



UNIVERSITÉ DE FRANCHE-COMTÉ

Construction d'un DST : autres propositions

par Christophe ENDERLIN

le 3 octobre 2014

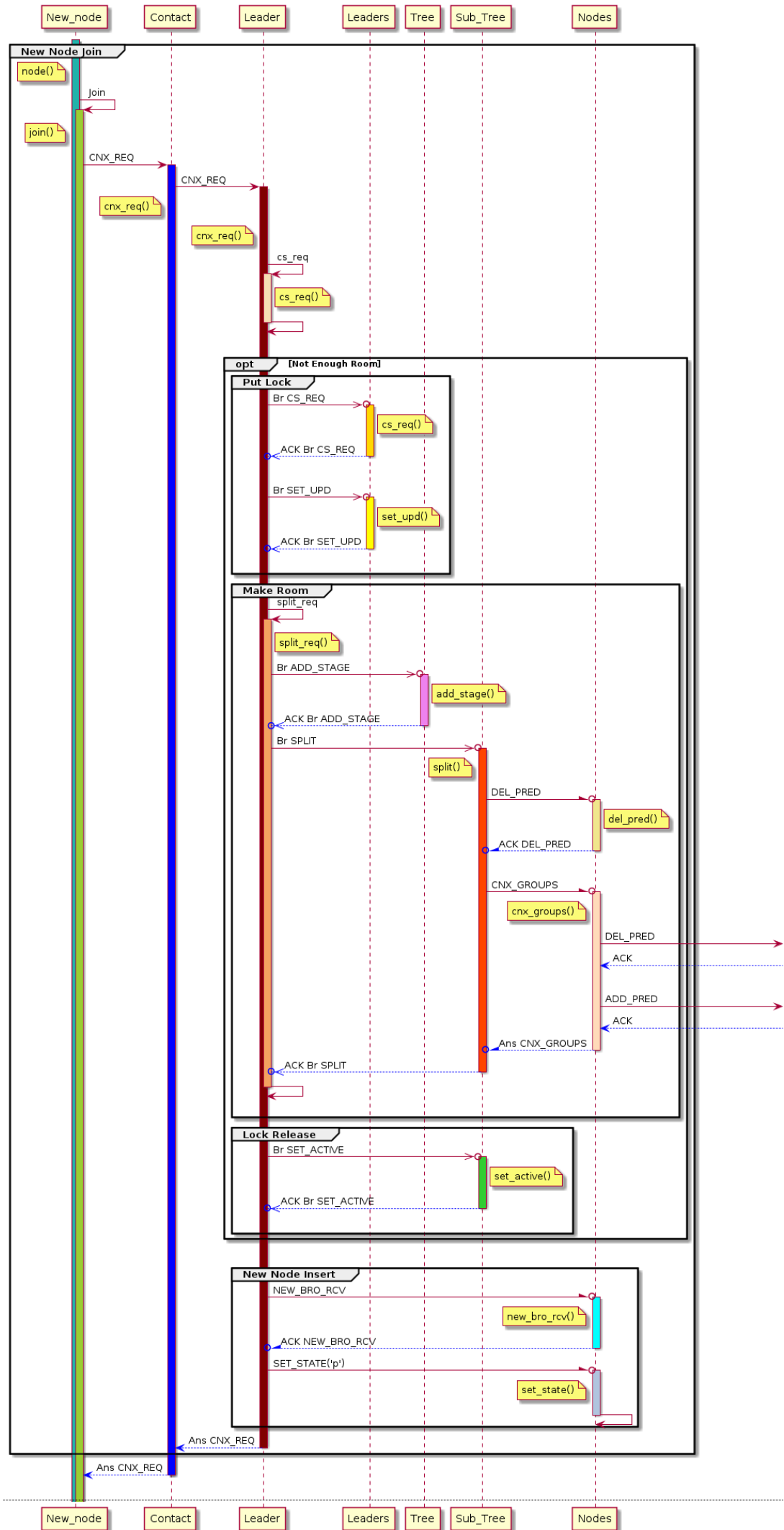
Chapitre 1

Contexte

1.1 Rappel

Pour mémoire, voici une représentation graphique du déroulement des opérations lors de l'insertion d'un nouveau nœud dans le DST, sans équilibrage de charge, pour simplifier. (voir page suivante)

DST : Node Arrival



1.2 Situation

À l'issue de nombreux tests, la première version de mon simulateur a montré son manque de robustesse : lorsqu'un grand nombre de nouveaux arrivants impactent la même zone du DST à un même moment, les choses se passent mal. L'ensemble des problèmes observés semble pouvoir être ramené aux deux points suivants :

1.2.1 Le foisonnement des requêtes/réponses se passe mal

Voici un exemple de cas de figure qui pose problème :

2 nœuds intègrent le DST via un même contact :

- *Node* 14 → *Node* 121 → *Node* 42 (intégration rapide, pas de scission requise)
- *Node* 249 → *Node* 121 → *Node* 42 (scissions et ajout d'étage requis)
- Alors que *Node* 121 est en attente de `ACK_CNX_REQ/249`¹ de 42, il reçoit `ACK_CNX_REQ/14` de 42. Il le stocke donc puisque ce n'est pas la réponse qu'il attend.
- *Node* 14 ne recevant pas la réponse de 121 reste en 'b'. À ce stade de l'intégration, il fait déjà partie du DST et il reçoit donc `ADD_STAGE/249` mais il ne peut y répondre \implies *deadlock*.

Comme on le voit, le fait que `ACK_CNX_REQ/14` soit stocké par 121 pose problème ici.

Un seul process (hébergé par *Node* 121) chargé de l'intégration de deux nouveaux nœuds différents attend deux réponses. Le problème vient alors du fait qu'il ne peut pas traiter les réponses du premier pendant qu'il est occupé avec le deuxième. Ce mécanisme n'est donc pas correct.

On pense alors à deux solutions possibles : *a*) utiliser la fonction `MSG_comm_waitany()` de Simgrid (elle permet de réagir à une réponse parmi celles attendues) *b*) utiliser plusieurs process dédiés, chacun gérant ses propres requêtes/réponses .

La deuxième solution semble permettre de bien séparer les tâches, ce qui aurait un double avantage : arrêter de mélanger les requêtes/réponses pour différents nouveaux arrivants, et présenter une trace d'exécution plus lisible.

1.2.2 L'échec d'une diffusion d'un SET_UPDATE est mal géré

Pour rappel, lorsque l'arrivée d'un nouveau nœud requiert des scissions, on diffuse un `SET_UPDATE` sur l'ensemble du sous-arbre impacté dans le but de placer un verrou (l'état 'u') sur l'ensemble de ses nœuds. Ainsi, les seules requêtes qu'ils accepteront seront celles qui concernent cet ajout (les autres sont soit refusées, soit différées).

1. Autrement dit, un accusé réception d'une requête de demande de connexion pour le nouveau nœud 249

Lorsque cette diffusion échoue (parce qu'on tombe sur une partie déjà verrouillée pour un autre arrivant, par exemple), on se retrouve dans la situation où une partie du sous-arbre a été verrouillée pour le nouveau nœud courant, et une autre pour un autre nouveau nœud. Il est donc important de remettre les choses en état (ôter les verrous posés par la diffusion en échec) pour ne pas bloquer les choses.

