MAF : un Protocole de Multicast Fiable

Prométhée Spathis et Kim L. Thai

Résumé– Cet article décrit la conception et la spécification de MAF, un nouveau modèle de communication point à multipoint actif, totalement fiable et adapté à des groupes de diffusion de grande taille. MAF utilise la technologie des Réseaux Actifs : les routeurs actifs de l'arbre de diffusion sont programmés pour maintenir une copie des données diffusées afin de pouvoir en réparer les pertes éventuelles. MAF s'affranchit des contraintes liées au facteur d'échelle en organisant les routeurs ainsi modifiés selon une structure hiérarchique arborescente, obtenue en divisant l'arbre de diffusion en sous-arbres. Une approche orientée émetteur est adoptée au sein de chaque sous-arbre. MAF a la particularité de s'accommoder de routeurs actifs à capacités de stockage finies, en agrégeant les acquittements positifs. Cet article décrit également l'implantation du protocole MAF sur la plate-forme nationale du projet RNRT Amarrage.

Motsclés–Communication de groupe, fiabilité totale, réseaux actifs, structuration hiérarchique, ACK agrégés, capacités de stockage finies.

***Abstract*— This paper describes the design and implementation of a novel reliable multicast protocol, totally reliable and scalable to large number of receivers. MAF relies on Active Networks technology: active routers in the multicast tree store sender's transmissions in order to be able to later retransmit them to repair downstream losses. To address scalability, MAF organizes those active routers into a hierarchical structure obtained by dividing the multicast tree into subtrees. Since a sender initiated approach is used within each of those subtrees, MAF has the particularity of operating correctly with finite buffers. This paper also describes the implementation of MAF over the active network platform deployed by the RNRT project AMARRAGE.**

***Index Terms*— totally reliable multicast, active networks, hierarchical structure, aggregated ACK, finite buffers.**

I. Introduction

Un service de communication point à multipoint offre un moyen *efficace* de diffuser des unités de données à un groupe de récepteurs, en ce sens qu'une seule copie de chacune de ces unités est envoyée. Cependant, au cours de la diffusion des données, certains récepteurs peuvent enregistrer des pertes, c'est-à-dire ne pas recevoir l'intégralité des données qui leur sont destinées. Ces pertes sont le plus souvent dues à de la congestion chez l'émetteur ou dans le réseau, et moins fréquemment à des erreurs de routage. Les applications multiparties fondées sur la diffusion fiable d'informations (ftp multidestinataire, mise à jour d'informations dupliquées réparties, tableau blanc, simulations distribuées, etc.) pouvant faire intervenir un nombre de récepteurs de l'ordre de plusieurs centaines de milliers, la conception de protocoles de diffusion fiable efficients doit tenir compte des contraintes imposées par la résistance au facteur d'échelle, telles que l'implosion de la source ou l'exposition des récepteurs à des retransmissions inutiles.

Avec l'apparition du paradigme des Réseaux Actifs, plusieurs protocoles de diffusion fiable, basés sur le soutien des routeurs de l'arbre de diffusion, ont récemment été proposés [KAS 00a], [LEH 98], [PAP 98], [SPE 01]. L'objectif de ces protocoles, que nous qualifierons d'actifs, est d'offrir une solution satisfaisante du point de vue de la résistance au facteur d'échelle. Ces protocoles diffèrent des approches de bout en bout traditionnelles par le rôle qu'ils attribuent aux routeurs de l'arbre de diffusion. En plus de dupliquer les unités de données émises par la source à l'ensemble du groupe de réception, ces routeurs sont modifiés pour participer aux mécanismes de recouvrement sur erreur. L'ensemble des routeurs ainsi modifiés que nous appellerons routeurs de réparation, est organisé selon une structure hiérarchique arborescente (multi-niveaux). Cette structure est obtenue en divisant en sous-groupes, l'arbre logique formé des routeurs de réparation de l'arbre de diffusion et des récepteurs du groupe de réception. Si ces protocoles minimisent les risques d'implosion et dispensent la source et les routeurs de l'arbre de diffusion de connaître l'ensemble de leurs fils, ils requièrent néanmoins, pour garantir une fiabilité totale, des capacités de stockage infinies au niveau de la source, voire même des routeurs de réparation.

Cet article décrit la spécification et la conception de MAF (Multicast Actif Fiable), un nouveau modèle de communication point à multipoint actif totalement fiable et adapté à des groupes de diffusion de grande taille. MAF se distingue des protocoles actifs existants par sa capacité à s'accommoder de routeurs actifs disposant de capacités de stockage finies. Notre approche s'affranchit des contraintes liées au facteur d'échelle en faisant intervenir les routeurs actifs de l'arbre de diffusion dans les mécanismes de recouvrement sur erreur et de réduction du retour d'information.

