Groupe de travail Réseau
Request for Comments : 3439
RFC mise à jour : 1958
Catégorie : Information

R. Bush
D. Meyer
décembre 2002
Traduction Claude Brière de L'Isle

Lignes directrices et philosophie de l'architecture de l'Internet

Statut de ce mémoire
Le présent mémoire apporte des informations pour la communauté de l'Internet. Il ne spécifie aucune sorte de norme de l'Internet. La distribution du présent mémoire n'est soumise à aucune restriction.

Résumé
Le présent document étend la RFC 1958 en soulignant certaines des lignes directrices de la philosophie auxquelles devraient adhérer les architectes et concepteurs des réseaux qui constituent l'épine dorsale de l'Internet. Nous décrivons le principe de simplicité, qui établit que la complexité est le principal mécanisme entravant une mise à niveau efficace, et nous discutons ses implications sur les questions d'architecture, de conception et d'ingénierie qui se rencontrent dans les réseaux dorsaux de grande échelle de l'Internet.

Table des matières

1. Introduction

La [RFC1958] décrit les principes sous-jacents de l'architecture de l'Internet. La présente note étend ce travail en soulignant certaines des lignes directrices philosophiques auxquelles les architectes et concepteurs des réseaux dorsaux de l'Internet

devraient adhérer. Bien que beaucoup des domaines auxquels on se réfère dans le présent document puissent être le sujet de controverses, le principe unificateur qu'on décrit ici, avec le contrôle de la complexité comme mécanisme de contrôle des coûts et de la fiabilité, ne devrait pas l'être. Dans les réseaux de transporteurs, la complexité peut provenir de nombreuses sources. Cependant, comme précisé dans [DOYLE2002], "Dans la plupart des systèmes, la complexité résulte du besoin de robustesse contre l'incertitude dans leurs environnements et leurs éléments composants beaucoup plus que des fonctionnalités de base". L'objet majeur du présent document, et donc d'attirer l'attention sur la complexité de certaines de nos architectures actuelles, et d'examiner les effets qu'a presque certainement une telle complexité sur la capacité de succès de l'industrie des transporteurs IP.

Le reste du présent document est organisé comme suit : la première section décrit le principe de simplicité et ses implications pour la conception de très grands systèmes. Le reste du document souligne les conséquences à haut niveau du principe de simplicité et comment il devrait guider les approches d'architecture et de conception des très grands réseaux.

2. Grands systèmes et principe de simplicité

Le principe de simplicité, qui a peut-être été formulé pour la première fois par Mike O'Dell, ancien architecte en chef de UUNET, déclare que la complexité est le principal mécanisme qui entrave une mise à l'échelle efficace, et a pour résultat d'être la principale cause de l'augmentation à la fois des dépenses en capital (CAPEX, *capital expenditures*) et des dépenses de fonctionnement (OPEX, *operational expenditures*). L'implication pour les transporteurs de réseaux IP est alors que pour aller vers le succès, les architectures et les concepts doivent tendre vers les solutions les plus simples possibles.

2.1 L'argument de bout en bout et la simplicité

L'argument de bout en bout, qui est décrit dans [SALTZER] (ainsi que dans la [RFC1958]), contient que le "concept de protocole de bout en bout ne devrait pas s'appuyer sur le maintien de l'état (c'est-à-dire, des informations sur l'état de la communication de bout en bout) à l'intérieur du réseau. Un tel état ne devrait être maintenu que dans les points d'extrémité, d'une façon telle que l'état ne puisse être détruit que lorsque le point d'extrémité lui-même le casse." Cette propriété a aussi été rapportée au concept de "sort partagé" de Clark [CLARK]. On peut voir que le principe de bout en bout conduit directement au principe de simplicité en examinant la formulation qu'on appelle "sablier" de l'architecture de l'Internet [WILLINGER2002]. Dans ce modèle, la taille fine du sablier est envisagée comme la couche (minimaliste) IP, et toute la complexité supplémentaire est ajoutée au-dessus de la couche IP. En bref, la complexité de l'Internet appartient aux bordures, et la couche IP de l'Internet devrait rester aussi simple que possible.

Finalement, noter que l'argument de bout en bout n'implique pas que le cœur de l'Internet ne va pas contenir et maintenir l'état. En fait, une énorme quantité d'états grossiers sont maintenus dans le cœur de l'Internet (par exemple, d'état d'acheminement). Cependant, le point qui importe ici est que cet état (grossier) est presque orthogonal à l'état maintenu par les points d'extrémité (par exemple, les hôtes). C'est cette minimisation de l'interaction qui contribue à la simplicité. Il en résulte que la considération de l'interaction d'état de "cœur contre point d'extrémité" est cruciale pour l'analyse de protocoles tels que ceux de traduction d'adresse réseau (NAT), qui réduisent la transparence entre le réseau et les hôtes.

2.2 Non linéarité et complexité du réseau

Les architectures et conceptions complexes ont été (et continuent d'être) les barrières et les défis les plus significatifs à la construction de réseaux IP à grande échelle qui soient rentables. Considérons, par exemple, la tâche de construire un grand réseau de paquets. L'expérience de l'industrie a montré que la construction d'un tel réseau est une activité différente (qui requiert donc un ensemble différent de compétences) de celle de la construction d'un réseau de taille petite ou moyenne, qui ont des propriétés intrinsèques différentes. En particulier, les plus grands réseaux présentent, à la fois en théorie et en pratique, des non linéarités architecturales, de conception, et d'ingénierie qui ne se présentent pas à une plus petite échelle. On appelle cela des non linéarités d'architecture, conception, ingénierie (ADE, *Architecture, Design, Engineering*). C'est à dire que des systèmes comme l'Internet pourraient être décrits comme extrêmement auto dissemblables, avec des échelles et niveaux d'abstraction très différents [CARLSON]. La propriété de non linéarité ADE se fonde sur deux principes bien connus tirés de la théorie des systèmes non linéaires [THOMPSON]:

2.2.1 Principe d'amplification

Le principe d'amplification déclare qu'il y a des non linéarités qui surviennent à grande échelle qui ne se produisent pas à des échelles petites ou moyennes.

Corollaire : Dans la plupart des grands réseaux, de petites choses peuvent causer, et causent, de grands événements. En termes de théorie des systèmes, dans de grands systèmes comme ceux là, même de petites perturbations à l'entrée d'un

processus peuvent déstabiliser le résultat du système.

Un important exemple du principe d'amplification est l'amplification de la résonance non linéaire, qui est un processus puissant qui transforme les systèmes dynamiques, comme les grands réseaux, de façon surprenante avec de petites fluctuations imperceptibles. Ces petites fluctuations peuvent s'accumuler lentement, et si elles sont synchronisées avec d'autres cycles, peuvent produire des changements majeurs. Les phénomènes de résonance sont des exemples de comportements non linéaires où de petites fluctuations peuvent être amplifiées et avoir des influences qui excèdent de loin leur taille initiale. Le monde de la nature est rempli de ces exemples de comportements de résonance qui peuvent produire des changements à l'échelle du système, tels que la destruction du pont de Tacoma Narrows (due à l'amplification de la résonance de petits coups de vent). Parmi d'autres exemples on citera les trous dans les ceintures d'astéroïdes et les anneaux de Saturne qui ont été créés par l'amplification de résonances non linéaires. Certaines caractéristiques du comportement humain et la plupart des systèmes de pèlerinages sont influencés par les phénomènes de résonance qui impliquent la dynamique du système solaire, les jours solaires, les cycles de 27,3 jours (sidéral) et de 29,5 jours (synodique) de la Lune ou le cycle de 365,25 jours du Soleil.

Dans le domaine de l'Internet, il a été montré qu'une interconnectivité accrue résulte en une convergence d'acheminement BGP pus complexe et souvent plus lente [AHUJA]. Il en résulte qu'une petite quantité d'inter connectivité est cause que le résultat d'un maillage d'acheminement est significativement plus complexe que son entrée [GRIFFIN]. Une importante méthode de réduction de l'amplification est de s'assurer que les changements locaux ont seulement un effet local (c'est à l'opposé des systèmes dans lesquels des changements locaux ont un effet global). Finalement, ATM fournit un excellent exemple d'effet d'amplification : si on perd une seule cellule, on détruit tout le paquet (et cela devient pire, car en l'absence d'un mécanisme comme l'élimination précoce de paquet [ROMANOV], on va continuer de transporter des paquets déjà endommagés).