Les tests ont montré que la méthode utilisée pour cela n'est pas correcte puisqu'on arrive malgré tout à générer des *deadlocks*, les deux sous-arbres se bloquant mutuellement.

Il faut donc trouver des solutions pour ces deux problèmes.

Chapitre 2

Foisonnement des messages : solution

Principes de base :

- Chaque demande d’insertion de nouveau nœud est traitée par un process distinct créé pour l’occasion
- Ce sous-process ne peut pas recevoir de réponses à des requêtes qu’il n’a pas émises. Dit autrement, il doit (et peut) traiter immédiatement toutes les réponses reçues. (plus besoin de différer ou de refuser)
- Plusieurs nouveaux nœuds utilisant le même contact ne pouvant pas être traités simultanément, un mécanisme de file d’attente pour les traiter séquentiellement est mis en place.

De plus, j’ai fait le choix d’utiliser aussi un tel sous-processus pour les diffusions de **SPLIT** et de **CS_REQ** parce que là encore, cela semble être un avantage de ne pas mélanger les réponses attendues dans le cas de diffusions croisées.

Le schéma général de cette solution à process multiples est présenté figure 2.1 en page 8.

Remarque préliminaire: `launch_fork_process()` est une fonction chargée de choisir comment exécuter les requêtes qui lui sont transmises. Elle les confie soit à un nouveau process (cas de **CNX_REQ**, **BR_SPLIT**, **BR_CS_REQ**), soit au process courant.

Explications : On utilise au plus trois process par nœud.

1. Main_Proc

- il se charge des initialisations et de la création du process **Tasks_Queue** (qui tourne tout le temps de la simulation).
- il héberge les files *tasks_queue* (les tâches **CNX_REQ** reçues en attente de traitement) et *delayed_tasks* (les tâches différées)
- lorsqu’il reçoit une requête, *a*) soit il la place dans la file *tasks_queue* (cas des **CNX_REQ**), *b*) soit il la transmet à `launch_fork_process()`, *c*) soit il la place dans la file *delayed_tasks*.

Process calls sequence

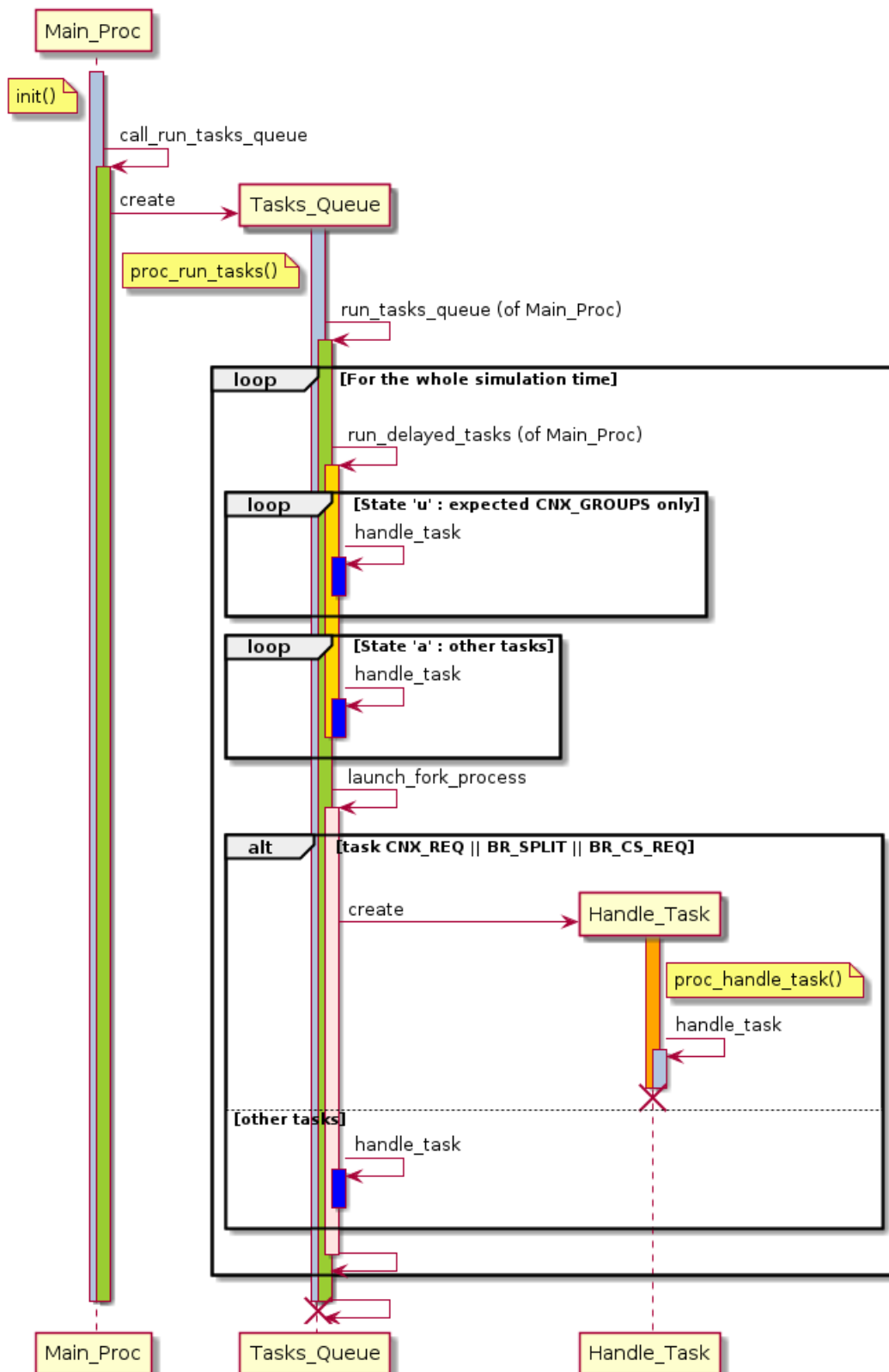


FIGURE 2.1 – Séquence d'appels des sous-process

- il héberge ses propres files *async_answers* et *sync_answers* (les réponses asynchrones et synchrones attendues aux requêtes qu'il a émises)

2. Tasks_Queue

Ce process est chargé d'exécuter la fonction `run_tasks_queue()` qui traite les files *tasks_queue* et *delayed_tasks*. Il héberge aussi ses propres files *async_answers* et *sync_answers*.

3. Handle_Task

C'est ce process qui est éventuellement créé par `launch_fork_process()` depuis `run_tasks_queue()` pour traiter les requêtes concernées. Il héberge aussi ses propres files *async_answers* et *sync_answers*.

2.1 Les fonctions de base

2.1.1 run_tasks_queue()

Voir figure 2.2 en page 10.

