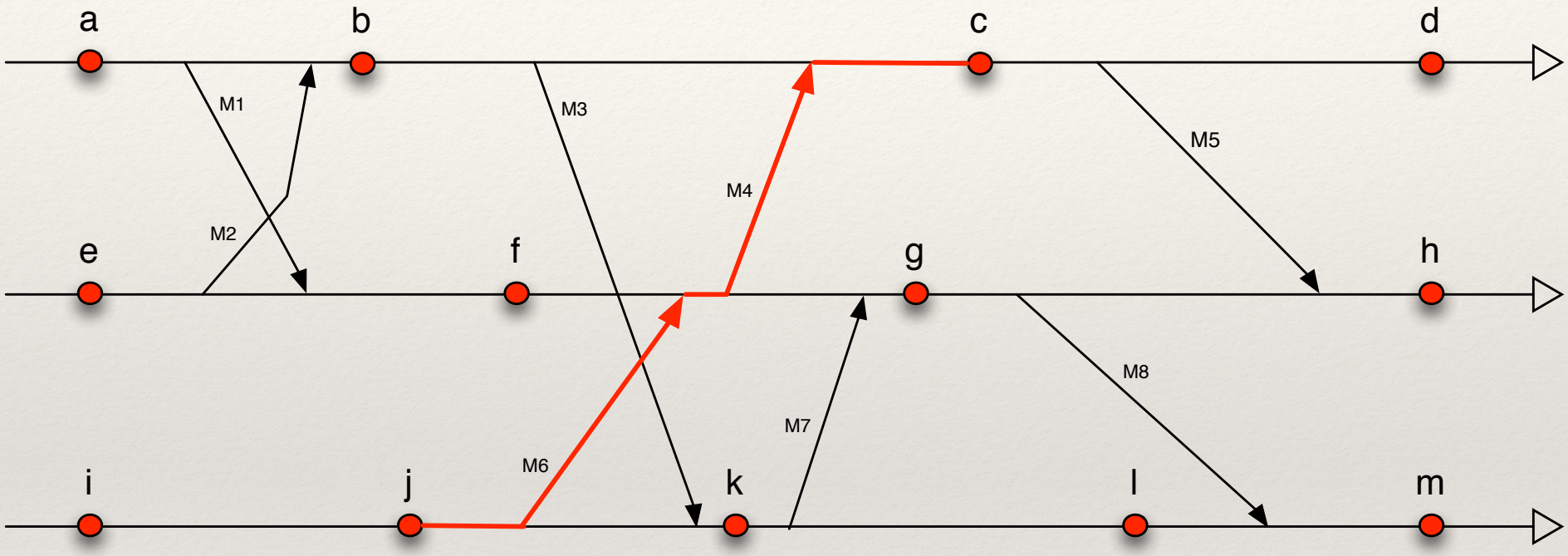
- Soit S un ensemble de PEE du système.
- Il existe un état global consistant incluant les PEE de S si et seulement si aucun PEE de S n'a de **zigzag** vers un autre PEE de S ou vers lui même.

Exemple



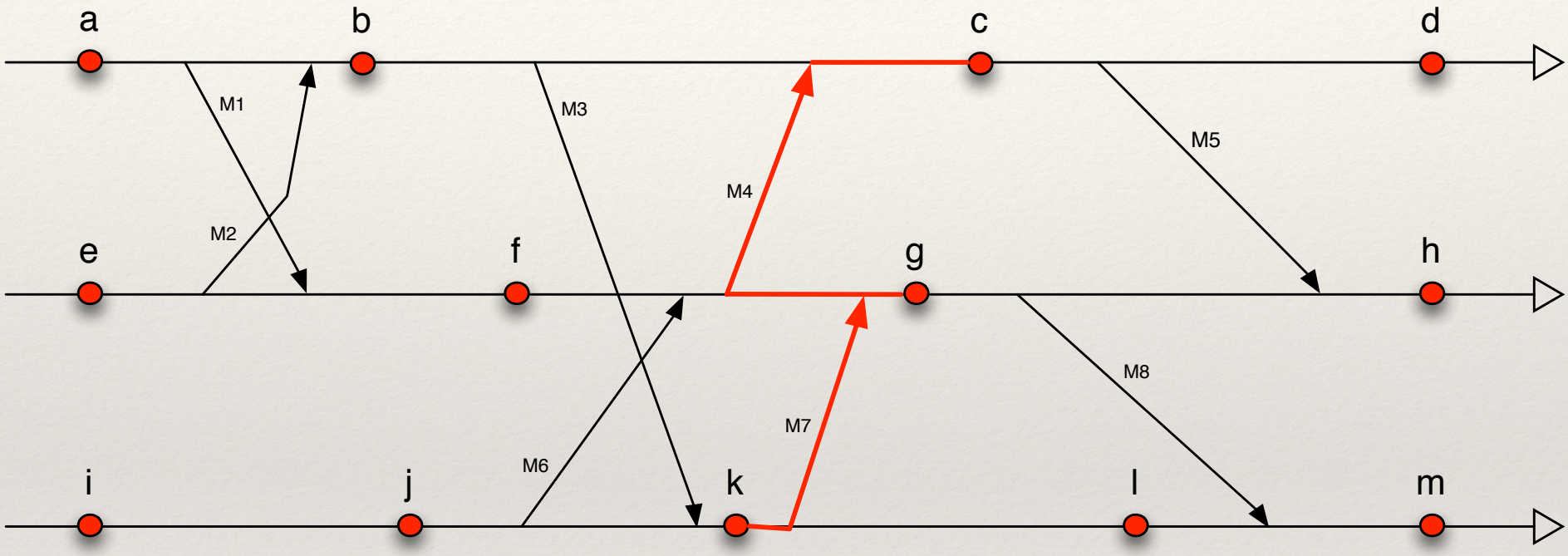
- {j,M6,M4,c} ?