Un autre intéressant exemple d'amplification vient du domaine de l'ingénierie, et il est décrit dans [CARLSON]. Il considère le Boeing 777, qui est un avion "informatisé", qui contient jusqu'à 150 000 sous-systèmes et approximativement 1000 CPU. On observe que bien que le 777 soit robuste aux grosses perturbations atmosphériques, aux limites de turbulences, et aux variations des charges embarquées (pour en nommer quelques unes) il pourrait être paralysé de façon catastrophique par des altérations microscopiques de quelques gros processeurs (ce qui, on le souligne, n'arrive heureusement qu'en de très rares occurrence). Cet exemple illustre la question que "la complexité peut amplifier de petites perturbations, et l'ingénieur qui conçoit le projet doit s'assurer que de telles perturbations sont extrêmement rares." [CARLSON]

2.2.2 Principe de couplage

Le principe de couplage déclare qu'à mesure que les choses deviennent plus grandes, elles présentent souvent une interdépendance accrue entre leurs composants.

Corollaire : Plus il y a d'événements qui surviennent simultanément, plus il est probable que deux ou plus vont interagir. Ce phénomène a aussi été appelé "interaction de caractéristiques imprévues" [WILLINGER2002].

Beaucoup des non linéarités observées dans les grands systèmes sont largement dues au couplage. Ce couplage a des composants à la fois horizontaux et verticaux. Dans le contexte du réseautage, le couplage horizontal se présente entre des composants de même couche de protocole, tandis que le couplage vertical survient entre les couches.

Le couplage se manifeste dans des systèmes naturels très divers, y compris des macro instabilités du plasma (hydromagnétique, par exemple, des effets de plissements, de cheminées volcaniques, de miroir, de ballonnement, de déchirures, de capture de particules) [NAVE], ainsi que diverses sortes de systèmes électrochimiques (considérer le problème de l'étiquetage des nucléotides fluorescents de synthèse et des acides nucléique [WARD]). Le couplage de la périodicité d'horloges physiques a aussi été observé [JACOBSON], ainsi que le couplage de divers types de cycles biologiques.

Il existe aussi plusieurs exemples canoniques dans des systèmes de réseau bien connus. On a comme exemples celui de la synchronisation de diverses boucles de contrôle, comme la synchronisation des mises à jour d'acheminement et la synchronisation du démarrage lent de TCP [FLOYD], [JACOBSON]. Un résultat important de ces observations est que le couplage est intimement relié à la synchronisation. Injecter de l'aléatoire dans ces systèmes est une façon de réduire le couplage.

Il est intéressant de noter qu'en analysant les facteurs de risque pour le réseau téléphonique public commuté (RTPC) Charles Perrow décompose le problème de la complexité le long de deux axes, qu'il appelle "interactions" et "couplage" [PERROW]. Perrow cite les interactions et le couplage comme des facteurs significatifs pour déterminer la fiabilité d'un système complexe (et en particulier, le RTPC). Dans ce modèle, interactions se réfère à la dépendance entre les composants

(linéaires ou non linéaires) alors que le couplage se réfère à la souplesse dans un système. Les systèmes avec des interactions linéaires simples ont des composants qui n'affectent que les autres composants qui sont fonctionnellement en aval. Les systèmes complexes ont des composants qui interagissent avec de nombreux autres composants dans des parties différentes et éventuellement éloignées du système. Les systèmes à couplage lâche sont dits avoir plus de souplesse à l'égard des contraintes de temps, de séquençage, et des conditions environnementales, que les systèmes à couplage étroit. De plus, les systèmes qui ont des interactions complexes et un couplage étroit auront plus probablement des états de défaillances imprévues (bien sûr, les interactions complexes permettent le développement de plus de complications qui rendent le système difficile à comprendre et prévoir) ; ce comportement est aussi décrit dans [WILLINGER2002]. Un couplage étroit signifie aussi que le système a moins de souplesse pour se récupérer dans les états de défaillance.

Le système de signalisation n° 7 (SS7) du RTPC pour le contrôle du réseau fournit un intéressant exemple de ce qui peut tourner mal avec un système complexe à couplage étroit. Les pannes comme la panne fameuse de 1991 du SS7 de AT&T démontrent le phénomène : la panne a été causée par des erreurs de logiciel dans le code de récupération de panne des commutateurs. Dans ce cas, un commutateur est tombé en panne à cause d'une casse de matériel. Quand ce commutateur est passé en sauvegarde, il a causé (plus un problème de file d'attente d'une probabilité raisonnable) l'engorgement de ses voisins. Lorsque les commutateurs voisins sont passés en sauvegarde, ils ont causé le crash des voisins, et ainsi de suite [NEUMANN] (la cause première s'est révélée être une commande "coupure" mal placée ; c'est un excellent exemple de couplage inter couches). Ce phénomène est similaire au couplage de phase des oscillateurs faiblement couplés, dans lesquels des variations aléatoires des temps de séquence jouent un rôle important dans la stabilité du système [THOMPSON].

2.3 La leçon de la complexité du vocal

Dans les années 1970 et 1980, les opérateurs vocaux rivalisaient d'ajout de dispositifs qui ont conduit à une augmentation substantielle de la complexité du RTPC, en particulier de l'infrastructure des commutateurs de classe 5. Cette complexité était normalement fondée sur le logiciel, pas sur le matériel, et avait donc des courbes de coûts pires que celle de la Loi de Moore. En résumé, les faibles marges sur les produits vocaux d'aujourd'hui sont dus aux fait que les coûts des OPEX et CAPEX ne chutent pas comme on aurait pu l'espérer de la part de mises en œuvre de matériels simples.

2.4 Mise à niveau du coût de la complexité

Considérons le coût de fourniture de nouveaux dispositifs dans un réseau complexe. Le réseau vocal traditionnel a peu d'intelligence dans ses appareils d'extrémité (les appareils téléphoniques) et un cœur très intelligent. L'Internet a des extrémité intelligentes, des ordinateurs avec des systèmes d'exploitation, des applications, etc., et un cœur simple, qui consiste en un plan de contrôle et des moteurs de transmission de paquets. Ajouter un nouveau service Internet est juste une question de distribution d'une application aux quelques ordinateurs consentants qui souhaitent l'utiliser. Comparer ceci à l'ajout d'un service au téléphone vocal, où on doit mettre à niveau le cœur du réseau entier.

3. La mise en couche considérée comme dommageable

Il y a plusieurs propriété génériques de la mise en couches, ou intégration verticale, telle qu'elle s'applique au réseautage. En général, une couche telle que définie dans notre contexte met en œuvre un ou plusieurs des éléments suivants :

Contrôle d'erreur : La couche rend le "canal" plus fiable (par exemple, couche de transport fiable)

Contrôle de flux : La couche évite d'inonder les homologues plus lents (par exemple, contrôle de flux ATM)

Fragmentation : Diviser de gros tronçons de données en plus petits morceaux, et les réassembler ensuite (par exemple, la fragmentation/réassemblage de TCP MSS)

Multiplexage : Permettre que plusieurs sessions de niveau supérieur partagent une seule "connexion" de niveau inférieur (par exemple, PVC ATM)

Établissement de connexion : Prise de contact avec un homologue (par exemple, prise de contact en trois phases de TCP, ILMI ATM)

Adressage/désignation : Des identifiants de localisation, de gestion, associés à des entités (par exemple, structure NSAP de GOSSIP 2 [RFC1629])