L'article s'organise de la manière suivante. Dans le second paragraphe, nous développons la problématique liée à la fiabilisation totale des communications de groupe et présentons son impact sur les choix de conception de MAF. Le paragraphe 3 décrit la spécification et la conception de MAF. Dans le paragraphe 4, nous détaillons l'implantation d'un prototype déployé à l'échelle géographique de la France. Le paragraphe 5 passe en revue les protocoles de diffusion fiable récemment proposés et qui exploitent un soutien des routeurs de l'arbre de diffusion. Dans le dernier paragraphe, nous présentons nos conclusions et donnons les perspectives de nos travaux.

Une partie de ce travail est réalisée dans le cadre du projet RNRT Amarrage Convention No 01.2.93.0128

Laboratoire d'Informatique de Paris 6 (LIP6), Université Pierre et Marie Curie, No 4, place Jussieu, 75005 Paris, France

E-mails : promethee.spathis,kim.thai@lip6.fr

II. Problématique

Les protocoles reposant sur un soutien des routeurs de l'arbre de diffusion distribuent le rôle de réparation des pertes le long de l'arbre de diffusion. En plus de la source, d'autres entités de l'arbre de diffusion se répartissent ainsi la réparation des pertes. L'ensemble de ces entités est organisé selon une structure hiérarchique arborescente qui résulte du découpage de l'arbre de diffusion en sous-arbres : à chaque sous-arbre est affectée une entité responsable des feuilles correspondantes. Tous ces protocoles adoptent, au sein de chaque sous-arbre, une approche orientée récepteur ; en effet, la détection des pertes y repose uniquement sur l'utilisation d'acquittements négatifs (NACK) et de temporisateurs de retransmission.

Si l'utilisation locale d'une telle approche permet de limiter les risques d'implosion de la source et dispense les entités chargées des réparations de connaître l'ensemble de leurs fils, le temps nécessaire pour détecter une perte n'est pas borné. Aucun mécanisme explicite ne permet à la source et aux entités de réparation de libérer leurs mémoires : une entité de réparation ne peut jamais être sûre que les feuilles de son sous-arbre ont correctement reçu les données diffusées ou encore que leurs NACK n'ont pas été perdus. Pour garantir une fiabilité totale, les entités chargées de réparer les pertes doivent théoriquement être en mesure de maintenir indéfiniment les données diffusées. Les protocoles existants supposent donc de la source et des entités de réparation des capacités de stockage infinies. En effet, il est montré dans [LEV 96] que les approches orientées récepteur basées uniquement sur l'utilisation de NACK ne fonctionnent correctement qu'en présence de capacités de stockage infinies.

Or, les entités disposent en pratique de capacités de stockage limitées. En l'absence de mémoire suffisante, une entité peut être amenée à remplacer des unités de données sans avoir la garantie que les membres du sous-groupe dont elle est responsable, les ont correctement reçues. Pour réparer les pertes des unités de données qu'elle a dû remplacer, l'entité doit alors s'en remettre aux entités situées en amont dans l'arbre de contrôle et en dernier recours, à la source. Dans ce cas, c'est la source qui doit disposer de capacités de stockage infinies pour que la diffusion des données soit garantie totalement fiable.

Nous pouvons noter que les politiques de remplacement des unités de données en mémoire étudiées dans [KAS 00b] n'écartent pas le risque qu'une entité remplace une unité de données avant qu'elle ait été correctement reçue en aval. En instanciant ces politiques, les entités de réparation minimisent le nombre de NACK qu'elles ne pourront satisfaire, faute d'avoir remplacé des unités de données perdues en aval.

MAF, le protocole que nous présentons dans cet article, garantit un service de diffusion totalement fiable, tout en consommant une quantité finie des capacités de stockage des entités de réparation utilisées.

III. Le protocole MAF

A. Vue générale

MAF exploite la flexibilité qu'introduit le paradigme des Réseaux Actifs [CLA 98] pour modifier le comportement des routeurs de l'arbre de diffusion. En effet, MAF rend un service de diffusion fiable performant et résistant au facteur d'échelle en définissant des mécanismes de recouvrement d'erreur que la technologie des Réseaux Actifs permet de déployer et d'implanter dynamiquement dans les routeurs de l'arbre de diffusion. Nous faisons l'hypothèse que MAF est exécuté au-dessus d'un réseau non fiable, où une unité d'information peut être perdue, dupliquée, retardée ou déséquencée. MAF repose sur l'existence d'un protocole de routage multidestinataire qui construit et maintient un arbre de diffusion optimal dont la racine est la source.