Exemple



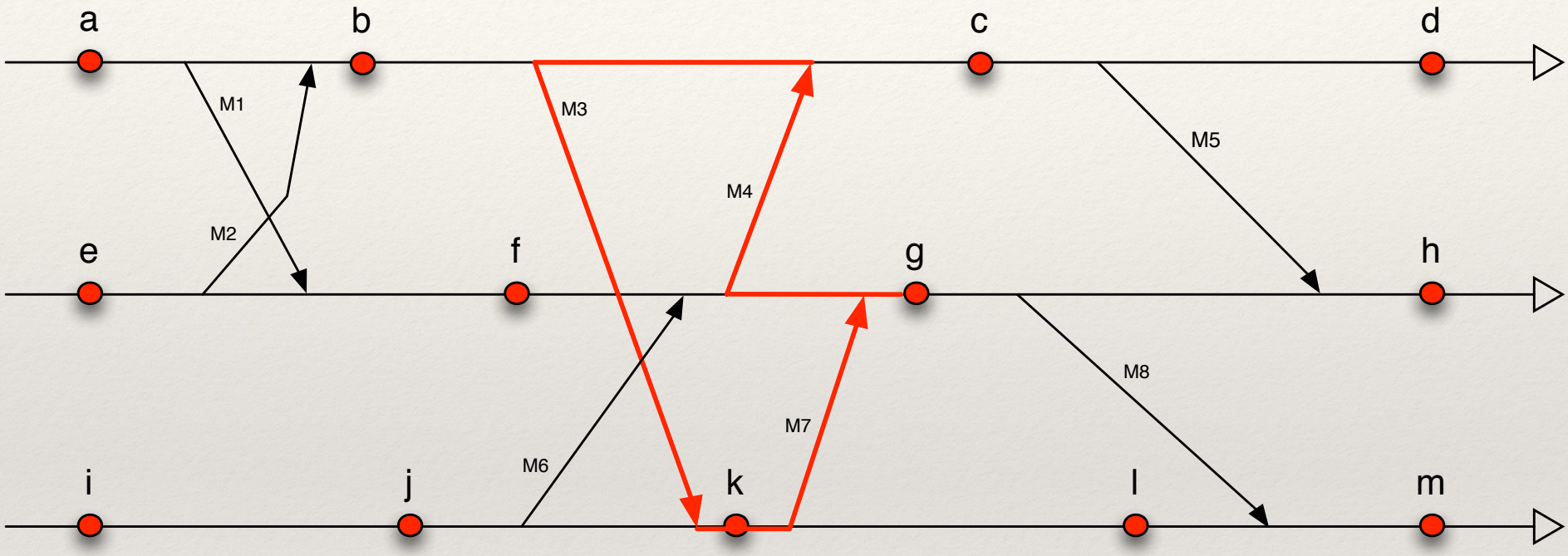
- {j,M6,M4,c} est un chemin causal donc un chemin zigzag
- {k,M7,M4,c}?

Exemple



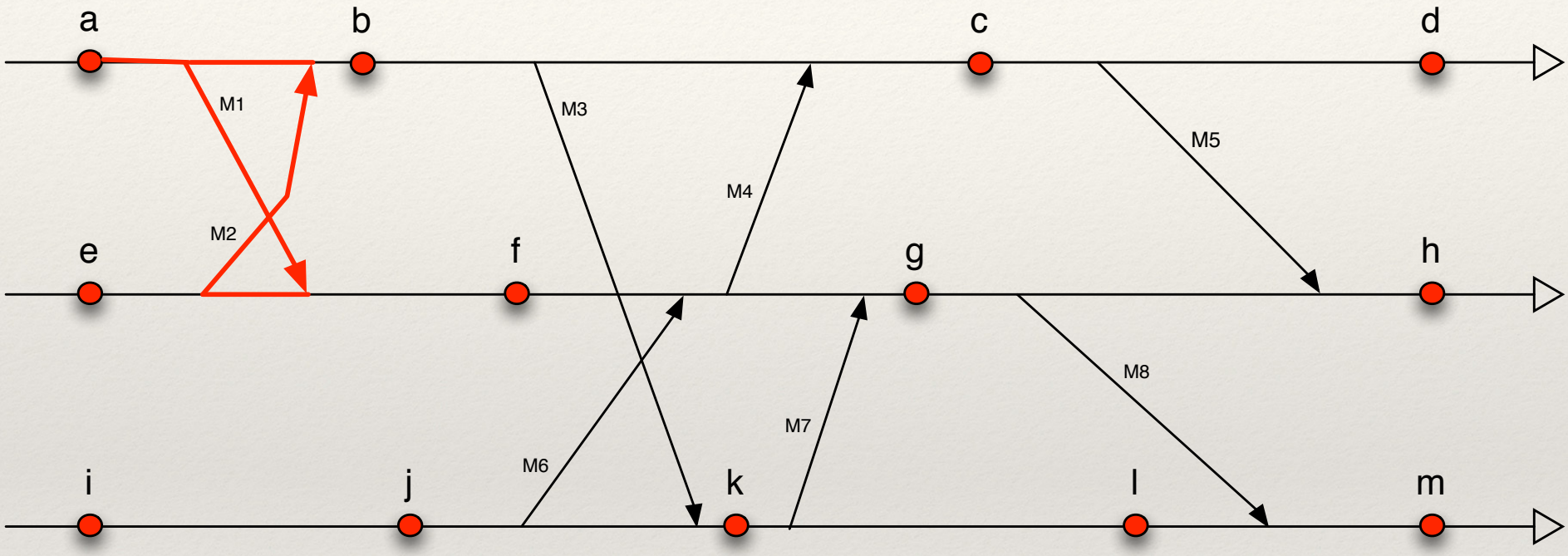
- {j,M6,M4,c} est un chemin causal donc un chemin zigzag
- {k,M7,M4,c} n'est pas un chemin causal, mais c'est un zigzag
- {k,M7,M4,M3,k}?