La mise en couche de ce type présente divers avantages conceptuels et de structuration. Cependant, dans le contexte de la

mise en couche de réseau, la mise en couche structurée implique que les fonctions de chaque couche sont entièrement mises en œuvre avant que l'unité de données de protocole soit passée à la couche suivante. Cela signifie que l'optimisation de chaque couche doit être faite séparément. De telles contraintes de rangement entrent en conflit avec l'efficacité de mise en œuvre des fonctions de manipulation des données. On pourrait accuser le modèle en couches (par exemple, TCP/IP et ISO OSI) de causer ce conflit. En fait, les opérations de multiplexage et de segmentation cachent toutes deux des informations vitales dont des couches inférieures peuvent avoir besoin pour optimiser leurs performances. Par exemple, la couche N peut dupliquer des fonctionnalités de niveau inférieur, par exemple, la récupération d'erreur bond par bond contre la récupération d'erreur de bout en bout. De plus, des couches différentes peuvent avoir besoin des mêmes informations (par exemple, d'horodatage) : la couche N peut avoir besoin des informations de la couche N-2 (par exemple, les tailles de paquet de la couche inférieure) et ainsi de suite [WAKEMAN]. Une remarque à ce sujet peut être encore plus ironique vient de l'article classique de Tennenhouse, "Le multiplexage en couche considéré comme dommageable" [TENNENHOUSE] : "L'approche ATM du réseautage haut débit est actuellement suivie au sein du CCITT (et ailleurs) comme mécanisme unificateur pour la prise en charge de l'intégration de services, de l'adaptation du débit, et du contrôle de la gigue au sein des couches inférieures de l'architecture du réseau. Cette prise de position est particulièrement concernée par le fait que la gigue sort du concept que les couches "moyennes" et "supérieures" fonctionnent au sein des systèmes d'extrémité et des relais des réseaux multi service (MSN, *multi-service network*)."

Par suite de l'interdépendance des couches, une mise en couche accrue peut rapidement conduire à une violation du principe de simplicité. L'expérience de l'industrie nous enseigne qu'une mise en couche accrue augmente fréquemment la complexité et conduit donc à une augmentation des OPEX, comme prédit par le principe de simplicité. Son corollaire est établit au paragraphe 2.5 de la [RFC1925] : "Il est toujours possible d'agglutiner plusieurs problèmes distincts en une seule solution d'interdépendance complexe. Dans la plupart des cas, c'est une mauvaise idée."

La première conclusion est alors qu'une séparation horizontale (par opposition à verticale) peut être plus économique et fiable à long terme.

3.1 L'optimisation considérée comme dommageable

Un corollaire des arguments de mise en couche ci-dessus est que l'optimisation peut aussi être considérée comme dommageable. En particulier, l'optimisation introduit de la complexité, tout comme elle introduit un plus étroit couplage entre composants et couches.

Un important effet de l'optimisation est décrit dans la Loi de diminution des retours, qui déclare que si un facteur de production est augmenté alors que les autres restent constants, le retour global va relativement diminuer après un certain point [SPILLMAN]. L'implication ici est qu'essayer d'augmenter l'efficacité au delà de ce point ne fait qu'ajouter de la complexité, et conduit donc à des systèmes moins fiables.

3.2 La richesse des caractéristiques est considérée comme dommageable

Bien qu'ajouter de nouvelles caractéristique puisse être considéré comme un gain (et en fait différencie fréquemment les fabricants des divers types d'équipements) il y a un danger. Ce danger est une augmentation de la complexité des systèmes.

3.3 Évolution de l'efficacité du transport pour IP

L'évolution des infrastructures de transport pour IP offre un bon exemple de la façon dont la diminution de l'intégration verticale a conduit à des efficacités diverses. En particulier,

```
  | IP sur ATM sur SONET  -->
  | IP sur SONET sur WDM  -->
  | IP sur WDM
  |
 \|/
Complexité décroissante, CAPEX, OPEX
```

Le point clé est ici que les couches sont retirées, ce qui résulte en efficacité de CAPEX et d'OPEX.

3.4 Convergence de mise en couche

La convergence se rapporte aux concepts de mise en couche décrits ci-dessus en ce que la convergence est réalisée via une "couche de convergence". L'état final de l'argument de convergence est le concept de "toute chose a sa couche" (EOSL,

Everything Over Some Layer). Conduit, DWDM, fibre, ATM, MPLS, et même IP ont tous été proposés comme couches de convergence. Il est important de noter que comme la mise en couche conduit normalement à une augmentation des OPEX, on peut s'attendre à ce que la convergence le fasse aussi. Cette observation est là aussi cohérente avec l'expérience de l'industrie.

Il y a de nombreux exemples notables d'échec de couche de convergence. Peut-être que l'exemple le plus approprié est celui d'IP sur ATM. La conséquence immédiate et la plus visible de la mise en couche d'ATM est ce qu'on appelle la taxe à la cellule ; noter d'abord que la réponse complète sur l'efficacité ATM est qu'elle dépend des distributions de taille de paquet. Supposons le schéma typique du trafic de type Internet, qui tend à avoir un fort pourcentage de paquets à 40, 44, et 552 octets. Des données récentes [CAIDA] montrent qu'environ 95 % des octets de WAN et 85 % des paquets sont TCP. Beaucoup de ce trafic est composé de paquets de 40/44 octets.

Maintenant, considérons le cas d'un cœur de réseau DS3 avec la procédure de convergence de la couche physique (PLCP, *Physical Layer Convergence Procedure*) activée. Le débit de cellule maximum est de 96 000 cell/s. Si on multiplie cette valeur par le nombre de bits de la charge utile, on obtient 96 000 cell/s * 48 octet/cell * 8 = 36 864 Mbit/s. Cela n'est cependant pas réaliste car cela suppose une mise en paquet parfaite de la charge utile. Il y a deux autres choses qui contribuent à la redondance d'ATM (la taxe à la cellule) : le gâchis de bourrage et les huit octets de l'en-tête de protocole d'accès à un sous-réseau (SNAP, *subnetwork access protocol*).

C'est l'en-tête SNAP qui cause le plus de problèmes (et on n'y peut rien) car il force la plupart des petits paquets à consommer deux cellules, dont la seconde est presque entièrement constituée de bits de bourrage (ce qui interagit très mal avec les données mentionnées ci-dessus, par exemple, que la plupart des paquets sont des accusés de réception TCP de 40 à 44 octets). Cela cause une perte d'environ 16 % supplémentaires sur le débit idéal de 36,8 Mbit/s.

De sorte que le débit total se trouve être, pour un DS3 :

débit de ligne DS3 :	44 736
redondance de PLCP :	- 4 032
en-tête par cellule :	- 3 840
en-tête SNAP et bourrage :	- 5 900
	=========
	30 960 Mbit/s

Résultat : avec un débit de ligne DS3 de 44 736 Mbit/s, le déchet total est d'environ 31 %.

Une autre façon de regarder cela est que comme une large fraction du trafic de WAN est composée d'accusés de réception TCP, on peut faire un calcul différent mais en rapport :
IP sur ATM exige
- les données IP (40 octets dans ce cas)
- les 8 octets de SNAP
- les 8 octets de bourrage AAL5
- 5 octets pour chaque cellule
+ tout ce qu'il faut pour remplir la dernière cellule.

Sur ATM, cela devient deux cellules - 106 octets pour porter 40 octets d'information. La prochaine taille la plus courante semble être une parmi les tailles de la gamme de 504 - 556 à 636 octets pour porter IP, TCP, et une charge utile TCP de 512 octets – avec les messages de plus de 1000 octets qui viennent en troisième.

On pourrait imaginer que 87 % de la charge utile (pour la taille de message de 556 octets) est mieux que 37 % de la charge utile (pour la taille d'accusé de réception TCP) mais ce ne sont pas les 95 à 98 % auxquels sont habitués les consommateurs, et la prédominance des accusés de réception TCP biaise la moyenne.

3.4.1 Note sur la mise en couche du protocole de transport

Les modèles de mise en couche de protocoles sont souvent présentés comme des modèles de "X sur Y". Dans ces cas, le protocole Y porte les unités de données de protocole du protocole X (et éventuellement les données de contrôle) sur le plan de données de Y, c'est-à-dire que Y est une "couche de convergence". Parmi les exemples, il y a le relais de trame sur ATM, IP sur ATM, et IP sur MPLS. Bien que la mise en couche de X sur Y n'ait rencontré qu'un succès marginal [TENNENHOUSE], [WAKEMAN], il y a eu quelques instances notables où on a pu gagner en efficacité. En particulier, "l'efficacité de X sur Y" peut être réalisée lorsque il y a une sorte "d'isomorphisme" entre le X et le Y (c'est-à-dire, il y a une petite couche de convergence). Dans ces cas, les données de X, et éventuellement du trafic de contrôle, sont "encapsulées" et transportées sur Y. Les exemples incluent le relais de trame sur ATM et le relais de trame, AAL5 ATM et Ethernet sur L2TPv3 [L2TPV3] ; les facteurs de simplification sont ici qu'il n'y a pas d'exigence qu'une horloge partagée

soit récupérée par les points d'extrémité de la communication, et que l'interfonctionnement du plan de contrôle est minimisé. Une solution de remplacement est de faire interfonctionner les plans de contrôle et de données de X de Y ; l'interfonctionnement de plan de contrôle est exposé plus loin.