La file *tasks_queue* contient toutes les demandes de connection (c'est à dire toutes les tâches CNX_REQ) reçues par le nœud qui héberge cette file. `run_tasks_queue()` est donc chargée de traiter les requêtes présentes dans cette file, par ordre de priorité.¹

Après avoir traité les tâches différées (voir détails `run_delayed_tasks()` plus loin), on s'occupe de la file *tasks_queue* proprement dite, à condition d'être actif (à l'état 'a').

Le nœud courant possède deux variables "de nœud" (c'est à dire globales à tous les process hébergés par ce nœud) :

- *Run_state* : Chaque requête de cette file sera traitée par un process distinct, mais il n'est pas possible ici d'insérer plus d'un nouveau nœud à la fois et il faut donc que ces process s'exécutent séquentiellement. C'est la raison d'être de cette variable *Run_state* : pour s'occuper de la tâche suivante, la tâche courante doit être terminée. (valeur *IDLE*)
- *Last_return* : Cette variable contient la valeur de retour de l'exécution courante de `connection_request()`.

Il y a deux sortes d'échec d'insertion possibles. Lors de l'arrivée d'un nouveau nœud, on a la séquence d'appels suivante : *nouveau nœud* → *contact* → *leader*. En cas d'échec, le contact doit refaire une tentative plus tard, mais son leader peut avoir changé entre temps. Le leader doit donc dépiler la requête alors que le contact doit la conserver dans sa file *tasks_queue*. Deux valeurs de retour en cas d'échec sont alors requises : un leader retourne la valeur *FAILED* alors qu'un contact retournera la valeur *UPDATE_NOK*.

Comme on peut le voir, dans le cas général (i.e. $cpt < MAX_CNX$), une valeur de retour *UPDATE_NOK* laisse la requête dans la file pour une nouvelle tentative un peu plus tard ("*Sleep for a while*"). Dans le cas contraire, on dépile pour passer à la requête suivante. (On voit donc qu'en cas de retour *FAILED*, la requête est bien dépilée.)

1. voir plus loin les remarques à ce sujet

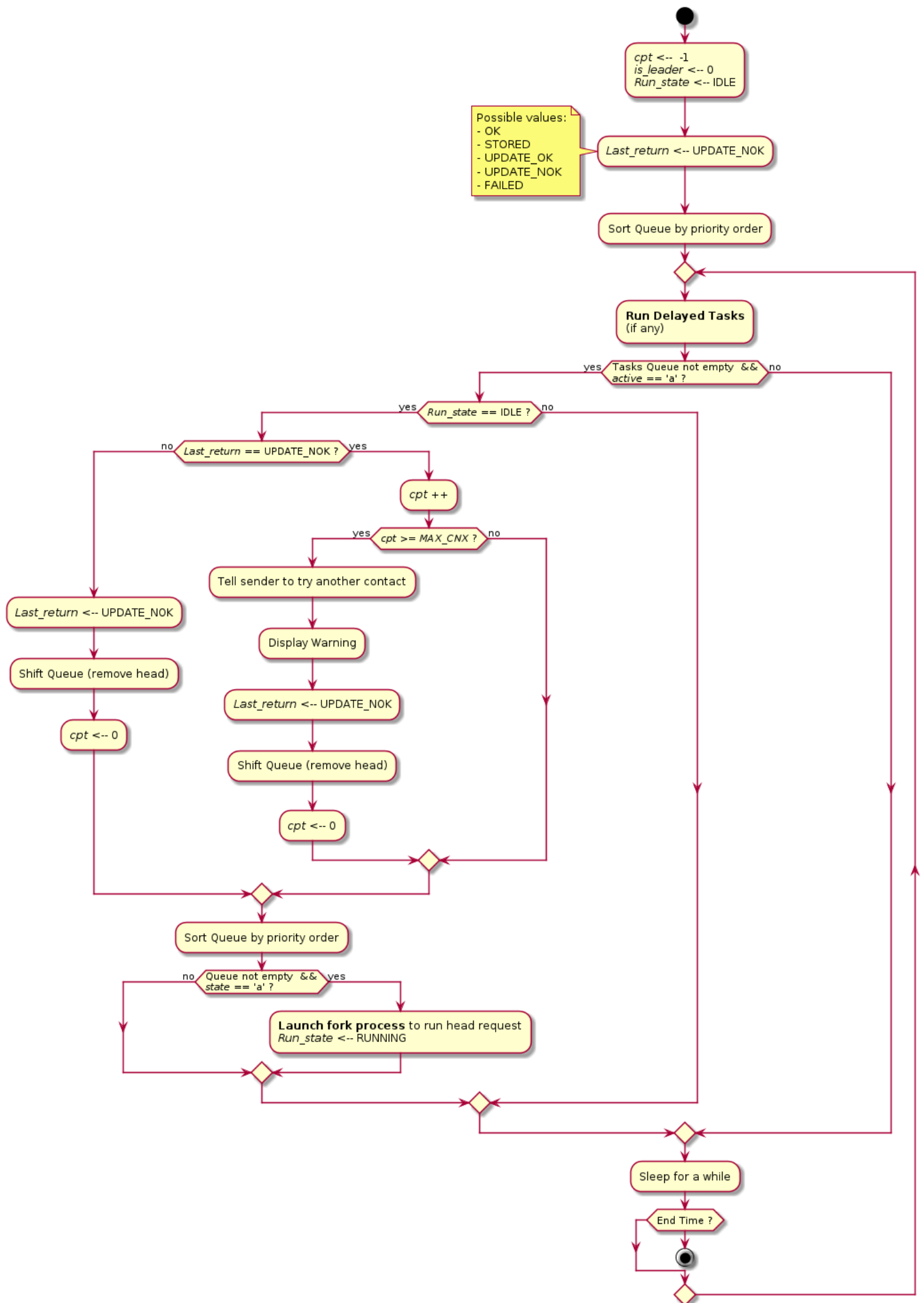


FIGURE 2.2 – run_tasks_queue()

En détails : À l'issue de l'exécution de `connection_request()`, une réponse est envoyée à l'émetteur seulement dans les cas où la requête est ici dépilée.

Lors d'un échec, voici alors ce qu'il se passe dans la séquence de retour *leader* → *contact* → *nouveau nœud* : le leader reçoit *FAILED* comme valeur de retour de `connection_request()`, il dépile donc sa requête et répond *UPDATE_NOK* à son contact. Celui-ci ne dépile rien et ne répond pas à *nouveau nœud*. La requête restant dans la file s'exécutera alors au plus *MAX_CNX* fois.

Si ce nombre est atteint, on décide alors de dépiler tout de même la requête et de répondre *UPDATE_NOK* au nouveau nœud. Il peut ainsi détecter l'échec et refaire d'autres tentatives avec d'autres contacts. (ces contacts sont alors choisis aléatoirement parmi les nœuds déjà intégrés au DST. Il s'agit d'un tableau global, donc introduisant un peu de centralisation dans cet algorithme.²)

Que la file soit dépilée ou pas, elle est à nouveau triée par ordre de priorité à ce moment-là. En effet, pendant le temps d'exécution de `connection_request()`, d'autres demandes d'insertion ont pu arriver dans la file et il faut s'assurer que la prochaine à exécuter soit bien la suivante en termes de priorité.