Exemple



- {j,M6,M4,c} est un chemin causal donc un chemin zigzag
- {k,M7,M4,c} n'est pas un chemin causal, mais c'est un zigzag
- {k,M7,M4,M3,k} est un cycle zigzag
- {a,M1,M2,a}?

Exemple



- {j,M6,M4,c} est un chemin causal donc un chemin zigzag
- {k,M7,M4,c} n'est pas un chemin causal, mais c'est un zigzag
- {k,M7,M4,M3,k} est un cycle zigzag
- {a,M1,M2,a} n'est pas un chemin zigzag car M2 est reçu après a.

Lélia Blin

Preuve

Algorithmique répartie
M1

Notations

- On notera:
  - a$\nrightarrow$b un chemin causal entre a et b
  - a$\not\rightarrow$b l'absence de chemin causal entre a et b
  - a$\rightsquigarrow$b un chemin zigzag entre a et b
  - a $\oplus$ b l'absence de chemin zigzag entre a et b

Preuve

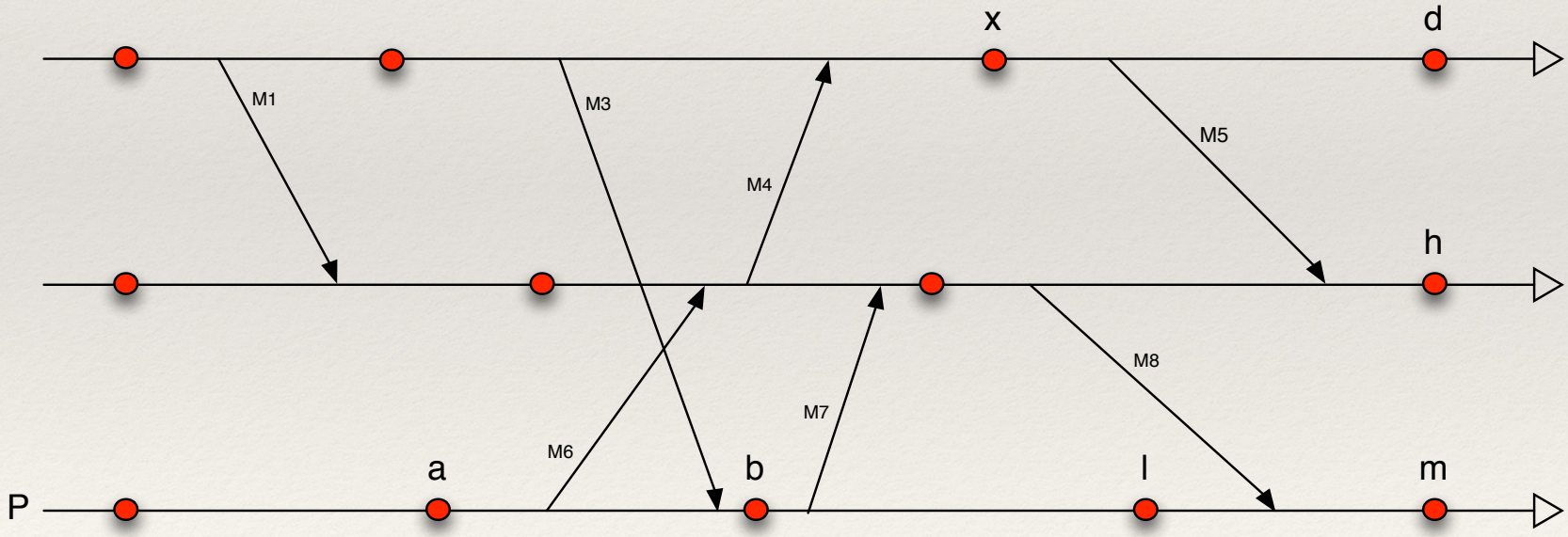
- Montrons que l'absence de zigzag entre membres de S implique que S peut-être étendu en un état global consistant

