

Chapitre III

Autoroutes de l'information

III.1 Introduction

Comme les années 80 ont établi les bases de systèmes répartis, les années 90 sont en train de fonder un nouveau domaine – celui des systèmes répartis de grande envergure. Le projet *Guide* est un exemple d'un système réparti fortement couplé, c'est à dire, un système où les sites très proches géographiquement sont connectés à un réseau local impliquant des communications rapides. Dans ce type de systèmes, les sites coopèrent de manière rapprochée pour réaliser une application répartie. Actuellement, on commence à voir apparaître des réseaux à haut débits qui étendent la possibilité de communiquer rapidement aux distances beaucoup plus grandes, voire à l'échelle mondiale. Les réseaux rapides sont à la base d'un nouveau type de systèmes répartis, à savoir les *systèmes de grande envergure*. Le facteur d'échelle est important – ces nouveaux réseaux peuvent relier des millions de machines comme c'est le cas actuellement d'Internet.

Une autre caractéristique importante de ces réseaux est l'apparition de nouveaux types de données. En plus des données traditionnelles comme le texte, des données *multimédia* (des images, de la vidéo et du son) pourront être échangées et traitées. Une analogie de ce réseau avec le réseau autoroutier est à l'origine du terme très médiatisé *des autoroutes de l'information (Information Superhighway)* [Reinhardt 94], [Verjus 94]. Les autoroutes de l'information seront l'effet de la convergence des réseaux de distribution de la télévision, des réseaux téléphoniques et des réseaux de transmission des données. Ils vont véhiculer des données multimédia et permettront le développement d'applications nouvelles comme le travail et l'enseignement à distance, la vidéo à la demande, des vidéo-conférences, le télé-enseignement, le commerce, les musées, les bibliothèques électroniques et beaucoup d'autres.

III.2 Autoroutes de l'information

L'émergence des autoroutes de l'information résulte de deux facteurs : l'apparition des réseaux à haut débit et le changement d'échelle du nombre de machines interconnectées. Les réseaux nouveaux comme *ATM (Asynchronous Transfer Mode)* [Rooholamini 94] offrent une gamme de débits très élevés allant de 34

Mbits/s à plusieurs Gbits/s et ils sont les plus appropriés à la transmission du multimédia, parce qu'ils permettent de négocier la qualité de service, par exemple une bande passante peut être garantie à chaque utilisateur [Gigabit 91], [Gigabit 92a], [Gigabit 92b]. D'autres paramètres de la qualité de service comme le débit, le taux d'erreurs, la gigue peuvent également être négociés. La qualité garantie de service permet de transporter des données très sensibles aux retards telles que la vidéo ou le son, par exemple, une image dans un flot de vidéo ne sera pas dégradée par des erreurs de transmission (ceci est vrai pour la vidéo non compressée), mais les images qui arrivent en retard ne sont plus utilisables. Selon le type de trafic, différents paramètres peuvent être choisis, par exemple, un transfert de fichier nécessite le moins d'erreurs possibles, mais il n'a pas besoin de la régularité de transfert. Les réseaux *ATM* pourront être utilisés à la fois comme des réseaux locaux, des réseaux métropolitains et des réseaux longues distances ce qui facilite leur intégration. En outre, *ATM* peut fonctionner sur tous les types de câbles allant de la paire torsadée à la fibre optique monomode. Son principe est basé sur le multiplexage des *cellules*, de paquets de longueur fixe de 53 octets, sur des circuits virtuels préalablement établis. Leur seul problème est la disponibilité – la normalisation d'*ATM* n'est pas encore terminée et les produits ne sont pas encore là. Néanmoins, il existe déjà des prototypes opérationnels comme *AURORA* [Clark 92] et diverses initiatives ont été déjà prises par les opérateurs de télécommunications [Reinhardt 94].