L'exécution d'une requête de la file est confiée à `launch_fork_process()` (voir plus haut)

2.1.2 `run_delayed_tasks()`

Voir figures 2.3 et 2.4 en pages 12 et 14.

Cette fonction est chargée de traiter la file des tâches différées (*delayed_tasks*). Lorsqu'une tâche ne peut pas être exécutée par un nœud, soit elle est refusée (et retournée à l'émetteur), soit elle est stockée dans cette file pour être exécutée plus tard. Cette fonction est donc appelée périodiquement. (Voir `run_tasks_queue()`)

On distingue deux cas : soit le nœud courant est à l'état 'u', soit il est à 'a'.

état 'u' (voir partie 1 - figure 2.3 en page 12)

Variables :

- *nb_elems* : contient le nombre d'éléments de la file au lancement de `run_delayed_tasks()`
- *cpt* : itérateur (de 0 à *nb_elems*)

Un nouveau nœud est alors en cours d'insertion et il faut examiner ce cas en premier pour minimiser les risques de blocages. Les seules tâches exécutables ici sont celles qui pourraient permettre de terminer cette insertion, c'est à dire les *CNX_GROUPS* pour le même nouveau nœud que celui en cours d'insertion. (c'est le test `task.args.new_node_id == state.new_node_id`)

On parcourt donc la file à la recherche de ces tâches pour les exécuter. (`handle_task(task)`)

2. Voir commentaires en conclusion

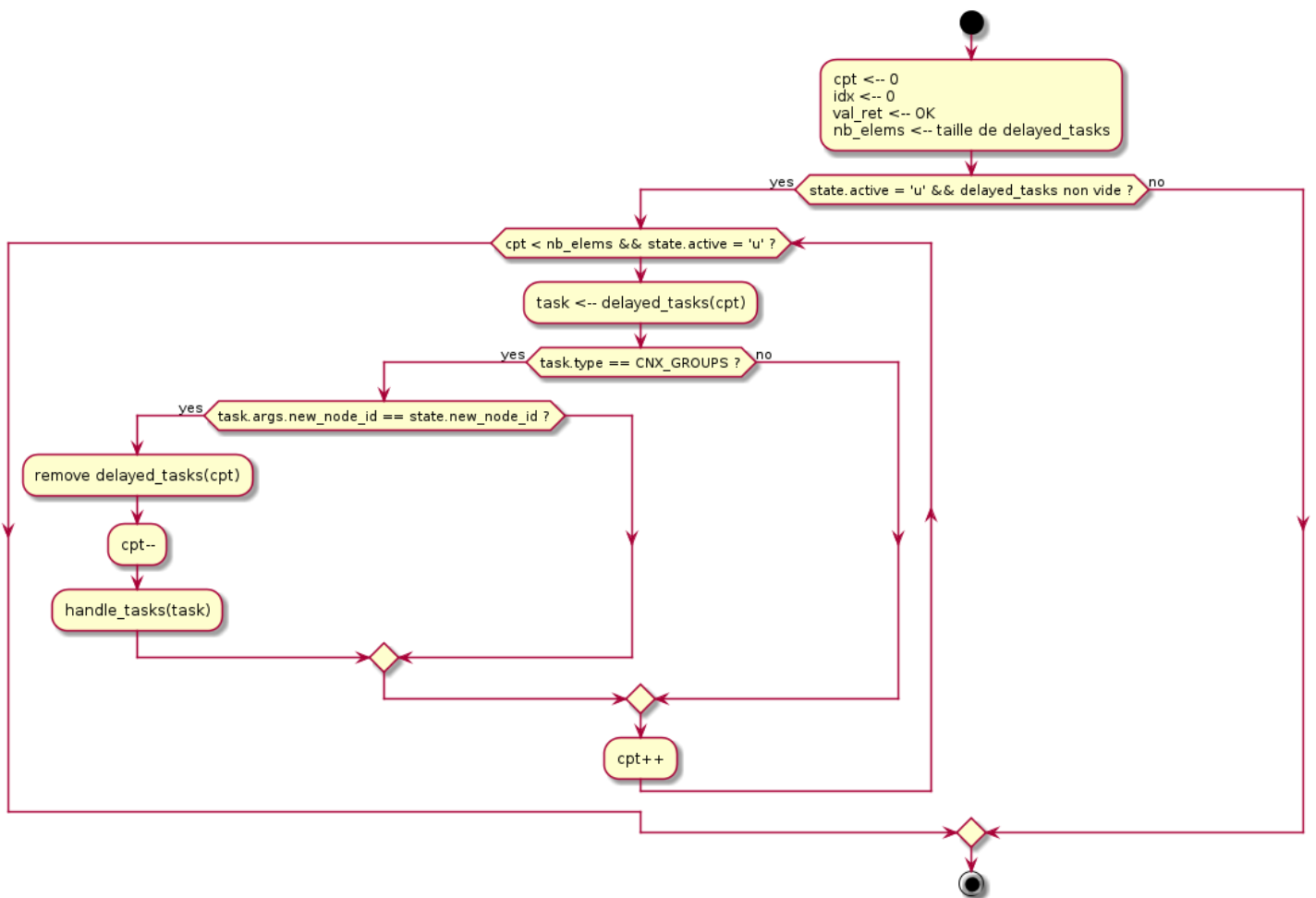


FIGURE 2.3 – Traitement des tâches différées - Partie 1

Lorsqu'on en trouve une, on peut l'ôter de la file sans s'assurer que son exécution a réussi ou pas puisque si elle échoue, elle est à nouveau stockée dans la file.

À noter : à l'issue de la fonction `handle_tasks()`, le nombre de tâches de la file a pu augmenter. Pourtant, on choisit de ne pas mettre à jour la variable `nb_elems` ici pour ne pas risquer une boucle sans fin. Si des tâches ont été ajoutées entre temps, elles seront traitées lors d'une prochaine exécution de `run_delayed_tasks()`.

état 'a' (voir partie 2 - figure 2.4 en page 14)

Ici, le nœud courant est prêt à exécuter n'importe quelle autre tâche et on peut traiter le reste de la file.

Variables :

- `nb_elems` : le nombre d'éléments restants de la file.
- `idx` : itérateur (de 0 à `nb_elems`)
- `is_contact` : vaut 1 si le nœud courant est le contact direct du nouveau nœud. (autrement dit, si l'émetteur de la tâche à exécuter est le nouveau nœud)
- `buf_new_node_id` : `task.args.new_node_id` est mémorisé dans cette variable parce qu'on en a besoin après la destruction de `task`.

On ôte la tâche courante de la file dans les cas suivants :

- l'exécution a réussi (*OK* et *UPDATE_OK*)
- la tâche a été stockée à nouveau (*STORED*)
- l'exécution a échoué mais le nœud courant n'est pas le contact direct du nouveau nœud. (*UPDATE_NOK* && *!is_contact*) Il s'agit du cas *FAILED* décrit plus haut.

À noter : si le nœud courant était verrouillé pour le même nouveau nœud que celui dont la tâche vient d'échouer, alors on ôte le verrou.