Partie 1

- On construit un état global noté G tel que S⊂G avec les règles suivantes:
  - Premièrement G:=S
  - Ensuite tant qu'il existe un site P n'ayant pas encore de PEE dans G appliquer les règles.
    1. Si P a un PEE a tel que ∃x∈S,a→x, faire G:=G∪{b} avec b le premier PEE de P tel que ∀x∈S,b ⊕ x
    2. Si P a un PEE a tel que ∀x∈S, a ⊕ x, faire G:=G∪{b} avec b le PEE initial de P

Première règle

1. Si P a un PEE a tel que $\exists x \in S, a \rightsquigarrow x$, faire $G := G \cup \{b\}$ avec b le premier PEE de P tel que $\forall x \in S, b \oplus x$



Première règle

- Pour la première règle, on est sûr de pouvoir construire un PEE adapté car P a au moins son PEE final qui correspond à la contrainte de zigzag vers les PEE de S.
- On constate qu'à la fin de la construction de G on a:
  - $\forall x \in G, \forall y \in S, x \oplus y$ (1)

Première règle: Preuve

- Prouvons qu'il n'y a pas de chemin causal entre membre de G
- Preuve par contradiction:
- Examinons chaque sous-cas induit par cette hypothèse:
  1. a∈S, b∈S et a⇢b. Ceci contredit (1)
  2. a∈G-S, b∈S et a⇢b. Ceci contredit (1)
  3. a∈S, b∈G-S et a⇢b. A étudier.
  4. a∈G-S, b∈G-S et a⇢b. A étudier.

Première règle: Preuve

3. a∈S, b∈G-S et a⇝b.

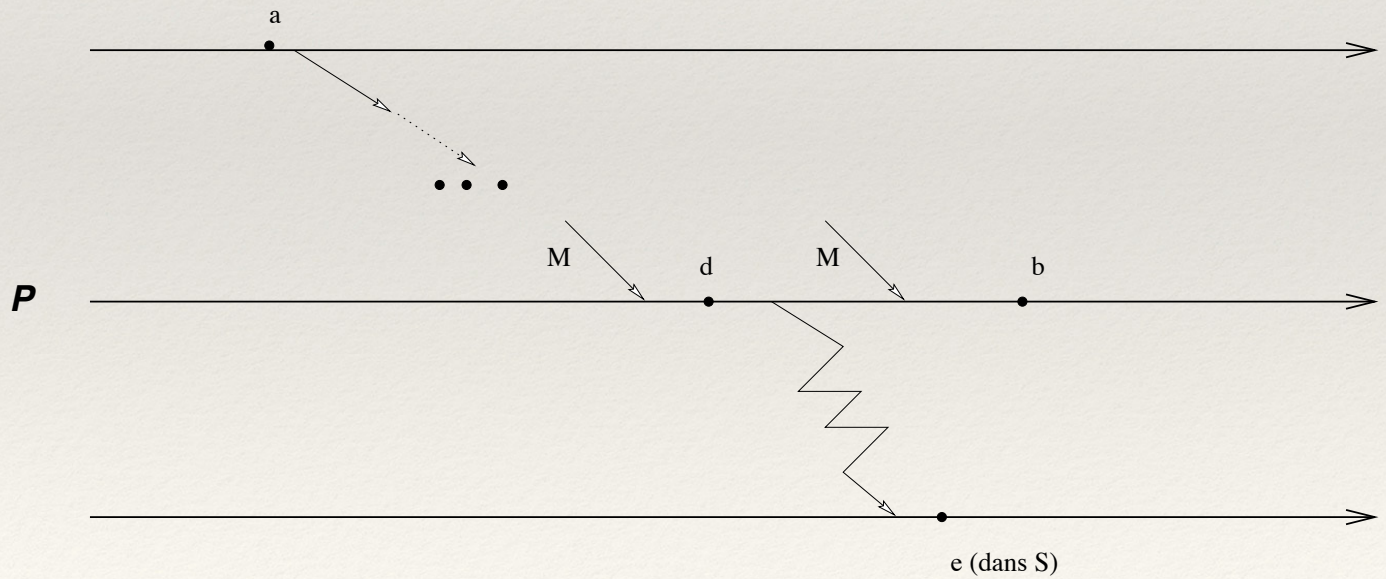
- Dans ce cas b ne peut pas être un PEE initial.
- Donc b a été rajouté par la règle 1 de construction et b est donc le premier PEE tel que
  - ∀x∈S, b ⊕ x
- Ainsi le PEE d précédant b sur le même site doit donc avoir un zigzag vers un PEE de S
  - ∃e∈S, d⇝e
- Or, comme a⇝b, il existe un chemin causal de a vers b composé d'envois et de réceptions de message jusqu'au site P.

Première règle: Preuve

3. a∈S, b∈G-S et a⇝b.

- Dans ce cas b ne peut pas être un PEE initial.
- Donc b a été rajouté par la règle 1 de construction et b est donc le premier PEE tel que
  - ∀x∈S, b ⊕ x
- Ainsi le PEE d précédant b sur le même site doit donc avoir un zigzag vers un PEE de S
  - ∃e∈S, d⇝e
- Or, comme a⇝b, il existe un chemin causal de a vers b composé d'envois et de réceptions de message jusqu'au site P.