Le deuxième facteur dans le développement des autoroutes de l'information est l'échelle. Pour illustrer ce problème, prenons l'exemple du réseau mondial Internet, qui actuellement voit sa taille grandir rapidement. Il connecte plus de 31000 réseaux séparés et tous les 10 minutes, un nouveau réseau est ajouté. On estime le nombre de machines interconnectées par Internet à 2 millions et environ 20 millions de personnes peuvent communiquer à travers ce réseau [Internet 94]. Vu que ces chiffres correspondent aux débits encore très limités d'aujourd'hui, le volume d'informations transportées sur les futurs réseaux à haut débit sera en augmentation de plusieurs ordres de grandeur. Cette utilisation massive va aussi entraîner la diminution du coût de la communication ce qui permettra d'utiliser les réseaux à longue distance comme aujourd'hui on utilise les réseaux locaux (certaines prévisions parlent d'une diminution de 1000 du coût de la communication dans la décennie prochaine [Neece 94]). Ceci sera possible essentiellement à cause de la large utilisation de la fibre optique. Donc, ce changement d'échelle des données du problème (le nombre de machines, le nombre d'utilisateurs, le volume du trafic, le débit du trafic) impliquera des changements au niveau de l'architecture, de protocoles et des services des futurs réseaux.

Le fait que dans le futur toute information sera numérique (la télévision incluse) montre l'importance du multimédia pour les autoroutes de l'information. En général, le terme *multimédia* désigne un vaste et nouveau domaine qui intègre différentes formes de données : texte, son, images, vidéo, et inclut des systèmes, des

services et des outils qui permettent de stocker, manipuler et présenter les données multimédia de façon uniforme. La nature des données multimédia est très différente de celle des données qui sont habituellement traitées par ordinateur, essentiellement par sa nature temporelle. Cela posent des problèmes nouveaux et nécessite des solutions bien adaptées.

Dès le début des années 90, les premières propositions de création d'un réseau d'information à grande échelle sont apparues [Dertouzos 91a][Dertouzos 91b]. Ces desseins partent d'une constatation simple : les ordinateurs sont de plus en plus nombreux, mais ils restent relativement isolés et ne peuvent pas coopérer. En outre, l'information à échanger est de plus en plus variée. Il faut donc des moyens d'interconnexion et de coopération pour créer une entité nouvelle – un réseau global d'information. Les réseaux futurs à haut débit permettent d'envisager de nouvelles applications, mais le terme d'*autoroutes de l'information* fait ressortir l'importance de la vitesse de transport en occultant l'aspect le plus important, à savoir l'infrastructure d'information. Il est vrai que pour transférer des données multimédia il faut disposer de réseaux rapides, mais un support de communication rapide n'est que le premier élément nécessaire pour un réseau global de l'information. Deux autres éléments sont des conventions communes de communication et des services communs qui constituent l'infrastructure d'information au dessus de laquelle on peut développer des applications. Ceci ressort des propositions de *National Information Infrastructure* [NII] et des plates-formes expérimentales comme *Smart Valley* ou *BAGNet* [Verjus 94].

Une architecture à trois niveaux a été proposée pour le réseau global d'information dans le projet I-95 (l'équipe de *Programming Systems Research* au MIT LCS dans laquelle j'ai travaillé pendant mon séjour sabbatique, a contribué à l'élaboration de ce projet) [I-95 93]. Au niveau bas, la couche de *substrat* fournit la base d'interconnexion pour la communication rapide. La qualité de ce service peut être modulée en fonction du type d'information ; on peut choisir par exemple le débit, le délai et le niveau de sécurité ou de fiabilité. La couche de *substrat* utilise la technologie des réseaux *ATM* et *RNIS* large bande. Au niveau intermédiaire, la couche de *pont* réalise des services communs qui sont nécessaires aux applications coopérantes et concourantes. Par exemple, cette couche peut fournir un service de répertoires, un service de stockage d'objets, un service d'accès associatif à l'information répartie, un service d'authentification, un service de monnaie électronique et tous les autres services communs aux applications. Enfin, le niveau le plus haut de l'architecture est constitué de la couche *dialogue* qui fournit un cadre pour pouvoir facilement assembler des applications à partir de modules bien définis. Des actions à effectuer par des modules sont décrites dans un langage de scénario (*script*) qui est interprété par des agents intelligents travaillant pour le compte de l'utilisateur.