3.5 Effets de second ordre

IP sur ATM fournit un excellent exemple d'effet de second ordre imprévu. En particulier, l'étude classique de Romanov et Floyd sur le débit TCP [ROMANOV] sur ATM a montré que des grandes mémoires tampon à débit binaire non spécifié (UBR, *Unspecified Bit Rate*) (plus grandes qu'une taille de fenêtre TCP) sont nécessaires pour réaliser des performances raisonnables, que les mécanismes d'élimination de paquet (comme l'élimination précoce de paquet, EDP, *Early Packet Discard*) améliorent l'utilisation effective de la bande passante et que des stratégies de service et d'abandon plus élaborées que FIFO + EPD, telles que la mise en file d'attente et le comptage par circuit virtuel, peuvent être nécessaires au goulet d'étranglement pour s'assurer à la fois d'une haute efficacité et d'un traitement équitable. Bien que toutes les études indiquent clairement qu'une taille de mémoire tampon non inférieure à une fenêtre TCP est nécessaire, la quantité de mémoire tampon supplémentaire exigée dépend naturellement du mécanisme d'élimination de paquet utilisé et la question fait encore l'objet de discussions.

Des exemples de ce type de problème avec la mise en couche abondent dans le réseautage pratique. Considérons, par exemple, l'effet des hypothèses implicites du transport IP sur les couches inférieures. En particulier :

- Perte de paquet : TCP suppose que les pertes de paquet sont des indications d'encombrement, mais parfois les pertes viennent d'une corruption sur une liaison sans fil [RFC3115].
- Paquets déclassés : TCP suppose qu'une réorganisation significative des paquets est une indication d'encombrement. Cela n'est pas toujours le cas [FLOYD2001].
- Durée d'aller-retour : TCP mesure les durées d'aller-retour et suppose qu'une absence d'accusé de réception pendant une certaine période, sur la base du temps d'aller-retour mesuré, est une perte de paquet, et donc une indication d'encombrement [KARN].
- Contrôle d'encombrement : le contrôle d'encombrement TCP suppose implicitement que tous les paquets dans un flux sont traités de la même façon par le réseau, mais ce n'est pas toujours le cas [HANDLEY].

3.6 Instanciation du modèle EOSL avec IP

Alors que IP est proposé comme transport pour presque tout, l'hypothèse de base, que tout va sur IP (EOIP, *Everything over IP*) va améliorer les OPEX et les CAPEX, exige un examen critique. En particulier, bien qu'il soit vrai que de nombreux protocoles peuvent être efficacement transportés sur un réseau IP (en particulier, ces protocoles qui n'ont pas besoin de récupérer la synchronisation entre les points d'extrémité de la communication, comme le relais de trame, Ethernet, et les AAL5 ATM) les principes de simplicité et de mise en couche suggèrent que EOIP pourrait ne pas représenter la stratégie de convergence la plus efficace pour tous les services. Une couche de convergence plus efficace en CAPEX et OPEX pourrait être très inférieure (là encore, ce comportement est prévisible par le principe de simplicité).

Un exemple où EOIP pourrait n'être pas le transport le plus efficace en OPEX et CAPEX serait le cas de services ou protocoles qui auraient besoin de temps de restauration de type SONET (par exemple, 50 ms). Il n'est pas difficile d'imaginer que cela coûterait plus cher de construire et faire fonctionner un réseau IP avec cette sorte de propriété de restauration et de convergence (si même cela était possible) que de construire d'abord le réseau SONET.

4. Éviter la fonction d'interfonctionnement universel

Bien qu'il y ait eu de nombreuses mises en œuvre de fonction d'interfonctionnement universel (UIWF, *Universal Interworking function*) les approches d'IWF ont été problématiques à grande échelle. Ce problème est codifié dans le principe d'intervention minimale [BRYANT] :
"Pour minimiser la portée des informations, et pour améliorer l'efficacité des flux de données à travers la couche d'encapsulation, la charge utile devrait, lorsque c'est possible, être transportée telle que reçue, sans modification."

4.1 Éviter l'interfonctionnement de plan de contrôle

Ce corollaire se comprend mieux dans le contexte de l'espace de solutions intégrées. Dans ce cas, l'architecture et la conception réalisent souvent le pire de tous les mondes possibles. Cela est dû au fait que de telles solutions intégrées ont de mauvaises performances aux deux extrémités du spectre performance/CAPEX/OPEX : les protocoles qui ont le moins de demande de commutation peuvent avoir à supporter le coût de ceux qui sont les plus chers, alors que les protocoles qui ont les exigences les plus contraignantes doivent souvent faire des concessions à ceux qui ont des exigences différentes.

Ajoutez à cela les diverses questions d'interfonctionnement de plan de contrôle et vous aurez de grosses opportunités de défaillances. En résumé, les fonctions d'interopération devraient se restreindre à l'interopération du plan des données et aux encapsulations, et ces fonctions devraient être réalisées à la bordure du réseau.

Comme on l'a décrit auparavant, les modèles d'interopération ont réussi dans les cas où il y a une sorte "d'isomorphisme" entre les couches qui doivent interfonctionner. Le compromis, qui est fréquemment décrit comme le "compromis de l'intégration contre les vaisseaux dans la nuit" a été examiné à divers moments et dans diverses couches de protocoles. En général, il y a peu de cas dans lesquels de telles solutions intégrées aient prouvé leur efficacité. BGP multi-protocoles [RFC2283] est une exception subtilement différente mais notable. Dans ce cas, le plan de contrôle est indépendant du format des données de contrôle. C'est-à-dire qu'aucune conversion des données du plan de contrôle n'est nécessaire, à l'opposé des modèles d'interfonctionnement de plan de contrôle comme l'inter fonctionnement ATM/IP envisagé par certains fabricants de logiciels de commutation, et le soi-disant interfonctionnement "SIN PNNI-MPLS" [ATMMPLS].

5. Commutation par paquet contre commutation de circuit : différences fondamentales

La sagesse conventionnelle estime que la commutation par paquets (PS, *packet switching*) est par nature plus efficace que la commutation de circuits (CS, *circuit switching*) principalement à cause de du gain d'efficacité qui peut être obtenu par le multiplexage statistique et le fait que les décisions d'acheminement et de transmission sont prises indépendamment dans le mode bond par bond [MOLINERO2002]. De plus, on estime généralement que IP est plus simple que la commutation de circuits, et donc devrait être plus économique à déployer et gérer [MCK2002]. Cependant, si on examine cette hypothèse et celles qui s'y rattachent, il émerge un tableau quelque peu différent (voir par exemple [ODLYZKO98]). Les paragraphes qui suivent discutent ces hypothèses.

5.1 PS est elle par nature plus efficace que CS ?

Il est bien connu que la commutation par paquets fait une utilisation efficace d'une bande passante rare [BARAN]. Cette efficacité se fonde sur le multiplexage statistique inhérent à la commutation par paquet. Cependant, on continue de s'interroger sur ce qu'on pense généralement être la faible utilisation des cœurs de réseau Internet. La première question qui se pose est quelle est l'utilisation moyenne actuelle des cœurs de réseau Internet, et comment cela se rapporte t-il à l'utilisation des réseaux vocaux longue distance ? Odlyzko et Coffman [ODLYZKO], [COFFMAN] reportent que l'utilisation moyenne des liaisons dans les réseaux IP était comprise entre 3 % et 20 % (les intranets d'entreprise sont dans la gamme des 3 %, alors que les cœurs de réseau d'Internet commercial fonctionnent dans la gamme des 15 à 20 %). D'un autre côté, l'utilisation moyenne des lignes vocales à longue portée est d'environ 33 %. De plus, pour 2002, l'utilisation moyenne des réseaux optiques (tous services) paraît osciller autour de 11 %, tandis que la moyenne historique est d'approximativement 15 % [ML2002]. La question devient alors pourquoi voit on de tels niveaux d'utilisation, en particulier à la lumière de l'affirmation que la PS est par nature plus efficace que la CS. Les raisons citées par Odlyzko et Coffman incluent que :

(i) le trafic Internet est extrêmement asymétrique et saccadé, mais les liaisons sont symétriques et de capacité fixe (c'est-à-dire, ne connaissent pas de matrice de trafic, ou de capacités de liaison exigées) ;
(ii) il est difficile de prédire la croissance du trafic sur une liaison, de sorte que les opérateurs tendent à ajouter la bande passante de façon agressive ;
(iii) la diminution des prix pour une granularité de bande passante plus grossière fait paraître plus rentable d'ajouter de la capacité par larges paliers.