- Le dernier message M arrivant en P peut être reçu avant ou après le PEE d,
- mais dans tout les cas, on peut construire un zigzag de a∈S vers e∈S, ce qui contredit (1).
- La figure donne une illustration de cette situation

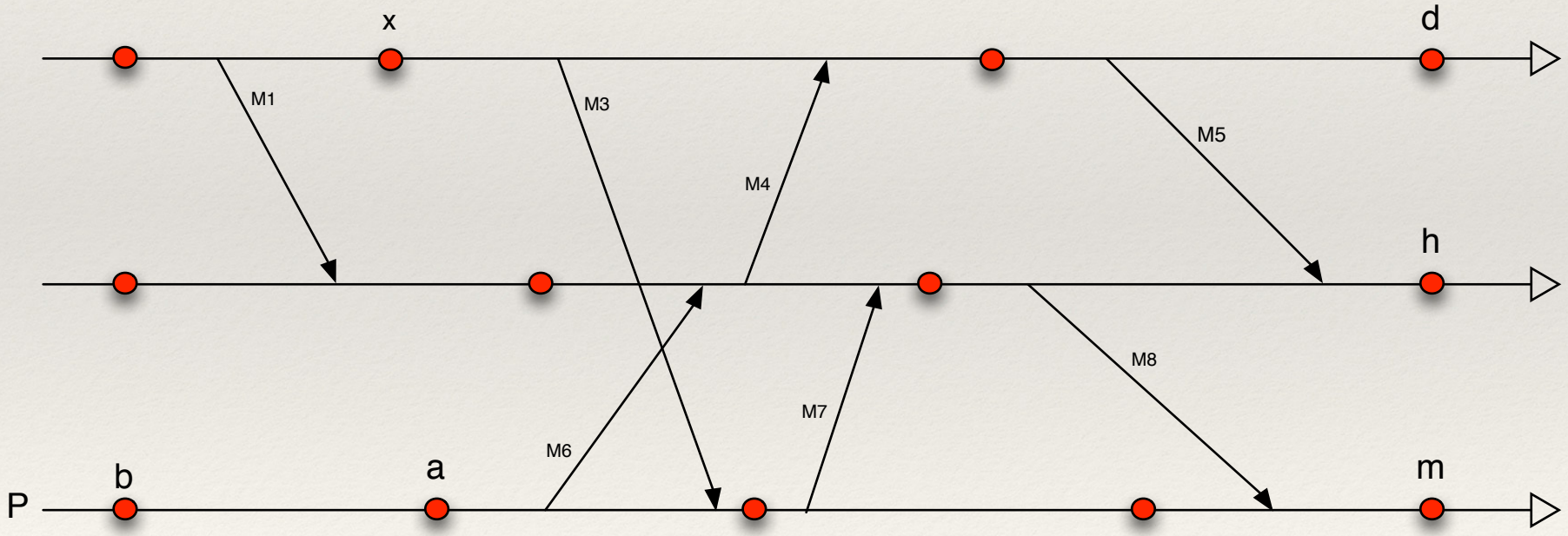


4. $a \in G\text{-}S$, $b \in G\text{-}S$ et $a \nrightarrow b$.

- Avec le même raisonnement que dans le début du cas 3 on peut conclure qu'il existe $e \in S$ tel que $a \rightarrow e$ ce qui contredit (1)

Partie 2

2. Si P a un PEE a tel que $\forall x \in S$, $a \oplus x$, faire $G := G \cup \{b\}$ avec b le PEE initial de P



Partie 2: Preuve

- Montrons maintenant que si un **zigzag** existe entre deux PEE de S alors S ne peut être un sous-ensemble d'un état global consistant.
- Supposons que $a \in S$, $b \in S$ et $a \rightsquigarrow b$
- Montrons par récurrence sur le nombre ℓ de messages composant le zigzag de a vers b, que a et b ne peuvent pas appartenir à un même état global consistant.

Partie 2: Preuve

- **Cas de base de la récurrence:** $\ell = 1$
  - Dans ce cas on a a→b ce qui est contraire à la définition d'un état global consistant contenant a et b.
- Hypothèse de récurrence:
  - Un zigzag composé de ℓ messages d'un PEE x vers un PEE y (éventuellement x=y) implique que x et y ne peuvent pas appartenir au même état global consistant.

Partie 2: Preuve

- Supposons maintenant qu'il y est un zigzag de a vers b composé de ℓ +1 messages $M_1,\ldots,M_{\ell+1}$
- Montrons que a et b ne peuvent pas être dans un même état global consistant.
- Pour cela supposons le contraire

Partie 2: Preuve

- Notons c le PEE suivant immédiatement la réception de M_ℓ sur un site P.
- Ainsi M_1,...,$M_{\ell+1}$ est un zigzag de longueur ℓ de a vers n'importe quel PEE après c sur le même site P (ou c lui même).
- Ainsi, par hypothèse de récurrence,
  - il y a un zigzag de longueur ℓ vers ces PEE et il ne peuvent donc pas être dans un même état global consistant.