MAF organise les récepteurs du groupe de diffusion et les routeurs actifs de l'arbre de diffusion selon une structure hiérarchique arborescente. Cette structure résulte du découpage en sous-arbres de profondeur 1, de l'arbre logique constitué par les routeurs actifs de l'arbre de diffusion et par les récepteurs du groupe de diffusion. La source et l'ensemble des routeurs actifs de l'arbre de diffusion comptent au plus D feuilles dont ils connaissent l'identité. Nous dirons de ce fait que la source et les routeurs actifs se répartissent la connaissance du groupe de diffusion. Cette limite D est définie à la création du groupe de diffusion. Passée en argument au protocole de routage multidestinataire, elle permet de réduire la charge de traitement de la source et des routeurs actifs de l'arbre de diffusion.

A la racine des sous-arbres ainsi construits, se trouve la source ou un routeur actif, alors que leurs feuilles peuvent être aussi bien des récepteurs que des routeurs actifs, eux-mêmes racines de sous-arbres de niveau inférieur. L'ensemble des feuilles d'un même sous-arbre constitue ce que nous appellerons un *sous-groupe* dont est responsable la racine du sous-arbre. En effet, MAF distribue le rôle de réparation des pertes constatées au sein d'un sous-groupe, à la racine du sous-arbre correspondant. Nous dirons de ce fait qu'une racine de sous-arbre est le *chef* du sous-groupe constitué de ses feuilles. Un chef joue le rôle qu'attribuent à la source les approches de bout en bout orientées émetteur : il assume, comme nous le verrons en détail plus tard, la détection et la réparation des pertes qui affectent le sous-groupe dont il est responsable. Pour réaliser ces fonctions, un chef doit connaître la composition du sous-groupe dont il est responsable.

Ainsi, un chef de sous-groupe est soit la source soit un routeur actif programmé pour agir en tant que source aux yeux de ses fils. Aux fonctions de la source, un routeur actif de l'arbre de diffusion ajoute celles des récepteurs : il retourne des acquittements positifs (ACK) à son propre chef, tout en étant responsable des chefs de niveau inférieur. Comme dans les approches orientées émetteur, la réception d'un ACK permet à un chef de détecter les éventuelles pertes. De plus, la réception d'un ACK agrégé constitue une indication explicite de libération de ses mémoires.

Une fois le groupe de diffusion établi, la source émet les unités de données le long de l'arbre de diffusion. MAF impose à la source un débit maximal de diffusion des données, fixé pour la durée de la session, à la création du groupe. En fixant le débit de la source, MAF prévient les situations de congestion dues aux rafales de trafic ainsi que les pertes liées à l'hétérogénéité des récepteurs du groupe. La suite du paragraphe présente les principaux mécanismes protocolaires

de MAF.

B. Découverte des chefs

MAF regroupe les routes qui permettent aux récepteurs de joindre la source, dans une structure logique que nous appellerons *arbre de contrôle*. Cette structure est logique en ce sens que seuls les routeurs actifs des routes qui séparent chacun des récepteurs de la source, constituent l'arbre de contrôle. Il s'agit des routeurs actifs que traversent les informations de contrôle retournées par les récepteurs à la source. Pour remplir son rôle dans les mécanismes de recouvrement sur erreur, MAF présuppose qu'un routeur actif qui diffuse des données, est en mesure d'intercepter les informations de contrôle que les routeurs actifs et récepteurs situés en aval retournent le long de l'arbre de contrôle : la séquence des routeurs actifs qui sépare dans l'arbre de contrôle, chacun des récepteurs de la source doit donc être l'inverse de celle que traverse la route inverse dans l'arbre de diffusion. MAF suppose donc que les routes séparant les récepteurs de la source dans l'arbre de contrôle, coïncident avec les routes inverses dans l'arbre de diffusion. Cette contrainte est due aux états qu'installe un routeur actif de l'arbre de diffusion en fonction des informations de contrôle reçues des récepteurs : c'est sur la base de ces états que sont réalisées la réparation des pertes, la réduction du nombre de récepteurs exposés à des retransmissions inutiles et l'agrégation des informations de contrôle.