Les autres facteurs statiques, qui incluent la lourdeur du protocole, d'autres sortes de granularité des équipements, la capacité de restauration, et le délai d'approvisionnement, contribuent tous à la nécessité de "sur provisionner" [MC2001].

5.2 PS est elle plus simple que CS ?

Le principe de bout en bout peut être interprété comme déclarant que la complexité de l'Internet appartient aux bordures. Cependant, les routeurs de cœur de réseau de l'Internet d'aujourd'hui sont extrêmement complexes. De plus, cette complexité progresse avec le débit de la ligne. Comme la complexité relative de la commutation de circuit et par paquet semble avoir résisté à l'analyse directe, on examinera plutôt plusieurs constructions de commutation par paquet et de circuit comme mesure de complexité. Parmi les métriques on peut regarder la complexité du logiciel, la complexité du fonctionnement global, la complexité du matériel, la consommation d'énergie, et la densité. Chacune de ces métriques est examinée ci après.

5.2.1 Complexité de logiciel/microcode

Une mesure de la complexité du logiciel/microcode est le nombre d'instructions requises pour programmer l'appareil.

L'image normale du logiciel pour un routeur Internet exige entre huit et dix millions d'instructions (y compris le microcode) tandis qu'un commutateur de transport normal exige en moyenne environ trois millions d'instructions [MCK2002].

Cette différence de complexité du logiciel a tendance à rendre les routeurs Internet non fiables, et a de notables autres effets de second ordre (par exemple, cela peut prendre beaucoup de temps pour réamorcer un tel routeur). Comme autre point de comparaison, considérons que le commutateur 5ESS de classe 5 d'AT&T (Lucent) qui a un énorme nombre de caractéristiques d'appel, n'exige que le double de nombre de lignes de code qu'un routeur de cœur de réseau Internet [EICK].

Finalement, comme les routeurs sont beaucoup plus du logiciel que du matériel, un autre résultat de la complexité du code est que le coût des routeurs bénéficie moins de la Loi de Moore que les appareils moins consommateurs de logiciel. Cela amène un compromis entre bande passante et appareil qui favorise la bande passante plus que les appareils peu consommateurs de logiciel.

5.2.2 Complexité des macro opérations

Une carte de ligne de routeur Internet doit effectuer de nombreuses opérations complexes, incluant le traitement de l'en-tête de paquet, la plus longue correspondance de préfixe, la génération des messages d'erreur ICMP, le traitement des options d'en-tête IP, et la mise en mémoire tampon du paquet afin que le contrôle d'encombrement TCP soit effectif (cela exige normalement une taille de mémoire tampon proportionnelle au taux de ligne fois le RTT (*round-trip time,* délai d'aller retour) de sorte qu'une mémoire tampon va tenir environ 250 ms de données de paquet). Cela n'inclut pas de filtrage de chemin et de paquet, ni de filtrage de qualité de service ou de réseau privé virtuel.

D'un autre côté, un commutateur de transport a seulement besoin de transposer les intervalles de temps d'entrée en intervalles de temps de sortie et interfaces, et peut donc être considérablement moins complexe.

5.2.3 Complexité du matériel

Une mesure de la complexité du matériel est le nombre de portes logiques sur une carte de ligne [MOLINERO2002]. Considérons le cas d'une carte de ligne de routeur Internet haut débit: Une carte de ligne de POS (*point of synchronisation*, point de synchronisation) de routeur OC192 contient au moins 30 millions de portes en ASIC *(Application-Specific Integrated Circuit,* circuit intégré spécifique d'application) au moins un CPU *(Central Processing Unit,* unité de traitement centrale) 300 Moctets de mémoire tampon de paquet, 2 Moctets de tableau de transmission, et 10 Moctets d'autre mémoire d'état. D'un autre côté, une carte de ligne commutée de transport comparable a 7,5 millions de portes logiques, pas de CPU, pas de mémoire tampon de paquet, pas de tableau de transmission, et une mémoire d'état intégrée dans la puce. La carte de ligne d'un commutateur de transport électronique contient plutôt normalement un trameur SONET, une puce pour transposer les intervalles de temps entrants en intervalles de temps sortants, et une interface avec le moteur de commutation.

5.2.4 Énergie

Comme les commutateurs de transport ont été traditionnellement construits à partir de plus simples composants matériels, ils consomment aussi moins d'énergie [PMC].

5.2.5 Densité

Les commutateurs de transport de plus forte capacité ont environ quatre fois la capacité d'un routeur IP [CISCO], [CIENA], pour un prix divisé par trois par Gigabit/s. La technologie optique (OOO) pousse plus loin cette différence de complexité (par exemple, lasers réglables, commutateurs MEM. Par exemple, les multiplexeurs[CALIENT]), et DWDM fournissent la technologie pour construire des commutateurs de transport à capacité extrêmement élevée, à faible consommation.

Une métrique en rapport est l'empreinte de pied physique. En général, du fait de leur densité supérieure, les commutateurs de transport ont une plus petite empreinte de pied physique "par gigabit".

5.2.6 Coûts fixes contre coûts variables

La commutation par paquet pourrait sembler avoir des coûts variable élevés, signifiant que cela coûte plus d'envoyer le n^{e} élément d'information en utilisant la commutation par paquet que cela ne coûterait dans un réseau à commutation de circuit. Une grande partie de cet avantage est due à la nature relativement statique de la commutation de circuit, par

exemple, la commutation de circuit peut tirer parti de l'arrivée préprogrammée des informations pour éliminer des opérations à effectuer sur les informations entrantes. Par exemple, dans le cas de la commutation de circuit, il n'est pas nécessaire de mettre en mémoire tampon les informations entrantes, d'effectuer la détection de boucle, de résoudre le prochain bond, de modifier les champs dans l'en-tête de paquet, et ainsi de suite. Finalement, de nombreux réseau de commutation de circuit combinent une configuration relativement statique avec des plans de contrôle hors bande (par exemple, SS7) ce qui simplifie considérablement la commutation au plan des données. La limite est que lorsque les débits de données deviennent élevés, il devient de plus en plus complexe de commuter les paquets, alors que la commutation de circuit s'adapte de façon plus ou moins linéaire.

5.2.7 Qualité de service

Bien que les composants d'une solution complète pour la qualité de service de l'Internet, incluant le contrôle de l'admission d'appel, une classification efficace des paquets, et des algorithmes de programmation, aient fait l'objet de recherches approfondies et de normalisation depuis plus de dix ans, la qualité de service signalée de bout en bout pour l'Internet n'est pas devenue une réalité. Autrement, la QS a fait partie de l'infrastructure de commutation de circuit presque depuis son commencement. D'un autre côté, la QS est normalement déployée pour déterminer les disciplines de mise en file d'attente à utiliser lorsque la bande passante est insuffisante pour supporter le trafic. Mais à la différence du trafic vocal, l'abandon de paquet ou de sévères retards peuvent avoir des conséquences beaucoup plus sérieuses sur le trafic TCP du fait de ses boucles de rétroaction sensibles à l'encombrement (en particulier, la temporisation/démarrage lent de TCP).