Partie 2: Preuve

- Donc, pour que a et b soient dans le même état global
  - il faut qu'il est un PEE d, avant c sur P qui soit causalement indépendant de a et b.
  - Or, dans ce cas là, d est placé sur P avant l'envoi du message $M_{\ell+1}$ ce qui se traduit par d→b, ce qui est incompatible avec ce que l'on cherche

Partie 2: Preuve

- Conclusion:
  - a et b ne peuvent pas être dans un même état global consistant
  - car on ne peut pas trouver de PEE x sur P tel que a$\nrightarrow$x et x$\nrightarrow$b.
  - Ce qui contredit l'hypothèse que a et b pouvaient être dans un même état global consistant
  - ce qui montre l'hypothèse de récurrence et fini de démontrer le théorème.

Conséquences pratiques

- Le théorème précédent est très important pour plusieurs raisons d'un point de vue théorique c'est une caractérisation des états globaux consistants
- En fait, c'**est beaucoup plus que cela.**
- C'est une façon constructive de prolonger **un ensemble de PEE sans zigzag interne vers un état global consistant.**
- La méthode de construction est donné dans la preuve du théorème.

Méthode

- Cette méthode est du type incrémentale
- Puisqu'on prolonge un ensemble de PEE ayant déjà de bonnes propriétés
  - Ici deux à deux sans zigzag
- vers un ensemble contenant d'autres PEE et qui possède encore la bonne propriété

Méthode

- De plus, partant d'un ensemble S, si on veut rajouter des PEE de deux nouveaux sites P et Q
- il suffira de tester les zigzag des PEE de P et Q vers les PEE de l'ensemble S
- mais pas l'indépendance des PEE entre P et Q

Remarque

- Un corollaire intéressant à ce propos est le suivant.
- Il montre que l'on peut partir d'un seul PEE qui ne soit pas un cycle zigzag pour construire un état global consistant.

Corollaire

- Soit a un PEE.
- Il existe un état global consistant contenant a si et seulement si a n'est pas impliqué dans un cycle zigzag.

Lélia Blin

Algorithme pour construire un état global

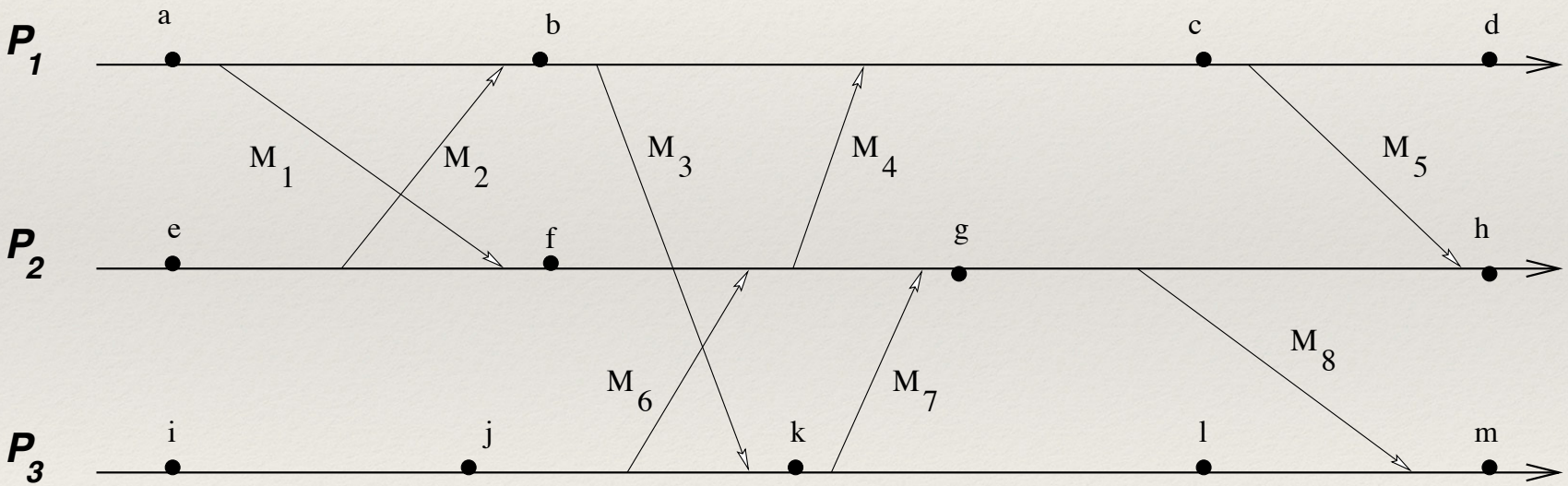
Algorithmique répartie
M1

Méthode

- Voyons comment déterminer par un algorithme s'il existe ou non un zigzag entre deux PEE.
- On suppose ici que l'on dispose de tous les PEE disponibles à la fin de l'exécution

Exemple

- Comment déterminer s'il existe un zigzag du PEE j vers le PEE b?