À l'issue de l'exécution de cette boucle (`idx == nb_elems`), la file n'est pas forcément vide. Elle contient toutes les tâches dont l'exécution a échoué et celles qui ont été stockées – y compris par d'autres process – entre temps. Cette boucle est alors à nouveau exécutée jusqu'à ce que `nb_elems` soit à 0, c'est dire jusqu'au succès de toutes les tâches qui se trouvaient dans cette file au moment de cet appel de `run_delayed_tasks()`. Comme indiqué précédemment, les tâches ajoutées entre temps ne seront pas traitées lors de cette exécution.

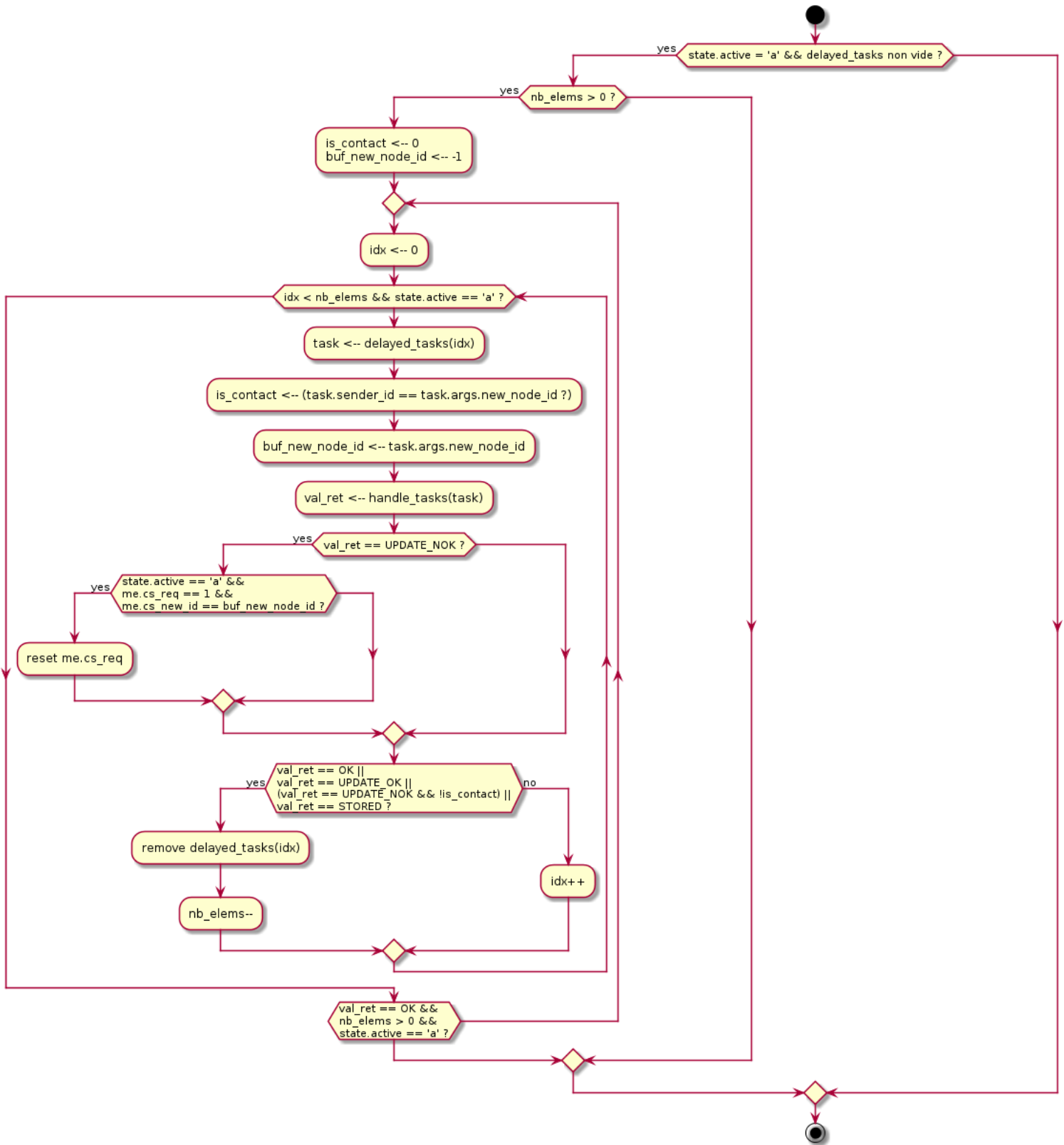


FIGURE 2.4 – Traitement des tâches différées - Partie 2

2.1.3 launch_fork_process()