5.2.8 Souplesse

Une métrique un peu plus difficile à quantifier est la souplesse inhérente à l'Internet. Bien que la souplesse de l'Internet ait conduit à sa croissance rapide, cette souplesse a un coût relativement élevé à la marge : le besoin d'un personnel de soutien très entraîné. Une règle standard d'approximation est que dans l'établissement d'une entreprise, une seule personne de soutien suffit à fournir le service du téléphone pour un groupe, alors qu'on a besoin de dix experts en réseautage informatique pour satisfaire aux exigences de réseautage du même groupe [ODLYZKO98A]. Ce phénomène est aussi décrit dans [PERROW].

5.3 Complexité relative

La complexité relative de calcul de la commutation de circuit par rapport à la commutation par paquet a été difficile à décrire en termes formels [PARK]. À ce titre, les paragraphes précédents cherchent à décrire la complexité en termes d'éléments observables. Avec cette idée en tête, il est clair que le facteur fondamental qui produit l'augmentation de complexité souligné ci-dessus est l'indépendance bond par bond (HBHI, *hop-by-hop independence*) inhérente à l'architecture IP. Cela s'oppose aux architectures de bout en bout telles que l'ATM ou le relais de trame.

[WILLINGER2002] décrit ce phénomène en termes d'exigence de robustesse de la conception originale de l'Internet, et comment cette exigence a conduit à la complexité du réseau. En particulier, ils décrivent une spirale de "complexité/robustesse" dans laquelle les augmentations de complexité créent d'autres sensibilités encore plus sérieuses, qui exigent alors une robustesse supplémentaire (d'où la spirale).

La leçon importante de ce paragraphe est que le principe de simplicité, bien qu'applicable à la commutation de circuit aussi bien qu'à la commutation par paquets, est crucial pour contrôler la complexité (et donc les propriétés d'OPEX et de CAPEX) des réseaux par paquets. Cette idée est renforcée par l'observation qu'alors que la commutation par paquets est une discipline plus jeune et moins mûre que la commutation de circuit, la tendance des commutateurs de paquet est à des cartes de ligne plus complexes, alors que la complexité des commutateurs de circuit apparaît suivre de façon linéaire les taux de ligne et la capacité agrégée.

5.3.1 HBHI et le défi OPEX

Par suite de HBHI, on doit approcher les réseaux IP d'une façon fondamentalement différente de celle des réseaux fondés sur le circuit. En particulier, le défi majeur des OPEX qui se pose au réseau IP est que de déboguer un réseau IP à grande échelle exige toujours un haut niveau d'expertise et d'intelligence, là encore du fait de l'indépendance bond par bond inhérente à une architecture par paquets (noter encore que cette indépendance bond par bond n'est pas présente dans des réseaux à circuit virtuel tels que ATM ou de relais de trame). Par exemple, on peut devoir visiter un grand ensemble de routeurs pour seulement découvrir que le problème est extérieur au réseau. De plus, les outils de déboguage utilisés pour faire le diagnostic des problèmes sont aussi complexes et quelque peu primitifs. Finalement, IP doit traiter avec des gens qui ont des problèmes avec leur DNS ou leur messagerie électronique ou les nouvelles ou quelque nouvelle application, tandis que ce n'est habituellement pas le cas pour TDM/ATM/etc. Dans le cas de IP, cela peut être facilité en améliorant l'automation (noter que beaucoup de ce que l'on mentionne ici est posé au consommateur). En général, il y a de

nombreuses variables externes au réseau qui affectent les OPEX.

Finalement, il est important de noter que la relation quantitative entre CAPEX, OPEX, et la complexité inhérente au réseau n'est pas bien comprise. En fait, il n'y a pas de métrique quantitative d'acceptation générale pour décrire la complexité d'un réseau, de sorte qu'on élude généralement la définition d'une relation précise entre CAPEX, OPEX, et complexité.

6. Le mythe du sur approvisionnement

Comme noté dans [MC2001] et ailleurs, beaucoup de la complexité qu'on observe dans l'Internet d'aujourd'hui est tournée vers une utilisation accrue de bande passante. Il en résulte que le désir des ingénieurs réseau de tenir l'utilisation du réseau au dessous de 50 % a été appelé "sur approvisionnement". Cependant, cette utilisation du terme sur approvisionnement est un abus de langage. Dans les cœurs de réseau de l'Internet moderne, la capacité inutilisée est en fait plutôt une capacité de protection. En particulier, on peut voir cela comme une "protection 1:1 à la couche IP". Vu de cette façon, on voit qu'un réseau IP provisionné pour fonctionner à 50 % d'utilisation n'est pas plus sur approvisionné que le réseau SONET normal. Cependant, les avantages importants qui découlent d'un réseau IP provisionné de cette façon incluent une vitesse de transmission proche de la vitesse de la lumière et une perte de paquet proche de zéro [FRALEIGH]. Ces bénéfices peuvent être vus comme un "effet collatéral" de l'approvisionnement de protection de 1:1.

Il y a aussi d'autres raisons, liées à la théorie des systèmes, pour fournir un approvisionnement de protection de type 1:1. La plus notable de ces raisons est que les réseaux à commutation par paquets avec boucles de contrôle dans la bande peuvent devenir instables et peuvent rencontrer des oscillations et désynchronisations en cas d'encombrement. Des interactions dynamiques complexes et non linéaires du trafic signifient que l'encombrement d'une partie du réseau va s'étendre à d'autres parties du réseau. Lorsque des paquets de protocole d'acheminement sont perdus du fait d'encombrement ou de surcharge de processeur d'acheminement, cela cause un état d'acheminement incohérent, et il peut en résulter des boucles de trafic, des trous noirs, et une perte de connexité. Donc, bien que le multiplexage statistique puisse en théorie donner une plus forte utilisation du réseau, en pratique, pour conserver des performances cohérentes et un réseau raisonnablement stable, la dynamique des cœurs de réseau de l'Internet plaide en faveur de l'approvisionnement 1:1 et son effet collatéral pour garder le réseau stable et le délai faible.

7. Le mythe de cinq neufs

Paul Baran, dans son article classique, "Quelques perspectives sur les réseaux -- passé, présent et avenir", déclarait que "les courbes d'équivalence entre coût et fiabilité du système suggèrent que les systèmes les plus fiables peuvent être construits d'élément relativement non fiables et donc de faible coût, si le problème est celui de la fiabilité du système pour le moindre coût global du système" [BARAN77].

Aujourd'hui, on se réfère à ce phénomène comme au "mythe des cinq neufs".
Précisément, ce qu'on appelle fiabilité des cinq neufs dans les éléments de réseau par paquets est considéré comme un mythe pour les raisons suivantes : d'abord, comme 80 % des pannes non programmées sont causées par des erreurs humaines ou de processus [SCOTT], il n'y a qu'une fenêtre de 20 % à optimiser. Donc, afin d'augmenter la fiabilité des composants, on ajoute de la complexité (l'optimisation conduit fréquemment à la complexité) qui est à la base de 80 % des pannes imprévues. Cela réduit effectivement la fenêtre de 20 % (c'est-à-dire qu'on augmente la probabilité de défaillance humaine et de processus). Ce phénomène est aussi caractérisé comme spirale "complexité/robustesse" [WILLINGER2002], dans laquelle les augmentations de complexité créent d'autres sensibilité encore plus sérieuses, qui à leur tour exigent une robustesse supplémentaire, et ainsi de suite (d'où la spirale).

La conclusion est alors que bien qu'un système comme l'Internet puisse atteindre une fiabilité du style cinq neufs, il n'est pas souhaitable (et vraisemblablement impossible) d'essayer de faire atteindre à tout composant individuel, en particulier les plus complexes, cette fiabilité standard.

8. Loi de proportionnalité des composants architecturaux

Comme on l'a noté à la section précédente, la complexité de calcul des réseaux à commutation de paquets comme l'Internet s'est révélée difficile à décrire en termes formels. Cependant, une définition intuitive, de haut niveau, de la complexité de l'architectural pourrait être que la complexité d'une architecture est proportionnelle à son nombre de composants, et que la probabilité de réaliser une mise en œuvre stable d'une architecture est inversement proportionnelle à son nombre de composants. Comme on l'a décrit ci-dessus, les composants incluent les éléments discrets tels que les éléments de matériel, les exigences d'espace et d'énergie, ainsi que les logiciels, microcodes, et protocoles qu'ils mettent en œuvre.