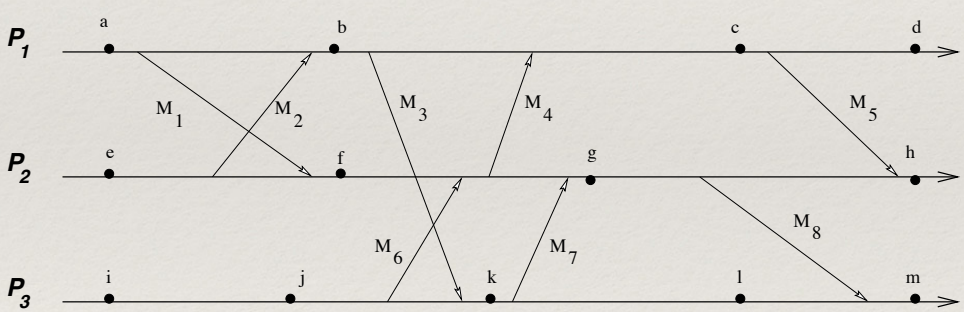


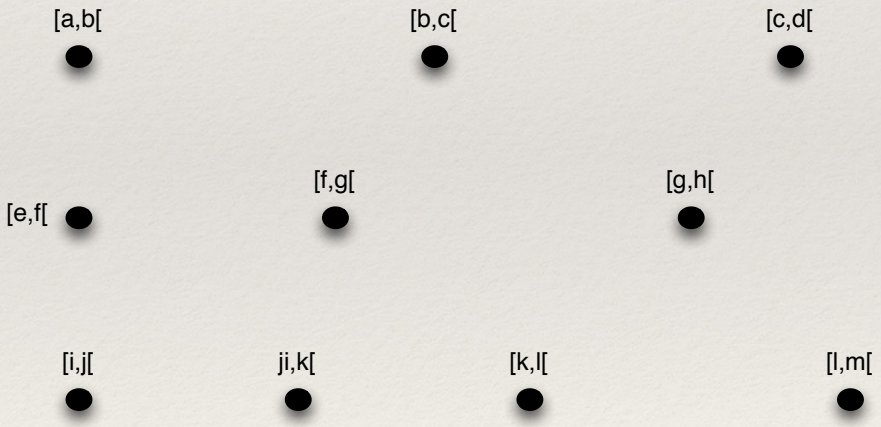
Algorithme

- On va appliquer une transformation du graphe des traces vers un graphe orienté.
- Puis chercher dans ce nouveau graphe s'il existe un chemin d'un certain sommet vers un autre
- Si oui ⇒ un zigzag existe
- Si non ⇒ un zigzag n'existe pas

Graphes des intervalles: sommets

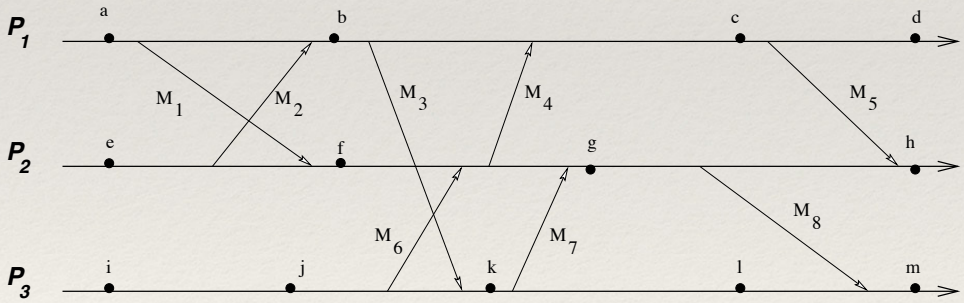
- Pour chaque site P ayant l_p PEE $C_{p,1},\ldots,C_{p,lp}$ créer les l_{p-1} sommets étiquetés $[C_{p,i},C_{p,i+1}[$ pour $i=1,\ldots,l_{p-1}$

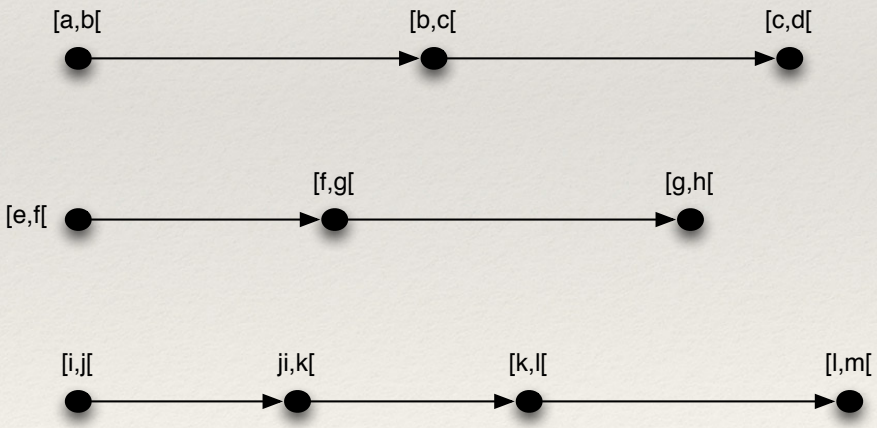




Graphes des intervalles: arcs

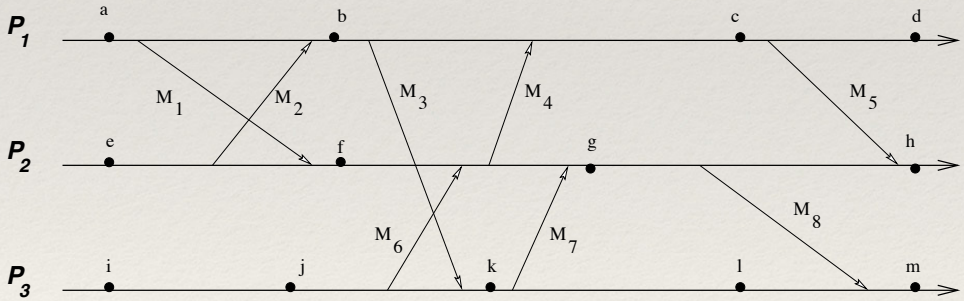
- Les arcs sont formés en appliquant les règles suivantes:
  - Un arc du sommet $[C_{p,i},C_{p,i+1}[$ vers le sommet $[C_{p,i+1},C_{p,i+2}[$.