Les arbres de diffusion et de contrôle pouvant être asymétriques, MAF adopte un mécanisme équivalent à celui qu'utilisent PGM (*Pragmatic General Multicast*) [SPE 01] et AER (*Active Error Recovery*) [KAS 00a], pour faire coïncider les routeurs actifs des deux arbres. MAF met en œuvre une signalisation légère qui utilise des messages appelés *Leader Discovery Message* ou LDM. Les LDM sont similaires aux *Source Path Message* de PGM. Cette signalisation permet à chacun des récepteurs et routeurs actifs de maintenir l'adresse du premier routeur actif situé immédiatement en amont dans l'arbre de contrôle. La source diffuse à intervalles réguliers les LDM à l'ensemble des récepteurs du groupe de diffusion. Un des champs de l'entête des LDM contient l'adresse du dernier routeur actif traversé dans l'arbre de diffusion. A la création des LDM, la source initialise ce champ en y positionnant sa propre adresse avant de les envoyer. Les routeurs actifs de l'arbre de diffusion et les récepteurs du groupe de diffusion enregistrent l'adresse que véhiculent les LDM reçus. Un routeur actif met à jour ce champ avec sa propre adresse avant de diffuser en aval le LDM intercepté. Les LDM étant diffusés régulièrement, l'adresse du routeur actif situé immédiatement en amont que maintiennent les récepteurs et routeurs actifs, reflète en permanence les changements des routes de l'arbre de diffusion.

Cette signalisation permet aux récepteurs et routeurs actifs de l'arbre de diffusion de connaître l'identité du chef courant dont ils dépendent.

C. Gestion des erreurs

MAF exécute au sein de chaque sous-arbre, une approche orientée émetteur, basée sur l'utilisation d'acquittements positifs (ACK) et de temporisateurs de retransmission. MAF délègue aux chefs de sous-groupes, la détection ainsi que la réparation des pertes locales au sous-arbre dont ils sont la racine: un chef de sous-groupe maintient en mémoire une copie des unités de donnée qu'il retransmet tant qu'elles ne sont pas acquittées par l'ensemble de ses fils. La décision d'effacer des copies d'unités de données en mémoire revient donc au chef et dépend des acquittements reçus de ses fils.

Au moyen des LDM régulièrement diffusés, les membres de sous-groupes connais- sent tous l'identité de leur chef respectif. Au sein d'un même sous-groupe, les membres retournent régulièrement des ACK directement adressés à leur chef, sans attendre les ACK de leurs propres fils. Pour réduire l'utilisation de bande passante des sous-arbres et la charge des chefs, ces acquittements sont périodiques et concernent l'ensemble des unités de données correctement reçues depuis le dernier acquittement émis. MAF requiert d'un chef de recevoir un ACK de chacun de ses fils avant d'être autorisé à libérer ses mémoires des unités de données acquittées. Une unité de données non acquittée par l'ensemble de ses fils, est interprétée comme étant perdue ; le chef de sous-groupe doit alors retransmettre cette unité.

Pour être en mesure de détecter les acquittements manquants, un chef utilise un temporisateur de retransmission et connaît la composition du sous-groupe dont il a la charge. Le temporisateur de retransmission représente l'intervalle de temps durant lequel ses fils sont supposés avoir retourné leurs acquittements. Un chef détermine T_{ack}, la valeur de son temporisateur de retransmission, en fonction du RTT qui le sépare de son fils le plus éloigné. Il enclenche son temporisateur de retransmission lorsqu'il diffuse la première unité de données. A l'épuisement du temporisateur de retransmission correspondant, les unités d'information non acquittées par l'ensemble de ses fils, sont interprétées comme étant perdues. Connaissant l'identité de ses D fils, un chef détermine ceux de ses fils dont il n'a pas reçu d'acquittement. Il retransmet alors les unités de données considérées comme perdues, uniquement à ceux de ses fils qui ne les ont pas acquittées et ce, jusqu'à ce qu'ils les aient positivement acquittées. Ce mécanisme préserve ainsi les fils qui n'ont pas enregistré la perte des retransmissions inutiles.

D. Gestions des mémoires des chefs

Comme nous venons de le voir ci-dessus, un chef de sous-groupe enregistre une copie des unités de données diffusées qu'il maintient en mémoire jusqu'à ce que l'ensemble de ses fils en aient accusé la réception. Il garantit ainsi la réparation systématique des pertes de ces unités de données affectant ses fils, sans avoir à s'en remettre aux chefs de niveaux supérieurs. Dès lors que l'ensemble de ses fils acquitte positivement des unités de données, un chef libère les mémoires correspondantes. Il est alors en mesure de recevoir les unités de données suivantes dont il pourra fiabiliser la diffusion à l'ensemble de ses fils. Un chef indique explicitement à son supérieur le nombre d'unités de données dont il libère ses mémoires, en retournant régulièrement des informations de contrôle appelées AACK (Aggregated ACK). Les AACK sont périodiques et