Dit de façon plus abstraite :
Soit
A une représentation de l'architecture A,
|A| le nombre de composants distincts sur le chemin de livraison du service de l'architecture A,
w une fonction d'accroissement monotone,
P la probabilité d'une mise en œuvre stable d'une architecture,

on a alors

Complexité(A) = O(w(|A|))
P(A) = O(1/w(|A|))

où
O(f) = {g:N->R | il existe c > 0 et n tel que g(n) < c*f(n)}

[C'est à dire, O(f) comprend l'ensemble des fonctions g pour lesquelles il existe une constante c et un nombre n, tels que g(n) soit inférieur ou égal à c*f(n) pour tout n. C'est à dire, O(f) est l'ensemble de toutes les fonctions qui ne croissent pas plus vite que f, sans considération des facteurs constants]

Il est intéressant de noter que le modèle de tolérance hautement optimisée (HOT, *Highly Optimized Tolerance*) [HOT] tente de caractériser la complexité en termes généraux (HOT est une tentative récente de développer un cadre général d'étude de la complexité, et est un membre de la famille d'abstractions généralement appelée "la nouvelle science de la complexité" ou "systèmes complexes adaptatifs"). La tolérance, dans la sémantique de HOT, signifie que la "robustesse dans les systèmes complexes est une quantité limitée et soumise à des contraintes, qui doit être gérée attentivement et protégée." Un des points de concentration du modèle HOT est de caractériser les distributions à grosse queue telles que la Complexité(A) dans l'exemple précédent (d'autres exemples portent sur des feux de forêt, des pannes de courant, et les distributions de trafic de l'Internet). En particulier, la Complexité(A) tente de cartographier l'extrême hétérogénéité des parties du système (Internet) et les effets de leur organisation en réseaux hautement structurés, avec des hiérarchies et des échelles multiples.

8.1 Chemins de livraison de service

La Loi de proportionnalité des composants architecturaux (ACPL, *Architectural Component Proportionality Law*) déclare que la complexité d'une architecture est proportionnelle à son nombre de composants.

COROLLAIRE : Minimiser le nombre de composants dans un chemin de livraison de service, où le chemin de livraison de service peut être un chemin de protocole, un chemin logiciel ou un chemin physique.

Ce corollaire est une conséquence importante de l'ACPL, car le chemin entre un consommateur et le service désiré est particulièrement sensible au nombre et à la complexité des éléments du chemin. Cela est dû au fait que le "lissage" de la complexité qu'on va trouver à de hauts niveaux d'agrégation [ZHANG] manque lorsque on se rapproche de la bordure, ainsi que lorsque on a des interactions complexes avec les systèmes de soutien et les systèmes de gestion de la relation client. Des exemples d'architectures qui n'ont pas trouvé de marché à cause de cet effet sont les systèmes de gestion de la relation client fondés sur TINA, et les architectures de services fondées sur CORBA/TINA. La leçon fondamentale à en tirer est que les seules possibilités pour le déploiement de ces systèmes étaient des "déploiements à échelle limitée (tels que) lorsque Starvision peut éviter de se confronter aux problèmes majeurs d'adaptabilité non résolus", ou "auraient autrement besoin d'investissements massifs (comme le gestionnaire d'objets distribués de niveau transporteur (ORB) construit presque à partir du néant)" [TINA]. En d'autres termes, ces systèmes avaient des chemins de livraison de service complexes, et étaient trop complexes pour être facilement déployés.

9. Conclusions

Le présent document essaye de codifier les principes de l'architecture de l'Internet qui sont bien admis. En particulier, le principe unificateur qui est décrit ici est mieux exprimé par le principe de simplicité, qui déclare que la complexité doit être contrôlée si on espère dimensionner efficacement un objet complexe. L'idée que la simplicité peut conduire par elle-même à une forme optimale a été un thème courant dans l'histoire, et a été formulée de bien d'autres façons et dans de nombreux environnements. Par exemple, considérons la maxime connue sous le nom de rasoir d'Occam, qui a été formulée par le philosophe médiéval anglais et moine franciscain William d'Ockham (ca. 1285-1349) et déclare "Pluralitas non est ponenda sine neccesitate" ou "la multiplicité ne devrait pas être postulée sans nécessité." (d'où l'appellation qu'on donne parfois au rasoir d'Occam de "principe de nécessité de la multiplicité " et de "principe de simplicité"). Une formulation peut-être un peu plus contemporaine du rasoir d'Occam déclare que la plus simple explication d'un phénomène est celle qui est par nature préférable. D'autres formulations de la même idée dans le principe "pourquoi faire simple quand on peut faire

compliqué" (KISS, *Keep It Simple Stupid*) et le principe du moindre étonnement (l'assertion que le système le plus fonctionnel est celui qui étonne le moins souvent les utilisateurs). [WILLINGER2002] donne un exposé plus théorique de la "robustesse par la simplicité", et dans un exposé sur le RTPC, [KUHN87] déclare que dans la plupart des systèmes, "un compromis peut être trouvé entre la simplicité des interactions et l'étroitesse du couplage".

Lorsque on l'applique aux architectures de réseaux à commutation de paquets, le principe de simplicité a des implications que certains peuvent considérer comme hérétiques, par exemple, que les approches très convergentes sont vraisemblablement moins efficaces que les solutions "moins convergentes". Autrement dit, la couche "optimale" de convergence peut être bien plus bas dans la pile de protocoles que ce qu'on croit habituellement. De plus, l'analyse ci-dessus conduit à plusieurs conclusions qui sont contraires aux conventions traditionnelles sur les réseaux de paquets. Celle qui est peut-être la plus significative est la croyance que la commutation de paquets est plus simple que la commutation de circuits. Cette croyance a conduit à des conclusions telles que "comme le paquet est plus simple que le circuit, cela doit coûter moins cher en fonctionnement". Cette étude prouve le contraire. En particulier, en examinant les métriques décrites ci-dessus, on trouve que la commutation de paquets est plus complexe que la commutation de circuits. Il est intéressant de voir que cette conclusion repose sur le fait que l'OPEX normalisé pour les réseaux de données est normalement significativement supérieur à celui des réseaux vocaux [ML2002].

Finalement, la conclusion importante de ce travail est que pour les réseaux de paquets qui sont à l'échelle de l'Internet d'aujourd'hui ou plus grands, on doit s'en tenir aux solutions les plus simples possibles si on espère construire des infrastructures rentables. L'idée est établie avec éloquence dans [DOYLE2002] : "L'évolution des protocoles peut conduire à une spirale robustesse/complexité/fragilité où la complexité ajoutée à la robustesse ajoute aussi de nouvelles fragilités, ce qui à son tour conduit à de nouvelles complexités qui alimentent donc la spirale." C'est exactement pour éviter ce phénomène qu'est conçu le principe de simplicité.

10. Considérations pour la sécurité

Le présent document n'affecte directement la sécurité d'aucun protocole Internet existant. Cependant, l'adhésion au principe de simplicité a un effet direct sur notre capacité à mettre en œuvre des systèmes sécurisés. En particulier, lorsque la complexité d'un système croît, il devient plus difficile à modéliser et analyser, et donc, il devient plus difficile de trouver et comprendre les implications pour la sécurité qui sont inhérentes à son architecture, sa conception et sa mise en œuvre.

11. Remerciements

Beaucoup des idées pour la comparaison de la complexité des réseaux à commutation de circuit à celle des réseaux à commutation de paquets ont été inspirées de conversations avec Nick McKeown. Scott Bradner, David Banister, Steve Bellovin, Steward Bryant, Christophe Diot, Susan Harris, Ananth Nagarajan, Andrew Odlyzko, Pete et Natalie Whiting, et Lixia Zhang ont apporté de nombreux commentaires utiles.

12. Références

[AHUJA] "The Impact of Internet Policy and Topology on Delayed Routing Convergence", Labovitz, et. al. Infocom, 2001.

[ATMMPLS]"ATM-MPLS Interworking Migration Complexities Issues and Preliminary Assessment", School of Interdisciplinary Computing and Engineering, University of Missouri-Kansas City, avril 2002

[BARAN] "On Distributed Communications", Paul Baran, Rand Corporation Memorandum RM-3420-PR, http://www.rand.org/publications/RM/RM3420", août, 1964.