L'architecture proposée dans le projet I-95 montre l'importance du niveau intermédiaire (*pont*) et du niveau le plus haut (*dialogue*) par rapport au niveau bas

(*substrat*). Des prototypes opérationnels du niveau le plus bas (*substrat*) exploitant la technologie ATM cités précédemment commencent à apparaître (*AURORA*, *BAGNet*). Mais le développement d'applications sur ces prototypes est difficile à cause du manque d'infrastructure. L'état actuel ressemble un peu à celui des années 50 dans le domaine des ordinateurs où on développait des applications directement sur le matériel en ne disposant pas du support des systèmes d'exploitation. En fait, il ne suffit pas de disposer de moyens d'interconnexion rapide pour arriver à la coopération des applications et à la banalisation des services nouveaux, encore faut-il avoir une infrastructure d'information (des services et des outils) qui les facilite.

À l'inverse de ces prototypes de réseaux rapides, on peut observer une tendance opposée dans le développement des réseaux existants. En fait, on peut voir le réseau Internet qui permet l'interconnexion des systèmes à l'échelle mondiale comme un substrat de communication lent et primitif. On constate que les niveaux élevés de l'architecture comme des protocoles et des services, commencent à se développer très rapidement : à côté des services traditionnels comme le courrier électronique, le transfert de fichiers et la connexion à distance, de nouveaux services facilitant la coopération et l'accès à l'information sur réseau commencent à apparaître, par exemple les serveurs d'information comme *archie* [Emtage 92], *WAIS* [Kahle 91], *World-Wide Web (WWW)* [Berners-Lee 92], *Gopher* [Alberti 92], et les protocoles comme *Z39.50* [Z39.50 88] et *HTTP* [Berners-Lee 93]. On peut aussi citer *Softbot* [Etzioni 94], *Indie* [Danzig 92], *Information Brokers* [Barbara 92], *GLOSS* [Gravano 93] et *Nomenclator* [Ordille 93].

La croissance du trafic engendré par ces nouveaux services est explosive. On peut comparer le succès de ces nouveaux services à celui rencontré par les tableurs au début des années 80 : ils sont nés "spontanément", à partir d'initiatives d'utilisateurs, et répondent à des besoins réels. Ces serveurs sont devenus très populaires essentiellement à cause de l'application *Mosaic* [Andreessen 93] qui permet d'accéder de manière conviviale à l'espace global d'information créé par l'ensemble des serveurs d'information.

En dépit du succès rencontré par ces services, on peut constater que les techniques informatiques utilisées pour la mise en œuvre des serveurs sont encore très primitives. On peut noter par exemple que les serveurs sont en général repérés par leur localisation ce qui complique la migration d'un serveur ; de même, la tolérance aux défaillances est en général faible. Néanmoins, ce développement fulgurant de l'espace d'information global sur Internet montre l'intérêt d'orienter les efforts de recherches et de développement vers les deux niveaux hauts de l'architecture pour créer une infrastructure de l'information. Il faut souligner l'importance du *WWW* qui a créé un cadre pour une coopération évoluée sur réseau. Ses concepts sont simples et robustes : un paradigme de l'*hypertexte*, un schéma de nommage à l'aide des *URL (Uniform Resource Locators)*, des *formulaires* pour des interactions avec l'utilisateur et le mécanisme des *scripts* pour l'extensibilité. Néanmoins, il y a

encore beaucoup à apporter à ce domaine, en particulier, pour faire converger ces débuts de l'infrastructure de l'information avec des réseaux à haut débit.