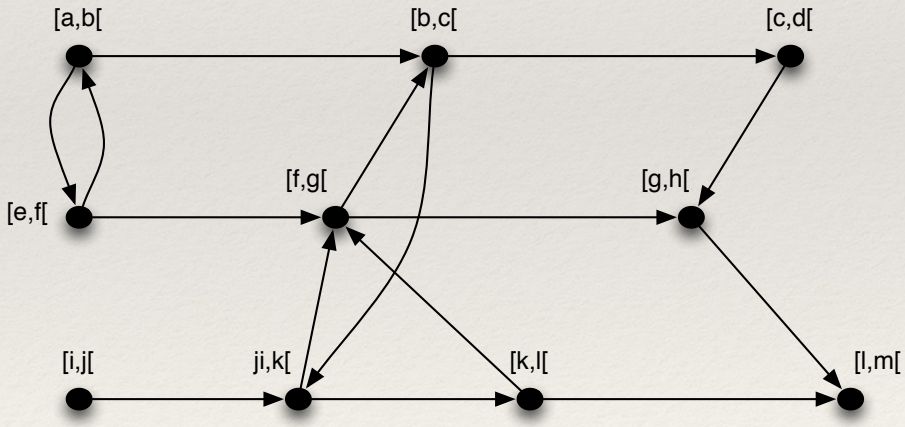




Graphes des intervalles: arcs

- Les arcs sont formés en appliquant les règles suivantes:
  - Un arc du sommet $[C_{p,i},C_{p,i+1}[$ vers le sommet $[C_{Q,j},C_{Q,j+1}[$ si le site P a envoyé (au moins) un message M vers le site Q entre les PEE $C_{p,i}$ et $C_{p,i+1}$ et si le site Q a reçu ce message M entre les PEE $C_{Q,j}$ et $C_{Q,j+1}$.





Graphes des intervalles: zigzag

- Ainsi, il y a un zigzag du PEE $C_{p,i}$ vers le PEE $C_{Q,i}$ si et seulement s'il existe un chemin orienté du sommet $[C_{p,i},C_{p,i+1}[$ vers le sommet $[C_{q,j-1},C_{q,j}[$ dans le graphes des intervalles.
- Cette recherche de chemin peut se faire par n'importe quel algorithme de recherche de chemin entre deux sommets

Algorithme réparti

- L'algorithme construit à la demande d'au moins un site un état global.
- Cet état est créer à la volée, directement
- Il n'y a donc pas besoin de faire d'analyse de traces d'exécution pour en construire un
- Du coup les PEE antérieurs peuvent être effacés (économie de mémoire)

Remarques

- On peut néanmoins noté les points suivants:
  - Cet algorithme nécessite l'utilisation d'un message de contrôle (utilisation de la ressource réseau)
  - Les sites ne sont plus libres de faire des enregistrements quand ils le veulent. Ils sont maintenant dépendants des uns des autres

Modèle

- On suppose
  - le réseau fiable
  - les liens sont FIFO
- Nous allons distinguer
  - le message <MARQUE> du protocole
  - des autres messages échangés par le système que nous appelons message de travail
- Application du protocole de manière atomique

Algorithme

Lorsqu'un site p quelconque veut créer un état global
Enregistrer un PEE;
$Enreg_p := True$;
Pour tout $v_p \in Voisins_p$ faire Envoyer($< MARQUE >$) à v_p;

Algorithme

```
Lors de la réception d'un message < MARQUE > d'un voisin q
 Si (Enreg_p = False) alors faire
      Enregistrer un PEE;
      Enreg_p := True;
      Pour tout v_p ∈ Voisins_p faire Envoyer(< MARQUE >) à v_p;
```

Remarques

- On peut montrer qu'il suffit qu'au moins un site exécute la procédure d'initialisation pour qu'un état global soit construit
- Même si plusieurs sites l'exécutent en même temps le résultat final sera bien un état global

Remarques

- Notons E={C_1,…,C_n} l'ensemble des n PEE produits par cet algorithme.
- E n'est pas consistant en général puisqu'il arrive qu'un PEE C_i soit enregistré sur un site i après l'arrivée d'un message <MARQUE> en provenance de j qui envoyé ce message après avoir enregistré C_j.
  - On a donc $C_j \rightarrow C_i$
- On peut cependant faire abstraction des messages <MARQUE> car ce sont des messages de contrôle qui n'ont pas de rapports direct avec l'activité du réseau sous surveillance.

Lemme

- Cet algorithme produit un état global consistant E dans l'exécution restreinte aux message de travail.