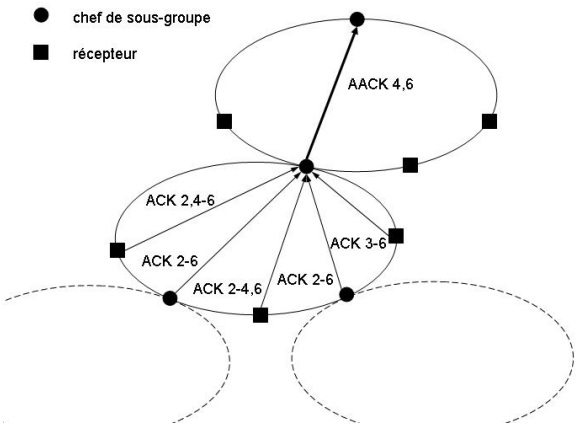


Fig. 1. Gestion des erreurs et des mémoires des chefs

concernent l'ensemble des unités de données correctement diffusées depuis le dernier AACK émis. T_{aack} correspond à l'intervalle de temps qui sépare l'émission de deux AACK consécutifs. Les chefs déterminent la valeur de T_{aack} de telle sorte que $T_{aack} = nT_{ack}$, avec $n > 1$.

Comme l'indique la figure 1, la construction des AACK revient pour un chef à agréger régulièrement les ACK qu'il reçoit de ses fils : il agrège les ACK dont la réception a provoqué la libération d'unités de données de ses mémoires. Les AACK reflètent donc le nombre d'unités de données correctement reçues par l'ensemble des membres des sous-groupes de niveau inférieur, depuis le dernier AACK émis. Le chef indique ainsi à son père le nombre d'unités de données qu'il s'attend à recevoir et dont il pourra enregistrer une copie sans avoir à remplacer des unités de données non encore acquittées par ses propres fils. Ce mécanisme permet à chacun des chefs de fiabiliser la diffusion des unités de données à l'ensemble de ses fils, tout en utilisant une quantité finie de sa mémoire. En effet, un chef déduit des AACK qu'il reçoit, le nombre d'unités de données *diffusables* ; il s'agit des unités de données pour lesquelles ses fils s'engagent à fiabiliser la diffusion au sein de leurs propres sous-groupes. Un chef contrôle ainsi la quantité de données que son propre chef lui envoie. Il évite ainsi la réception d'unités de données dont il ne pourrait enregistrer de copie pour en fiabiliser la diffusion ou dont l'enregistrement provoquerait le remplacement des copies d'unités de données non encore acquittées par l'ensemble de ses fils. Nous pouvons noter que les ACK et les AACK ne transitent qu'au sein du sous-arbre où ils ont été générés.

Au moyen des AACK qu'ils adressent à leur propre chef, les chefs sont en mesure de maintenir une copie des unités de données diffusées, jusqu'à ce qu'elles aient été positivement acquittées par l'ensemble de leurs fils. Les chefs garantissent ainsi la réparation systématique de l'ensemble des pertes enregistrées par leurs sous-groupes, sans avoir à s'en remettre aux chefs situés en amont dans l'arbre de contrôle.

IV. Implantation et déploiement

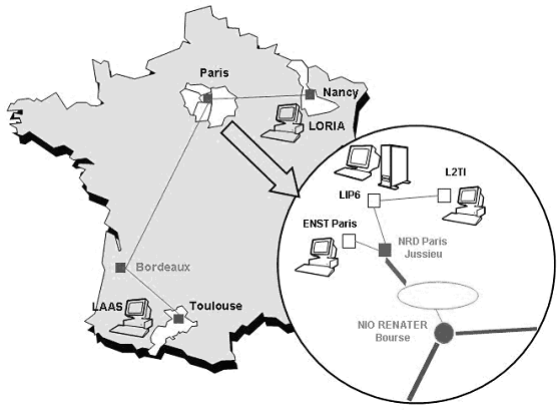
Nous avons développé un prototype du protocole MAF conforme à l'architecture de réseaux actifs spécifiée dans le cadre du projet RNRT Amarrage[1] auquel le LIP6 participe, les autres partenaires étant Thalès (Gennevilliers), France Télécom R& D (Issy-les-Moulineaux), l'ENST (Paris), le LAAS (Toulouse), le LORIA (Nancy), le L2TI (Villetaneuse). Nous avons utilisé la plate-forme nationale de démonstration mise en place dans le cadre du projet. La plate-forme Amarrage est une infrastructure de réseau actif virtuel : elle consiste en un prototype réaliste supportant le déploiement à l'échelle géographique de la France et la validation de composants logiciels actifs. L'architecture des nœuds actifs de cette plate-forme est construite autour des trois composants décrits ci-dessous :