Au cours de mon détachement au MIT, j'ai travaillé au sein de l'équipe *Programming Systems Research* sur deux aspects des autoroutes de l'information qui sont au cœur du développement de l'infrastructure de l'information pour les futurs réseaux. Le premier concerne la recherche de l'information sur réseaux – nous avons défini une architecture pour le *routage par contenu* qui permet de manière conviviale d'accéder à un ensemble de serveurs d'information sur réseau. Le deuxième aspect concerne la représentation des données multimédia – nous avons défini la *vidéo algébrique*, une représentation de la vidéo qui facilite l'accès associatif. La suite de chapitre présente ces activités.

III.3 Recherche de l'information sur réseaux

Il apparaît actuellement sur le réseau Internet un nombre grandissant de serveurs (cités ci-dessus) qui offrent l'accès aux données textuelles, au son, aux images et à la vidéo. A l'aide de *WAIS* on peut indexer des fichiers et les retrouver en posant des requêtes formulées en langue naturelle [Kahle 91], [Stein 91]. La recherche peut être affinée en spécifiant les documents qui s'avèrent correspondre le plus à la question posée. Pour trouver un serveur *WAIS* relevant à une requête, on peut consulter un répertoire des serveurs qui fournit une description succincte de chaque serveur. *WAIS* utilise un protocole standardisé *Z39.50* [Z39.50 88] pour la communication. Un autre exemple est *archie* [Emtage 92] qui collecte de l'information sur des fichiers archivés sur des serveurs *FTP*. Une requête *archie* retourne les noms de serveurs *FTP* qui archivent les fichiers pertinents ce qui permet ensuite de les récupérer via *FTP*.

Les systèmes de navigation sur réseau permettent d'organiser des serveurs d'information pour faciliter la navigation et la découverte d'information. *Gopher* structure des serveurs d'information en une arborescence qui ressemble à un système de fichier traditionnel [Alberti 92]. L'utilisateur peut naviguer dans des répertoires qui correspondent aux serveurs distants. Le *WWW* supporte la navigation dans un univers de documents hypertextes répartis [Berners-Lee 92]. En suivant des liens dans un document hypertexte, on accède aux autres documents éventuellement répartis. La communication entre des serveurs *WWW* se déroule selon le protocole *HTTP* qui spécifie comment on accède à un document hypertexte. Ce protocole permet aussi d'envoyer des demandes d'exécution d'actions sur des serveurs distants. Ce mécanisme est à la base des *documents virtuels*, qui sont générés comme le résultat d'une action, par exemple la réponse à une requête [Houh 94].

Il est difficile de trouver ou de localiser de l'information en utilisant les systèmes *Gopher* ou *WWW*, parce qu'ils sont surtout destinés à la navigation. Néanmoins, certains outils d'aide à la recherche commencent à apparaître. *Veronica* maintient

un index des titres de documents obtenus à partir des menus des serveurs *Gopher* ce qui fournit une aide pour découvrir les serveurs *Gopher* pertinents [Hahn 94]. Le *Jump Station* indexe les titres des documents hypertextes dans l'univers du WWW [URLc].

Tous ces serveurs d'information constituent une ressource potentiellement énorme, mais très mal utilisée à cause de la complexité du réseau et du manque d'outils adéquats de recherche et de navigation. Si l'on sait où chercher, on peut alors contacter et consulter les serveurs concernés, mais l'inconvénient de cette approche est que la recherche ne peut être effectuée que si l'on sait à quel serveur s'adresser. La solution idéale consiste à ce qu'une requête formulée par l'utilisateur soit automatiquement envoyée vers l'ensemble des serveurs susceptibles de fournir une réponse. Une première approche de cette solution est le *système de routage par contenu* décrit ci-dessous [Sheldon 94], [Duda 94]. Le contenu de chaque serveur est résumé dans une *étiquette de contenu* et un routeur fait suivre des requêtes de recherche selon les étiquettes des serveurs qui se sont enregistrés chez lui.

III.3.1 Routage par contenu

Cette section présente l'architecture du routage par contenu et son premier prototype réalisé pour les serveurs *WAIS*.

