Détection de scans
Configuration et principes de fonctionnement

Rappels sur le collecteur

Une documentation complète existant sur ce domaine (*Collecteur & fichier de configuration netmet.conf*), nous nous contenterons de rappeler certains principes de fonctionnement du collecteur qui nous paraissent important pour la bonne compréhension de ce qui va suivre. Il est conseillé au lecteur non averti de se reporter à la documentation existante.

Chaque service netMET (métrologie, statistiques et sécurité) a son/ses collecteurs spécifiques. Les collecteurs sont chargés de réceptionner les paquets *UDP NetFlow* (processus *netMETcll*) en provenance d'un routeur ou du duplicateur de flows netMET (*netMETdup*), de les épurer de leurs informations inutiles (pour la problématique netMET), puis de gérer en temps réel un certains nombre de compteurs de métrologie (processus *netMETacc*).

Le fonctionnement du collecteur et de ses deux processus ainsi que la circulation des différentes informations mise en oeuvre sont illustrés par le schéma Figure -1-.

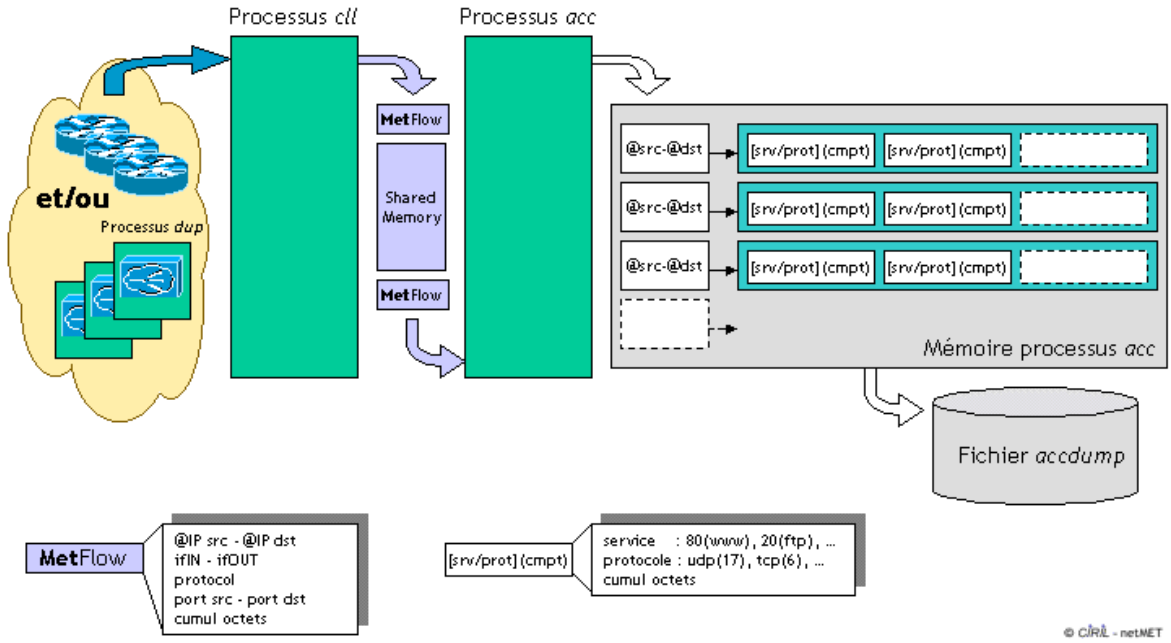


Figure -1- Fonctionnement de *netMETcll* et *netMETacc*

Lors de l'installation, les collecteurs netMET dédiés à la sécurité ont été installé en mode non-agrégé, c'est-à-dire qu'ils n'agrègent pas les *NetFlow* en provenance/destination d'une interface particulière du routeur avec une adresse symbolique (adresse d'agrégation ou "trou noir"). Ainsi, les communications entre les machines de l'Internet et les machines "intérieures" seront perçues distinctement, il est donc évident que le nombre de couple (@src-@dst) peut facilement augmenter... La taille du processus d'accounting ainsi que celle du fichier *dump* risque en mode non-agrégé d'être importante mais toutes les informations utiles à la détection de scans sont ainsi conservées. Il est en particulier impossible d'identifier la machine source d'un scan provenant de l'Internet lorsque les adresses sont agrégées car celle-ci est alors remplacée par l'adresse symbolique.

Principes de détection de scans

On l'aura compris, la détection de scans se fait à l'aide des fichiers *dump* produit par les processus d'accounting du service de sécurité. On distinguera deux sortes de scans différents, le scan en "hauteur" et le scan en "largeur".