- Le système d'exploitation des nœuds (*NodeOS*) est un noyau Linux (2.4.0) étendu de façon à intégrer à IPv6, le protocole standard d'encapsulation de paquets actifs ANEP (*Active Network Encapsulation Protocol*) [ALE 97]. Contrairement au réseau mondial d'expérimentation, l'Abone [BER 00] qui est construit comme un réseau d'*overlay* au-dessus de tunnels UDP, le *NodeOS* de l'architecture Amarrage implante des canaux de communication basés sur le protocole IPv6. La suppression d'UDP retire un niveau d'indirection dans les traitements des données et permet alors aux services actifs d'opérer directement sur IP.
- L'environnement d'exécution est une machine virtuelle Java rendue entièrement compatible IPv6 : elle offre au programmeur une interface pour l'accès à la programmation réseau sur IPv6 et aux fonctionnalités d'IPv6.
- L'interface de programmation réseau qu'exporte le *NodeOS* dérive de celle de ANTS [WET 98]. L'architecture initiale développée au MIT, a été en effet étendue afin de supporter les besoins spécifiques d'applications telles que MAF.

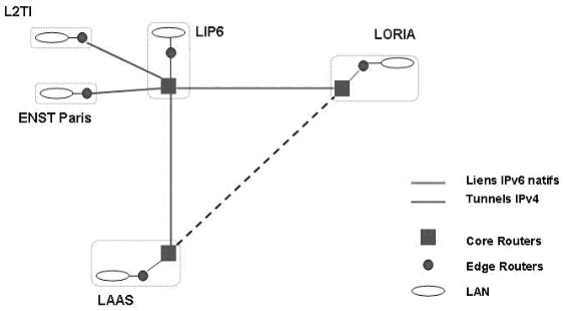
L'ensemble de ces composants est implanté dans les routeurs actifs d'accès des sites du projet connectés à la plate-forme. L'environnement est multi-sites et distribué entre Paris, Toulouse et Nancy. En raison des choix technologiques, la plate-forme est bâtie sur l'infrastructure RENATER (épine dorsale RENATER 2 bis) et bénéficie du service pilote IPv6 de RENATER. Le type de liaisons mises en place entre les sites dépend du type de connectivité des sites impliqués: le LIP6 et le LORIA sont dotés de connexions IPv6 natives à l'épine dorsale RENATER 2 bis, alors que l'ENST, le L2TI et le LAAS sont connectés par l'intermédiaire de tunnels IPv4 les reliant au NRD RENATER qui leur est le plus proche. Les figures 2(a) et 2(b) donnent respectivement une vue globale et logique de l'architecture de la plate-forme Amarrage utilisée dans le cadre des expérimentations.

Cette infrastructure de réseau actif virtuel nous a permis de déployer le prototype de MAF à l'échelle géographique de la France. Les expérimentations que nous avons menées à l'échelle nationale, ont permis de valider le modèle de communication fiable point à multipoint actif présenté dans cet article. Du fait des choix de conception de MAF, le délai de réparation des pertes observé pour chacun des récepteurs, correspond au RTT qui le sépare du chef de son sous-groupe. Par ailleurs,

[1] `http://www-rp.lip6.fr/amarrage/`



(a) Vue globale du réseau AMARRAGE



(b) Vue logique du réseau AMARRAGE

Fig. 2. La plate-forme AMARRAGE

les informations de contrôle (ACK et AACK) traitées par la source et les routeurs actifs, proviennent uniquement de leurs D fils : le volume des informations de contrôle est donc indépendant de la taille du groupe de diffusion. Au cours de nos expérimentations, aucun des routeurs actifs n'a été amené à remplacer d'unités de données non acquittées par l'ensemble de ses fils et nous n'avons observé aucune perte locale à un sous-arbre que le routeur actif situé à la racine n'ait pu réparée. Par conséquent, les informations de contrôle et les unités de données retransmises restent localisées au sein des sous-arbres où elles ont été générées. Nous poursuivons notre campagne de mesures pour mettre en relation le débit global de MAF et la quantité de mémoire dont les routeurs actifs de l'arbre de diffusion disposent.

V. Travaux antérieurs

Dans ce paragraphe, nous passons en revue les protocoles de diffusion fiable basés sur le soutien des routeurs de l'arbre de diffusion, récemment proposés dans la littérature. AER [KAS 00a] et ARM [LEH 98] reposent sur des architectures de réseaux programmables ; de leur côté, PGM [SPE 01] et LMS (*Light-weight Multicast Services*) [PAP 98] demandent un codage *en dur* des mécanismes de recouvrement sur erreur, au niveau de routeurs de l'arbre de diffusion.