La difficulté de fournir l'accès associatif au contenu des serveurs d'information répartis réside essentiellement dans le problème d'échelle. Ce problème peut être abordé de deux points de vue extrêmes. Le premier consiste à constater que pour pouvoir trouver des serveurs pertinents à une requête, il faut disposer de l'information sur le contenu de tous les serveurs. Dans ce cas, il faudrait construire un index global qui serait consulté pour le routage des requêtes. L'index pourrait assurer que tous les serveurs pertinents à une requête serait trouvés. Cette solution peut suffire à petite échelle, mais elle ne s'adapte pas à la taille du réseau Internet et de son accroissement constant. L'espace de stockage nécessaire pour l'index est également trop important et son maintien à jour est difficile.

Une autre solution à l'opposée de la précédente consisterait à envoyer chaque requête à tous les serveurs. Même si cette méthode assure aussi que tous les serveurs pertinents seront trouvés, elle implique trop de trafic sur le réseau et trop de charge de travail aux serveurs. En fait, il faut une solution intermédiaire qui permettrait d'acheminer une requête à un ensemble réduit de serveurs et ceci sans accéder à un index global.

Un autre problème est celui du nombre de réponses qui peuvent être obtenues d'une recherche sur un nombre important des serveurs. Si l'utilisateur reçoit trop de réponses à sa question, il ne pourra pas les bien exploiter. Ce qu'il lui faut, c'est un moyen de la découverte progressive de l'espace d'information et la spécialisation des questions. Pour ces raisons, nous pensons que l'accès à l'information répartie

doit reposer sur un support à la fois du *routing de requêtes* et d'*aide à la composition de requêtes*.

Notre solution au problème d'accès à l'information répartie est basée sur l'architecture de *routing par contenu* [Sheldon 94]. Le système de routing par contenu peut être vu comme l'extension d'un système de recherche d'information traditionnel qui fournit un accès associatif aux documents. En plus de documents (qui peuvent être multimédia), notre système supporte un nouveau type de document que l'on appelle *collection*. Une collection est un ensemble de documents (qui peuvent être aussi des collections) et une *étiquette de contenu*. L'étiquette de contenu est un prédicat qui résume le contenu de l'ensemble de documents. Une collection peut être vue comme une organisation hiérarchique de documents dans un graphe acyclique (cf. Fig. 3.1, EC désigne l'étiquette de contenu).