- scan en "hauteur" : il y a scan en "hauteur" lorsqu'une machine communique avec plus de n machines d'un même réseau. Ce phénomène se traduit au niveau des structures de données manipulées par un grand nombre d'entrées @src-@dst différentes.

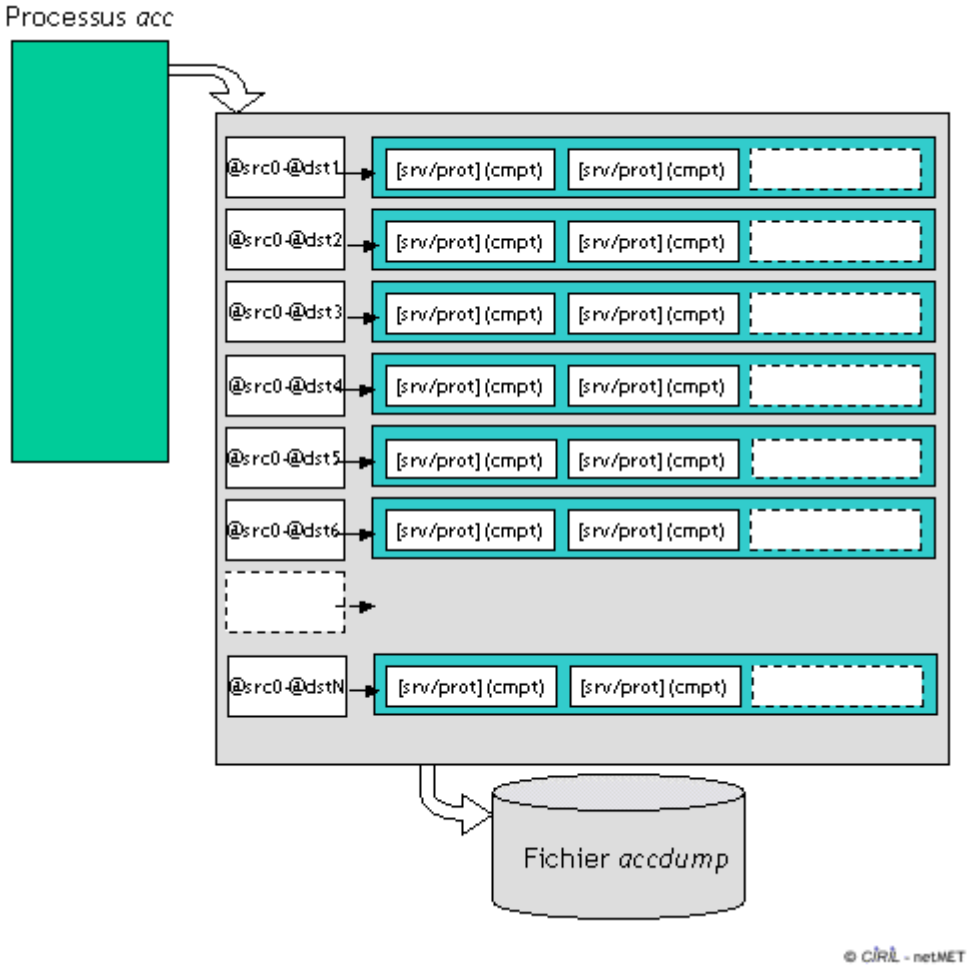


Figure -2- Table d'accounting typique d'un "scan en hauteur"

- scan en "largeur" : il y a scan en "largeur" lorsqu'une machine communique avec une seule machine destination mais sur un grand nombre de service. Ainsi, pour une entrée @src-@dst on trouvera une longue liste de services d'où le terme de scan en "largeur".