Comme MAF, ces protocoles distribuent le rôle de réparation des pertes, le long de l'arbre de diffusion. En plus de la source, d'autres entités de l'arbre de diffusion se répartissent la réparation des pertes. L'ensemble de ces entités est organisé selon une structure hiérarchique arborescente comparable à celle qu'adopte MAF : à chaque sous-arbre est affectée une entité responsable des membres du sous-groupe correspondant. Dans ARM, ces entités sont des routeurs actifs de l'arbre de contrôle, programmés pour intercepter et traiter les informations de contrôle que les récepteurs situés en aval dans l'arbre de diffusion adressent à la source.

Dans LMS et PGM, la réparation des pertes est assumée par des récepteurs du groupe de diffusion, appelés respectivement *repliers* et *designed local receivers* (DLR) : au sein du sous-arbre dont ils sont la racine, les routeurs sont chargés de sélectionner un récepteur auquel ils redirigent systématiquement les informations de contrôle des autres récepteurs du sous-arbre.

Dans AER, les entités chargées de réparer les pertes sont des serveurs dédiés appelés serveurs de réparation, connectés à des routeurs de l'arbre de diffusion : un routeur doté d'un serveur de réparation implante les mécanismes qui lui permettent de détecter les informations nécessitant un traitement par son serveur de réparation. Le routeur intercepte ces informations et les redirige au serveur de réparation qui lui est connecté.

Si au sein de chaque sous-groupe, les protocoles existants préconisent tous l'exécution d'une approche orientée récepteur, ils diffèrent selon qu'en plus des récepteurs, les entités responsables des sous-groupes de niveau inférieur, assument ou non la détection des pertes. En effet, la détection des pertes locales à un sous-arbre est assumée dans AER aussi bien par les récepteurs que par les serveurs de réparation responsables des sous-arbres de niveau inférieur tandis qu'elle est à la charge des récepteurs uniquement dans ARM, PGM et LMS.

Quel que soit le type des entités responsable de réparer les pertes, leur détection repose uniquement sur l'utilisation d'acquittements négatifs (NACK) et de temporisateurs de retransmission. Ces protocoles diffèrent également par le mécanisme de retransmission des unités de données perdues mis en œuvre pour limiter, voire écarter l'exposition des récepteurs à des réparations inutiles.

Dans ARM et PGM, les routeurs maintiennent des états appelés respectivement *NACKrecord* et *repair state* où sont enregistrées pour chaque unité de données à réparer, les interfaces par lesquelles arrivent les NACK en notifiant la perte. Les routeurs déterminent ainsi les interfaces qui leur permettent de joindre les seuls récepteurs affectés par une perte qu'ils n'ont pu réparer.

Dans AER, les unités de données que répare un serveur de réparation sont diffusées par son routeur d'attache, à l'ensemble de ses fils, y compris à ceux qui n'en ont pas enregistré la perte. Les serveurs de réparation réduisent toute-

fois l'exposition des récepteurs à des réparations inutiles en supprimant les unités de données retransmises qu'ils reçoivent et dont ils n'ont pas enregistré la perte.
Dans LMS, un replier qui reçoit des NACK d'un routeur, lui retourne autant de copies de l'unité de données perdue que d'instances reçues de chaque NACK, copies que le routeur transmet en point à point à chacun des récepteurs affectés par cette perte.

Ces mêmes entités sont chargées de réduire le volume des informations de contrôle en supprimant les NACK dupliqués. Dans ARM et PGM, les routeurs suppriment les NACK qui indiquent des pertes pour lesquelles ils maintiennent respectivement un *NACKrecord* et un *repair state*.
Comme AER, les routeurs PGM utilisent également les techniques de découragement et de temporisation : un serveur de réparation qui reçoit de l'un de ses fils, un NACK notifiant une perte qu'il ne peut réparer, diffuse immédiatement ce NACK à l'ensemble de ses fils dans l'arbre de diffusion pour décourager ceux de ses fils qui s'apprêteraient à lui transmettre un NACK identique.

Parmi l'ensemble de ces protocoles, seuls les auteurs de PGM notent que le choix d'une approche orientée récepteur ne permet pas de garantir une fiabilité totale. En fait, PGM garantit qu'un récepteur qui ne reçoit pas toutes les unités de données, est toutefois capable de détecter les pertes non réparables. Les autres propositions de protocoles ne fonctionnent correctement que sur l'hypothèse de capacités de stockage infinies.