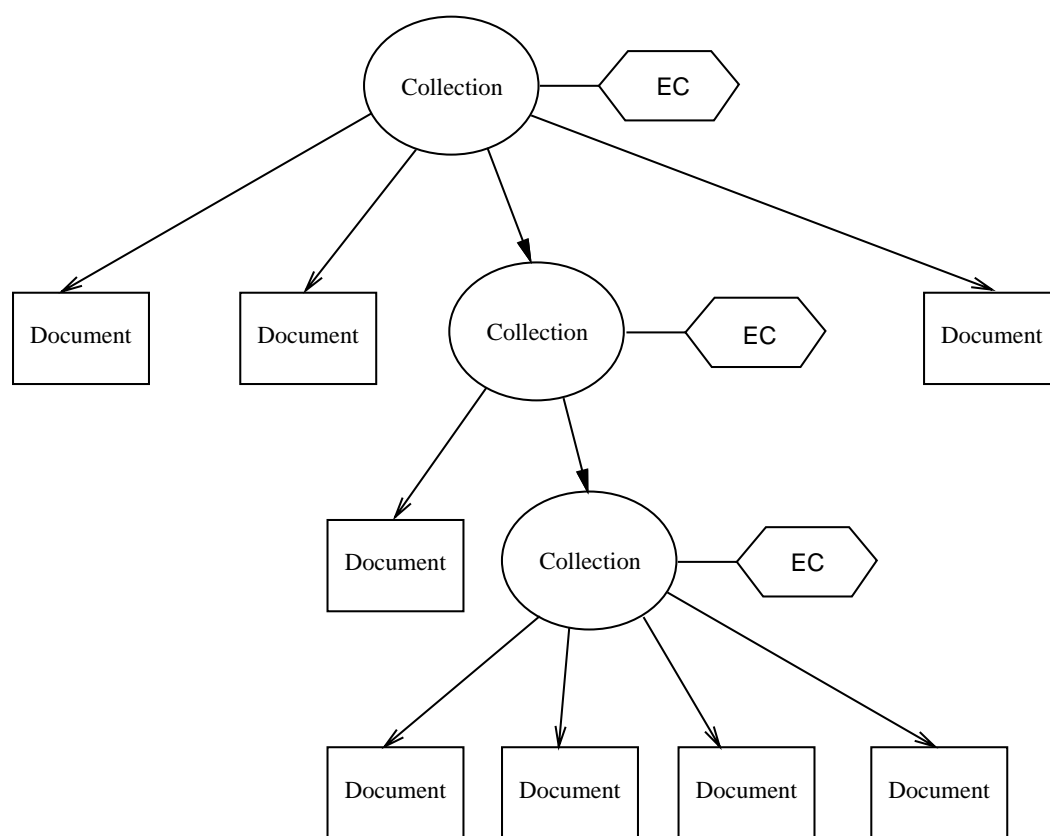


Fig. 3.1 : Organisation hiérarchique de collections

Un système de routing par contenu est une composition hiérarchique de serveurs. Les feuilles de cette architecture sont des serveurs d'information qui stockent des documents et les retournent en réponse des requêtes, tandis que les nœuds internes sont des routeurs qui supportent des collections. Des serveurs d'information se font enregistrer auprès des routeurs en fournissant les étiquettes de contenu.

Un routeur fournit une interface pour des opérations sur des collections :

expand

cette opération remplace une collection par ses documents membres ; sur un document qui n'est pas une collection cette opération n'a pas d'effet.

search

cette opération initialise une recherche dans l'espace des documents où toutes les collections sont remplacées par leurs documents membres.

select

cette opération choisit les documents qui satisfont à un prédicat dans une collection.

completion

cette opération aide à la composition de requêtes en proposant des termes à ajouter.

retrieve

cette opération accède à un document ; appliquée sur une collection, elle retourne l'étiquette de contenu de la collection.

Pour expérimenter cette approche à grande échelle, nous avons développé un prototype du routeur par contenu pour les serveurs *WAIS* [Duda 94]. Il existe actuellement environ 500 serveurs *WAIS* qui gèrent de documents de domaines très variés allant des archives des *news*, jusqu'à l'information sur le génome humain. La difficulté de la construction d'un routeur pour *WAIS* consiste à l'élaboration des étiquettes de contenu des serveurs. La solution idéale consisterait à avoir l'accès à l'index *WAIS* de chaque serveur ce qui nécessite la coopération de tous les administrateurs – il n'y a pas de moyen d'accéder de manière automatique à ces index. Même si l'on peut les récupérer, leur taille devient prohibitive parce qu'en moyenne les index *WAIS* sont de taille des données utilisées pour l'indexation. Nous avons donc décidé d'utiliser les fichiers catalogues qui contiennent les en-têtes des documents, par exemple le sujet de message, le titre d'un document ou le titre d'une entrée bibliographique. Par conséquent, ils contiennent une information significative sur le contenu des documents. En outre, la taille des catalogues est une fraction de la taille des index et ils peuvent être récupérés automatiquement en réponse à une requête vide.

A partir des fichiers catalogue, les étiquettes de contenu des serveurs sont créées, indexées et utilisées ensuite pour le routage et pour l'aide à la composition de requêtes. L'utilisateur commence d'abord par composer une requête avec des termes généraux et le routeur retourne un ensemble des serveurs *WAIS* qui sont susceptibles de fournir une réponse. Si l'utilisateur veut affiner sa requête, le système lui propose un choix de termes qui sont liés aux termes déjà spécifiés : ce sont les termes