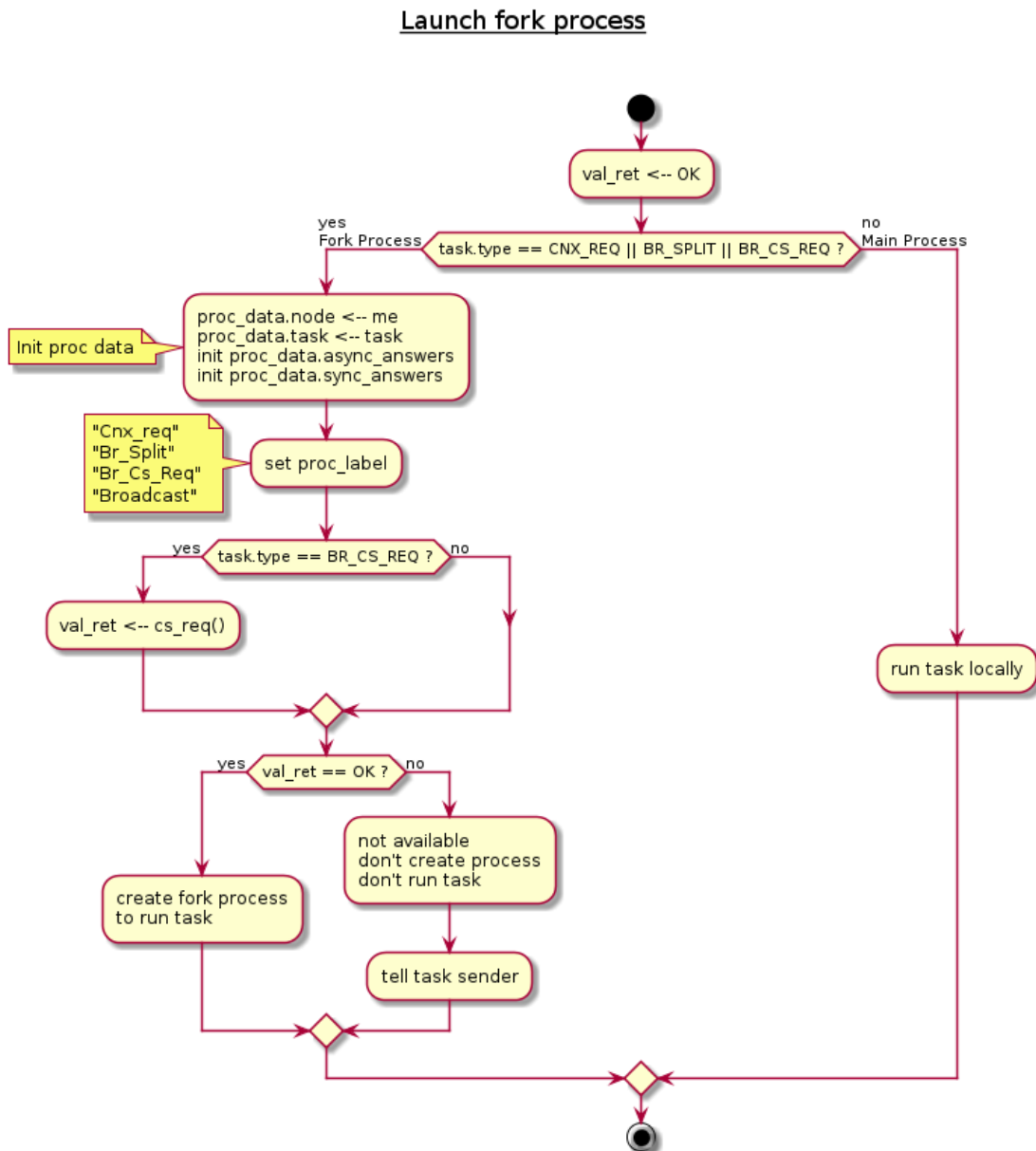


FIGURE 2.5 – Launch Fork Process

Si la tâche transmise à cette fonction n'est ni une demande de connexion (`CNX_REQ`), ni une diffusion de `SPLIT` ou de `CS_REQ`, alors elle est exécutée localement, c'est à dire par le process courant. Sinon, elle est confiée à un nouveau process créé pour l'occasion.

Ce nouveau process doit posséder ses propres files d'attente de réponses attendues (*async_answers* et *sync_answers*) de sorte qu'elles ne puissent plus être mélangées avec celles requises pour l'insertion d'un autre nouveau nœud.

Il est labelisé ainsi : `xxx-nom_process-yyy` où :

- `xxx` est l'id du nœud courant
- `nom_process` est le nom attribué par cette fonction (tel que `"Cnx_req"` ou `"Br_split"`, etc)
- `yyy` est l'id du nouveau nœud en cours d'insertion

On s'assure ainsi de l'unicité de l'étiquette de ce process.

Dans le cas où la tâche est une diffusion de `CS_REQ` (demande d'entrée en section critique), on commence par s'assurer que le nœud courant est disponible (c'est à dire répondrait favorablement à un `CS_REQ`) pour ne pas créer de process fils inutilement. En cas de non-disponibilité, il faut en informer l'émetteur de la requête.

2.2 Les fonctions de communication

2.2.1 envoi synchrone - `send_msg_sync()`

La requête synchrone est utilisée lorsqu'une réponse est attendue pour poursuivre l'exécution du programme.

Cas d'emplois principaux :³

- `CNX_REQ` (fonction `join()`) La réponse attendue est la table de routage du contact qui a réussi l'insertion du nouveau nœud
- `BROADCAST` (fonction `handle_task(): BROADCAST`) La réponse attendue ici est la valeur de retour de la fonction `handle_task()`, c'est à dire le succès ou l'échec de la requête diffusée.

De plus, on utilise aussi ce type de requête pour retransmettre au leader les requêtes `CNX_REQ` et `SPLIT_REQ`.⁴

L'attente de la réponse ne devant pas être bloquante, toutes les requêtes ou autres réponses arrivant dans l'intervalle doivent être traitées.

Déroulement simplifié des opérations

L'émetteur de la requête appelle `send_msg_sync()` qui exécute les tâches suivantes :

- crée la requête
- encapsule la requête dans une tâche
- empile la requête sur `sync_answers`
- envoie la tâche au destinataire

3. Pour simplifier, la phase d'équilibrage de charge est ignorée ici.

4. Pas forcément utile dans le cas de `SPLIT_REQ`, à voir

- attend la réponse
- traite tout ce qui est reçu qui n'est pas la réponse
- dès que la réponse est reçue, dépile la requête de *sync_answers*
- retourne la réponse à l'appelant

Dans le détail

La fonction `send_msg_sync()` est chargée d'envoyer une requête à un autre nœud, puis d'en attendre la réponse.

Chaque process hébergé par un nœud possède deux files : *sync_answers* et *async_answers*.

Ces deux files servent à stocker les requêtes synchrones et asynchrones (respectivement) émises, le temps qu'elles reçoivent une réponse et qu'elle soit lue. Ces requêtes sont dépilées dès que leur réponse a été prise en compte.⁵

Lorsqu'une réponse est reçue, il y a trois possibilités :

1. il s'agit de **la réponse attendue** : Dans ce cas, la réponse est prise en compte, la requête correspondante est dépilée et le traitement se poursuit.
2. il s'agit d'**une autre réponse attendue** : Elle est alors enregistrée dans la bonne pile, avec la requête correspondante, et on se remet en écoute.
3. il ne s'agit d'**aucune réponse attendue** : Elle est simplement ignorée (cas de certains accusés réception, par exemple)

Première partie envoi de la requête (Voir figure 2.6 - page 18)

Après avoir construit la tâche contenant la requête, celle-ci est envoyée au moyen de la fonction `isend()` de Simgrid, puis stockée dans la pile *sync_answers* du process courant. On attend ensuite la fin de la communication.

Deux traitements différents des cas d'échecs :

- la communication ne se termine pas (cas du **process** destinataire arrêté) ou elle a le statut `MSG_TRANSFER_FAILURE` (cas de l'**hôte** destinataire arrêté) : La fonction s'arrête en retournant une erreur de transmission.
- la communication se termine avec le statut `TIMEOUT` (cas d'un destinataire trop occupé, par exemple) : On recommence l'émission un certain nombre de fois (*loop_cpt*). Si *max_loops* est atteint, la fonction est également arrêtée avec une erreur de transmission.

5. Pour mémoire, la réponse à une requête synchrone contient les données demandées alors que la réponse à une requête asynchrone est simplement un accusé réception.