VI. Conclusion

Dans cet article, nous avons présenté la spécification et la conception de MAF, un nouveau modèle de communication point à multipoint. MAF rend un service de diffusion totalement fiable et adapté à des groupes de diffusion de grande taille, en définissant des mécanismes de recouvrement d'erreur que la technologie des Réseaux Actifs permet de déployer et d'implanter dynamiquement dans les routeurs actifs de l'arbre de diffusion.

La structuration hiérarchique arborescente, obtenue en divisant l'arbre de diffusion en sous-arbres, permet de réduire la charge de la source en attribuant la réparation des pertes aux routeurs actifs envoie situés à la racine des sous-arbres où ces pertes ont eu lieu : les routeurs actifs maintiennent une copie des données diffusées qu'ils retransmettent jusqu'à ce qu'elles aient été localement acquittées. Le retour des ACK permet aux routeurs actifs de détecter les pertes, celui des AACK consiste en une indication explicite de libération de leurs mémoires. En agrégeant les ACK qu'ils reçoivent, les routeurs actifs indiquent le nombre d'unités de données dont ils sont en mesure de maintenir une copie, jusqu'à ce que toutes les feuilles du sous-arbre dont ils sont responsables les aient correctement reçues. Les routeurs actifs garantissent ainsi la réparation systématique des pertes locales aux sous-arbres dont ils sont responsables, sans pour autant disposer de capacités de stockage infinies.

En structurant hiérarchiquement les routeurs actifs de l'arbre de diffusion, MAF minimise la latence de réparation des pertes. La structuration dispense également la source et les routeurs actifs de connaître la composition totale du groupe de diffusion. La quantité de traitement que MAF requiert de la source et des routeurs actifs, est indépendante de la taille du groupe de diffusion. Les informations de contrôle échangées pour détecter et réparer les pertes ne le sont que localement aux sous-arbres où ces pertes ont lieu : l'ensemble des ACK, des ACK agrégés et des retransmissions que provoque leur éventuelle absence, est toujours localisé au sein du sous-arbre où ces informations de contrôle ont été générées. MAF écarte ainsi les risques d'implosion de la source et réduit la consommation en bande passante des liens de l'arbre de diffusion. L'agrégation des ACK propres à chacun des sous-arbres permet à MAF de garantir une fiabilité totale tout en prévenant les situations de blocage.

Le prototype de MAF que nous avons implanté au-dessus de l'architecture de réseaux actifs développée dans le cadre du projet RNRT Amarrage, a été déployé à l'échelle géographique de la France. Les expérimentations que nous avons menées, ont permis de valider le modèle de communication fiable point à multipoint actif présenté dans cet article.

References

[ALE 97] Alexander D., Braden B., Gunter C., Jackson A., Keromytis A.Minden G.and Wetherall D., Active Network Encapsulation Protocol (ANEP), RFC Draft, 1997.

[BER 00] Bersen S., Braden B.Ricciulli L., Introduction to the ABone, 2000.

[CLA 98] Clark D., The Design Philosophy of the DARPA Internet Protocols, *Proceedings of ACM SIGCOMM'88*, Stanford, CA, 1998.

[KAS 00a] Kasera S., Bhattacharyya S., Keaton M., Kiwior D., Kurose J., Towsley D.Zabele S., Scalable Fair Reliable Multicast Using Active Services, *IEEE Network Magazine (Special Issue on Multicast)*, 2000.

[KAS 00b] Kasera S., Kurose J.Towsley D., Buffer Requirements and Replacement Policies for Multicast Repair Service, *Proceedings of NGC 2000*, Stanford, CA, 2000.

[LEH 98] Lehman L., Garland S.Tennenhouse D., Active Reliable Multicast, *Proceedings of IEEE INFOCOM'98*, San Francisco, CA, 1998.

[LEV 96] Levine B.Garcia-Luna-Aceves J. J., A Comparison of Known Classes of Reliable Multicast Protocols, *Proceedings of ICNP'96*, Columbus, OH, 1996.

[PAP 98] Papadopoulos C., Parulkar G.Varghese G., An Error Control Scheme for Large-Scale Multicast Applications, *Proceedings of IEEE INFOCOM'98*, San Francisco, CA, 1998.

[SPE 01] Speakman T., Crowcroft J., Gemmell J., Farinacci D., Lin S., Leshchiner D., Luby M., Montgomery T., Rizzo L., Tweedly A., Bhaskar N., Edmonstone R., Sumanasekera R.Vicisano L., PGM Reliable Transport Protocol Specification, RFC 3208, 2001.

[WET 98] Wetherall D., Guttag J.Tennenhouse D., ANTS: A Toolkit for Building and Dynamically Deploying Network Protocols, *Proceedings of IEEE OPENARCH'98*, San Francisco, CA, 1998.