[BARAN7] "SOME PERSPECTIVES ON NETWORKS--PAST, PRESENT AND FUTURE", Paul Baran, Information Processing 77, North-Holland Publishing Company, 1977,

[BRYANT] "Protocol Layering in PWE3", Bryant et al, Travail en cours.

[CAIDA] http://www.caida.org

[CALLIENT] http://www.calient.net/home.html

[CARLSON] "Complexity and Robustness", J.M. Carlson and John Doyle, Proc. Natl. Acad. Sci. USA, Vol. 99, Suppl.

1, 2538-2545, 19 février 2002. http://www.pnas.org/cgi/doi/10.1073/pnas.012582499

[CIENA] "CIENA Multiwave CoreDiretor", http://www.ciena.com/downloads/products/coredirector.pdf

[CISCO] http://www.cisco.com

[CLARK] "The Design Philosophy of the DARPA Internet Protocols", D. Clark, Proc. of the ACM SIGCOMM, 1988.

[COFFMAN] "Internet Growth: Is there a 'Moores Law' for Data Traffic", K.G. Coffman and A.M. Odlyzko, pp. 47-93, Handbook of Massive Data Stes, J. Elli, P. M. Pardalos, and M. G. C. Resende, Editors. Kluwer, 2002.

[DOYLE2002] "Robustness and the Internet: Theoretical Foundations", John C. Doyle, et. al. Travail en cours..

[EICK] "Visualizing Software Changes", S.G. Eick, et al, National Institute of Statistical Sciences, Technical Report 113, décembre 2000.

[FLOYD] "The Synchronization of Periodic Routing Messages", Sally Floyd and Van Jacobson, IEEE ACM Transactions on Networking, 1994.

[FLOYD2001] "A Report on Some Recent Developments in TCP Congestion Control, IEEE Communications Magazine, S. Floyd, avril 2001.

[FRALEIGH] "Provisioning IP Backbone Networks to Support Delay-Based Service Level Agreements", Chuck Fraleigh, Fouad Tobagi, and Christophe Diot, 2002.

[GRIFFIN] "What is the Sound of One Route Flapping", Timothy G. Griffin, IPAM Workshop on Large-Scale Communication Networks: Topology, Routing, Traffic, and Control, mars 2002.

[HANDLEY] "On Inter-layer Assumptions (A view from the Transport Area), slides from a presentation at the IAB workshop on Wireless Internetworking", M. Handley, mars 2000.

[HOT] J.M. Carlson and John Doyle, Phys. Rev. E 60, 1412-1427, 1999.

[ISO10589] "Intermediate System to Intermediate System Intradomain Routing Exchange Protocol (IS-IS)".

[JACOBSON] "Congestion Avoidance and Control", Van Jacobson, Proceedings of ACM Sigcomm 1988, pp. 273-288.

[KARN] "TCP vs Link Layer Retransmission" in P. Karn et al., Advice for Internet Subnetwork Designers, Travail en cours.

[KUHN87] "Sources of Failure in the Public Switched Telephone Network", D. Richard Kuhn, EEE Computer, Vol. 30, No. 4, avril, 1997.

[L2TPV3] Lan, J., et. al., "Layer Two Tunneling Protocol (Version 3) -- L2TPv3", Travail en cours..

[MC2001] "U.S Communications Infrastructure at A Crossroads: Opportunities Amid the Gloom", McKinsey&Company for Goldman-Sachs, août 2001.

[MCK2002] Nick McKeown, communication personnelle avril, 2002.

[ML2002] "Optical Systems", Merril Lynch Technical Report, avril, 2002.

[MOLINERO2002] "TCP Switching: Exposing Circuits to IP", Pablo Molinero-Fernandez and Nick McKeown, IEEE janvier, 2002.

[NAVE] "The influence of mode coupling on the non-linear evolution of tearing modes", M.F.F. Nave, et al, Eur. Phys. J. D 8, 287-297.

[NEUMANN] "Cause of AT&T network failure", Peter G. Neumann, http://catless.ncl.ac.uk/Risks/9.62.html#subj2

[ODLYZKO] "Data networks are mostly empty for good reason", A.M. Odlyzko, IT Professional 1 (no. 2), pp. 67-69, mars/avril 1999.

[ODLYZKO98A] "Smart and stupid networks: Why the Internet is like Microsoft". A. M. Odlyzko, ACM Networker, 2(5), décembre 1998.

[ODLYZKO98] "The economics of the Internet: Utility, utilization, pricing, and Quality of Service", A.M. Odlyzko, juilllet 1998. http://www.dtc.umn.edu/~odlyzko/doc/networks.html

[PARK] "The Internet as a Complex System: Scaling, Complexity and Control", Kihong Park and Walter Willinger, AT&T Research, 2002.

[PERROW] "Normal Accidents: Living with High Risk Technologies", Basic Books, C. Perrow, New York, 1984.

[PMC] "The Design of a 10 Gigabit Core Router Architecture", PMC-Sierra, http://www.pmc-sierra.com/products/diagrams/CoreRouter_lg.html

[RFC1629] R. Colella, R. Callon, E. Gardner et Y. Rekhter, "Lignes directrices pour allocations de NSAP OSI dans l'Internet", mai 1994.

[RFC1925] R. Callon, "Les douze vérités du réseautage", 1er avril 1996. *(Information)*

[RFC1958] B. Carpenter, éd., "Principes de l'architecture de l'Internet", juin 1996. (*MàJ par* RFC3439) *(Information)*

[RFC2283] T. Bates, R. Chandra, D. Katz, Y. Rekhter, "Extensions multiprotocoles pour BGP-4", février 1998. (*Obsolète, voir* RFC2858) *(P.S.)*

[RFC3155] S. Dawkins, G. Montenegro, M. Kojo et N. Vaidya, "Implications des liaisons avec des erreurs sur les performances de bout en bout", août 2001. (BCP0050)

[ROMANOV] "Dynamics of TCP over ATM Networks", A. Romanov, S. Floyd, IEEE JSAC, vol. 13, No 4, pp.633-641, mai 1995.

[SALTZER] "End-To-End Arguments in System Design", J.H. Saltzer, D.P. Reed, and D.D. Clark, ACM TOCS, Vol 2, Number 4, novembre 1984, pp 277-288.

[SCOTT] "Making Smart Investments to Reduce Unplanned Downtime", D. Scott, Tactical Guidelines, TG-07-4033, Gartner Group Research Note, mars 1999.

[SPILLMAN] "The Law of Diminishing Returns:, W. J. Spillman and E. Lang, 1924.

[STALLINGS] "Data and Computer Communications (2nd Ed)", William Stallings, Maxwell Macmillan, 1989.

[TENNENHOUSE] "Layered multiplexing considered harmful", D. Tennenhouse, Proceedings of the IFIP Workshop on Protocols for High-Speed Networks, Rudin ed., North Holland Publishers, mai 1989.

[THOMPSON] "Nonlinear Dynamics and Chaos". J.M.T. Thompson and H.B. Stewart, John Wiley and Sons, 1994, ISBN 0471909602.

[TINA] "What is TINA and is it useful for the TelCos?", Paolo Coppo, Carlo A. Licciardi, CSELT, EURESCOM Participants in P847 (FT, IT, NT, TI)

[WAKEMAN] "Layering considered harmful", Ian Wakeman, Jon Crowcroft, Zheng Wang, and Dejan Sirovica, IEEE Network, janvier 1992, p. 7-16.

[WARD] "Custom fluorescent-nucleotide synthesis as an alternative method for nucleic acid labeling", Octavian Henegariu*, Patricia Bray-Ward and David C. Ward, Nature Biotech 18:345-348 (2000).

[WILLINGER2002] "Robustness and the Internet: Design and evolution", Walter Willinger and John Doyle, 2002.

[ZHANG] "Impact of Aggregation on Scaling Behavior of Internet Backbone Traffic", Sprint ATL Technical Report TR02-ATL-020157 Zhi-Li Zhang, Vinay Ribeiroj, Sue Moon, Christophe Diot, février 2002.

13. Adresse des auteurs

Randy Bush
mél : randy@psg.com

David Meyer
mél : dmm@maoz.com

14. Déclaration complète de droits de reproduction

Remerciement

Le financement de la fonction d’édition des RFC est actuellement fourni par la Internet Society.