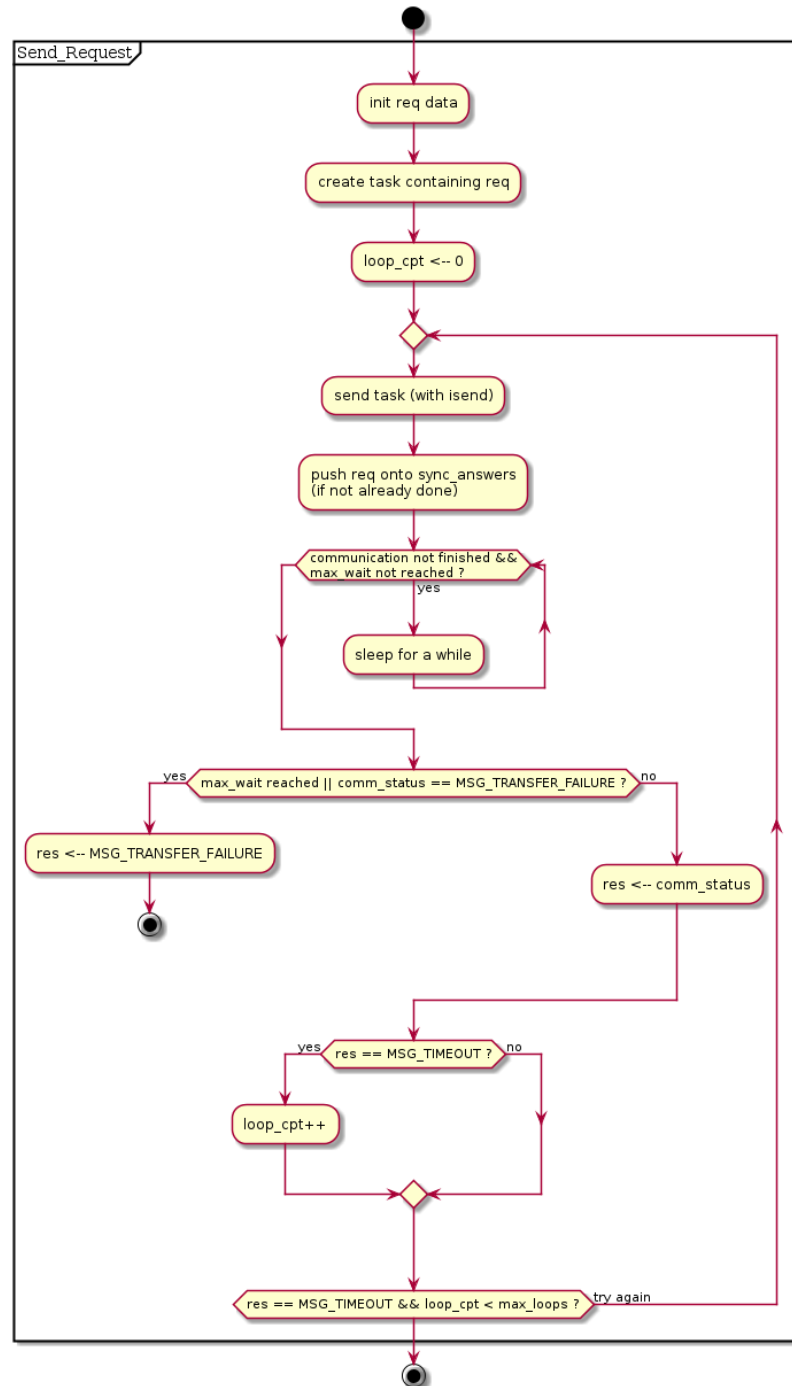
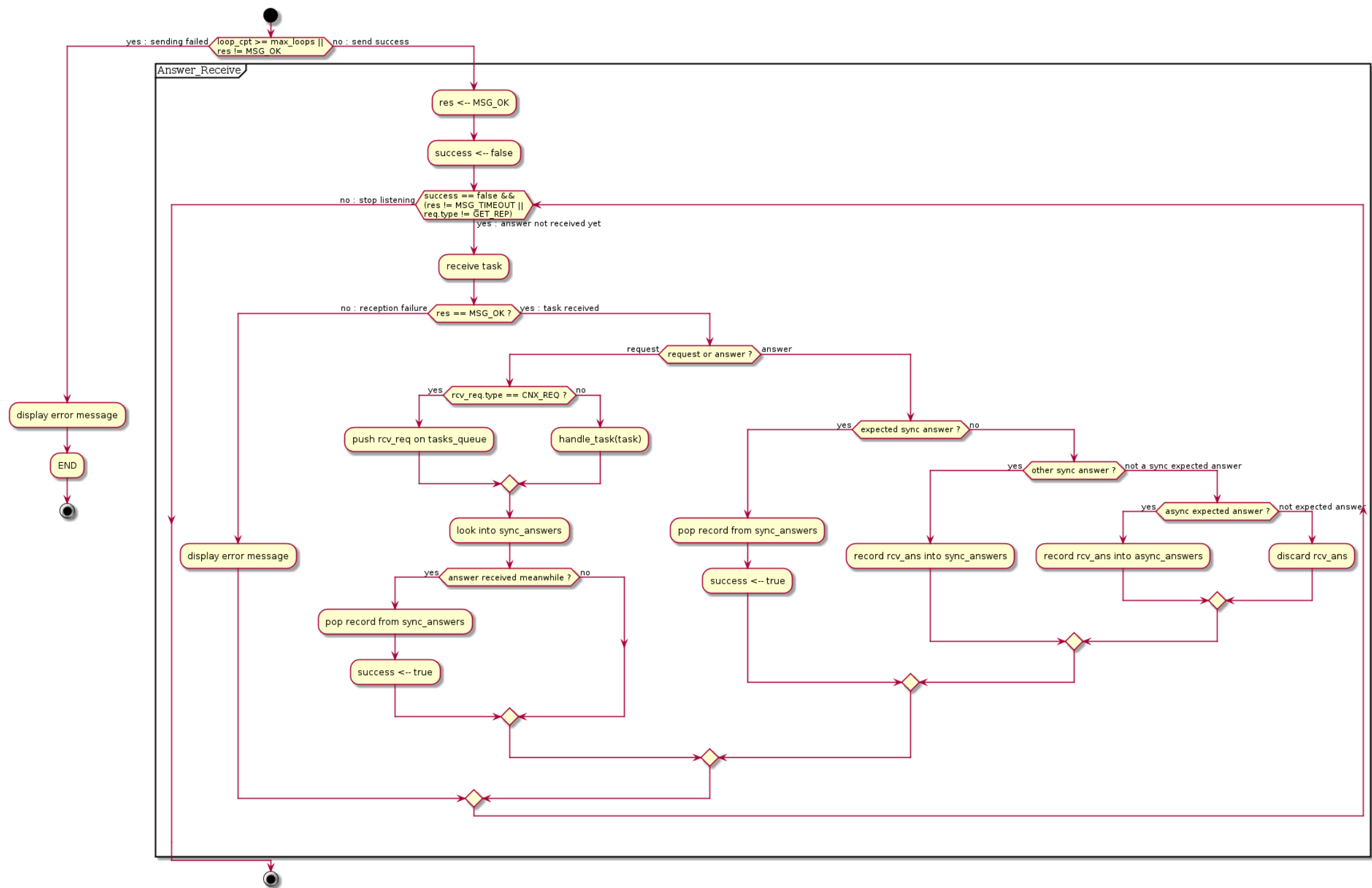


FIGURE 2.6 – Algo `send_msg_sync` - Partie 1

Deuxième partie réception de la réponse (Voir figures 2.2.1 et 2.7 - pages 19 et 20)

Dans cette partie, le process courant se met en écoute de la réponse. En cas d'erreur de communication (`res != MSG_OK`), on la signale et on se remet en écoute, sauf s'il s'agit d'un `TIMEOUT` pour une requête `GET_REP`.⁶ Dans ce cas, on arrête la fonction là en retournant simplement l'erreur pour la signaler à l'appelant.

6. `GET_REP` demande à un nœud de fournir un autre représentant moins chargé pour un étage donné. S'il met trop de temps à répondre et qu'on abandonne, ça n'a pas d'autre conséquence que de conserver le même représentant. On évite ainsi un éventuel *deadlock*



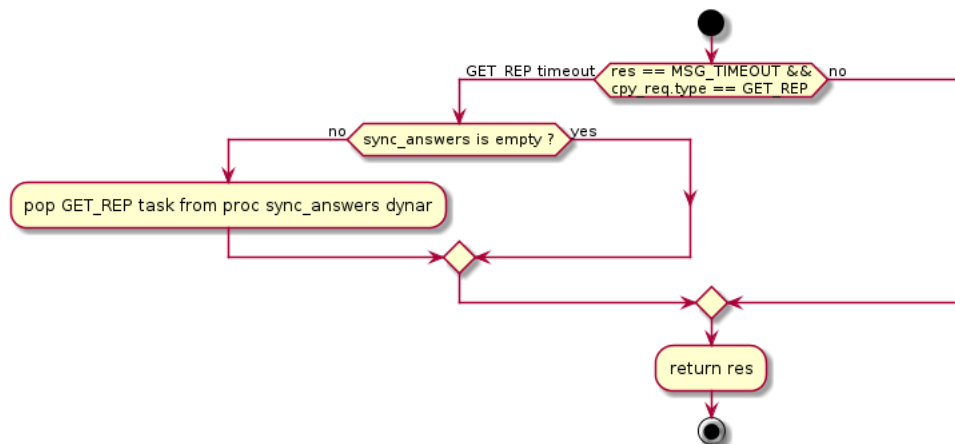


FIGURE 2.7 – Algo `send_msg_sync` - Partie 3

Deux types de tâches peuvent être reçus : soit une requête, soit une réponse.

Requête : Si c'est une demande de connexion (`CNX_REQ`), elle est simplement stockée dans `tasks_queue`.⁷ Sinon, on l'exécute avec `handle_task()`.

Ensuite, on regarde dans la pile `sync_answers` pour voir si la réponse attendue n'a pas été reçue entre temps. Si oui, elle est dépilée et on quitte la boucle de réception. Sinon, on se remet en écoute.

Réponse : On commence par regarder s'il s'agit de la réponse synchrone attendue. Si oui, elle est dépilée et on quitte la boucle de réception. Sinon, il peut s'agir d'une autre réponse attendue (synchrone ou asynchrone). Autrement dit, la requête correspondante à cette réponse est trouvée dans une des deux piles `sync_answers` ou `async_answers`. Dans ce cas, la réponse est enregistrée avec sa requête dans sa pile et on se remet en écoute. S'il ne s'agit d'aucune réponse attendue, on peut l'ignorer simplement et se remettre en écoute.

Remarques

1. Dans cette fonction, la requête en attente de réponse se trouve toujours au sommet de la pile du fait de ces trois éléments :
 - un process donné ne peut exécuter qu'une seule instance de `send_msg_sync()` à la fois
 - `send_msg_sync()` est la seule fonction susceptible d'empiler une requête dans `sync_answers`
 - une pile `sync_answers` n'est associée qu'à un seul process.

De ce fait, lorsqu'on reçoit une réponse à une requête synchrone attendue, il est facile de déterminer s'il s'agit de la réponse attendue ou d'une autre : la requête est au sommet de la pile ou pas.

7. On rappelle que cette file est hébergée par le nœud courant.

2. Si `send_msg_sync()` n'est pas appelée par le process principal, elle ne peut pas recevoir d'autre message que celui qui est attendu.
-

2.2.2 envoi asynchrone - `send_msg_async()`

Dans ce type de requête, le destinataire renvoie un simple accusé réception à l'émetteur une fois la requête demandée effectuée. Ces accusés réception seront pris en compte ou pas selon les besoins de synchronisation.

Dans les deux cas, les opération se déroulent de cette façon : L'émetteur appelle `send_msg_async()` qui exécute les tâches suivantes :

- crée la requête
- encapsule la requête dans une tâche
- envoie la tâche au destinataire

Puis deux cas se présentent :

Un accusé réception est attendu Cas des requêtes suivantes :

- BROADCAST (fonction `broadcast()`)
- NEW_BROTHER_RCV (fonction `connection_request()`)
- DEL_PRED (fonctions `connectSplittedGroups()`, `split()`)
- ADD_PRED (fonction `connectSplittedGroups()`)
- CNX_GROUPS (fonction `split()`)

L'émetteur empile alors la requête sur `async_answers` puis, lorsque plusieurs requêtes ont été émises, il appelle `wait_for_completion()` pour arrêter le déroulement du programme jusqu'à ce que l'ensemble des accusés réception ait été reçu.⁸ C'est cette fonction qui dépile les requêtes de `async_answers` au fur et à mesure de leur réception et comme `send_msg_sync()`, elle traite aussi tout ce qu'elle reçoit entre temps.

Pas d'accusé réception Dans ce cas, l'appelant ne fait qu'exécuter `send_msg_async()`.

`send_msg_async()` en détails *****

Cette fonction réalise la même chose que la première partie de `send_msg_sync()`, à la différence qu'on n'empile pas la requête puisqu'on laisse le soin à l'émetteur de décider s'il faut le faire ou pas.

Une autre différence est qu'on ignore un éventuel défaut d'émission (`res != MSG_OK`) alors que le programme est arrêté dans `send_msg_sync()`. Ce n'est pas volontaire, je n'ai juste pas encore étudié la tolérance aux pannes.

`wait_for_completion()` en détails *****

8. voir détails de `wait_for_completion()` plus loin

On attend *ans_cpt* réponses. Si *max_wait* est atteint avant que toutes les réponses aient été reçues et traitées, on quitte sur erreur, ce cas ne devant pas se produire.

`wait_for_completion()` pouvant s'appeler elle-même, il faut commencer par vérifier si des réponses attendues ont déjà été traitées par ces appels récursifs, et si oui, s'il en reste encore à traiter. C'est l'appel à `check_async_nok()` du début.

Si oui, on se met en écoute, et comme dans `send_msg_sync()`, il faut traiter tout ce qui est reçu dans l'intervalle.

- Requête : *****

Si c'est `CNX_REQ`, on l'empile sur *tasks_queue*, sinon, on l'exécute avec `launch_fork_process()`. On regarde ensuite si des réponses ont été reçues entre temps avec `check_async_nok()`, puis on se remet éventuellement en écoute.

- Réponse : *****

- Réponse asynchrone correspondant à une requête empilée : - si elle fait partie de celles qui sont attendues, la requête correspondante est dépilée d'*async_answers*, après avoir mémorisé un éventuel échec de `BROADCAST` ou `SET_UPDATE` dans *ret* (pour du log) On se remet ensuite en écoute s'il reste des réponses à recevoir. (`ans_cpt > 0`)

- sinon, elle est enregistrée dans *async_answers*

- Réponse synchrone attendue : - si oui, elle est enregistrée dans *sync_answers* - sinon, elle est ignorée.

Chapitre 3

Échec d'une diffusion d'un verrou : solution

Parler du mécanisme de CS_REQ.