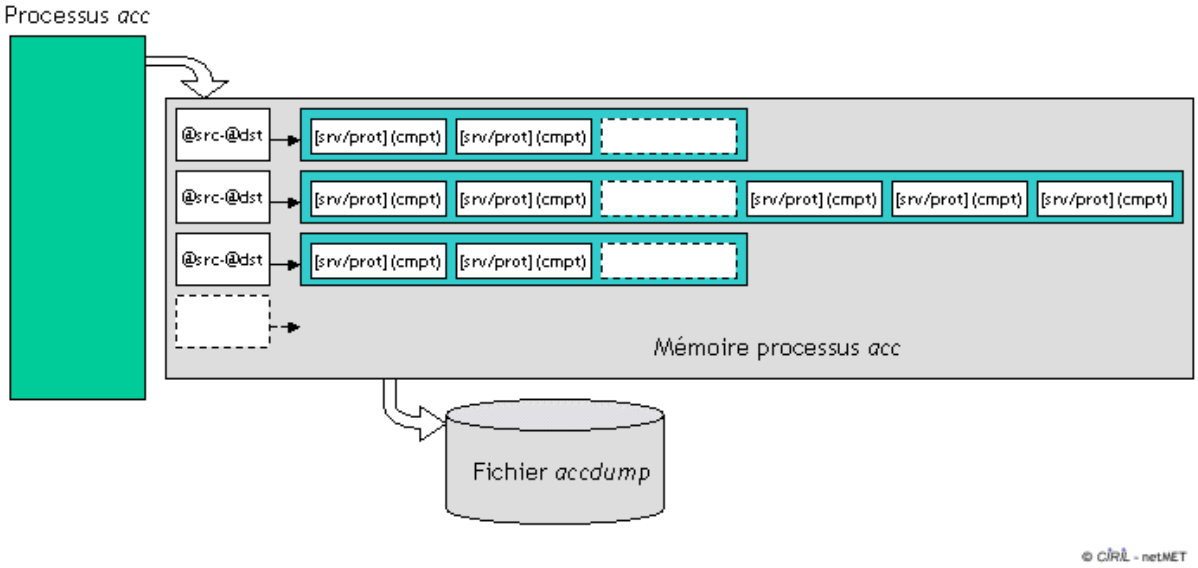


Figure - 3- Table d'accounting typique d'un "scan en largeur"

Nous considérerons donc comme suspectes toutes machines communiquant avec plus de n machines d'un même réseau ou toutes machines demandant plus de m services à une même adresse IP. La seule difficulté est de choisir les nombres n et m afin de bien distinguer un comportement normal, d'une tentative de scan : il n'est a priori pas étrange qu'une machine communique avec 20 machines d'un réseau mais il est étrange que cette machine communique avec 255 machines d'un réseau de classe C. Ainsi, pour la détection de scan en hauteur, le seuil doit être fonction de la taille du réseau : un scan en hauteur sur un réseau de classe C n' a pas le même impact qu'un scan sur un réseau de classe B. De même, le nombre minimum de services demandés entrainant un "avis de scan" doit être choisi de manière à exclure le fonctionnement normal des serveurs mais doit tout de même permettre la détection des comportements suspects.

Configuration et Fonctionnement

La configuration du service de détection de scan consiste essentiellement en la configuration du fichier "explt.conf", que l'on trouve dans le répertoire "netMet/etc" sous le répertoire de travail.

Les seuils choisis par défaut déclenchant une alerte de scan en hauteur sont :

- plus des 2/3 (=176) des machines d'un réseau de classe C ont été contactées par une même source
- plus de 800 machines du même réseau de classe B ou A ont été contactées par la même machine

Nous avons choisi de fixer un seul seuil commun pour les réseaux de classe A et B car aucun organisme (tout du moins en France) ne possède de réseau de classe A, et, lorsque plus de 800 machines d'un même réseau ont été contactées par la même machine-source, sur le même type de service, il n'y a plus aucun doute à avoir sur les intentions du suspect, il est alors inutile d'attendre qu'un plus grand nombre de machine soit touché.

Le seuil fixé par défaut déclenchant une alerte de scan en largeur est atteint lorsque plus de 500 services ont été demandés pour un couple @src-@dst donné. Comme nous l'avons dit précédemment, ce seuil doit éviter de retrouver systématiquement les plus gros serveurs parmis la liste des machines subissant des scans en largeur.

<table>
<tr><th>Etape</th><th>Commandes</th><th>Explications</th></tr>
<tr><td></td><td>netmet> cd /home/netmet/netMet</td><td>Positionnement dans le répertoire de travail</td></tr>
<tr><td></td><td>netmet> vi etc/explt.conf</td><td>Dans le paragraphe DESIGN - DESIGN- DESIGN, initialisez par vos valeurs ou libellés les variables suivantes :

<table>
<tr><td>NETMET_SCANS_THRESHOLD_B_A</td><td>Nombre minimum de machines touchées dans un réseau de classe B ou A entrainant un avertissement de scan en hauteur.
(par défaut: 800)</td></tr>
<tr><td>NETMET_SCANS_THRESHOLD_C</td><td>Nombre minimum de machines touchées dans un réseau de classe C entrainant un avertissement de scan en hauteur.
(par défaut: 176)</td></tr>
<tr><td>NETMET_SCANS_PORT</td><td>Nombre minimum de services demandés pour une même machine entrainant un avertissement de scan en largeur.(par défaut: 500)</td></tr>
</table>
</td></tr>
</table>

Le service de détection de scan s'active automatiquement une fois par jour. Il fait un compte-rendu le jour j des scans détectés le jour j-1, le rapport généré est accessible via le web et est rangé dans html/aaaa-mm/aaaa-mm-jj/SCANS/ sous le nom de SCANS.html.

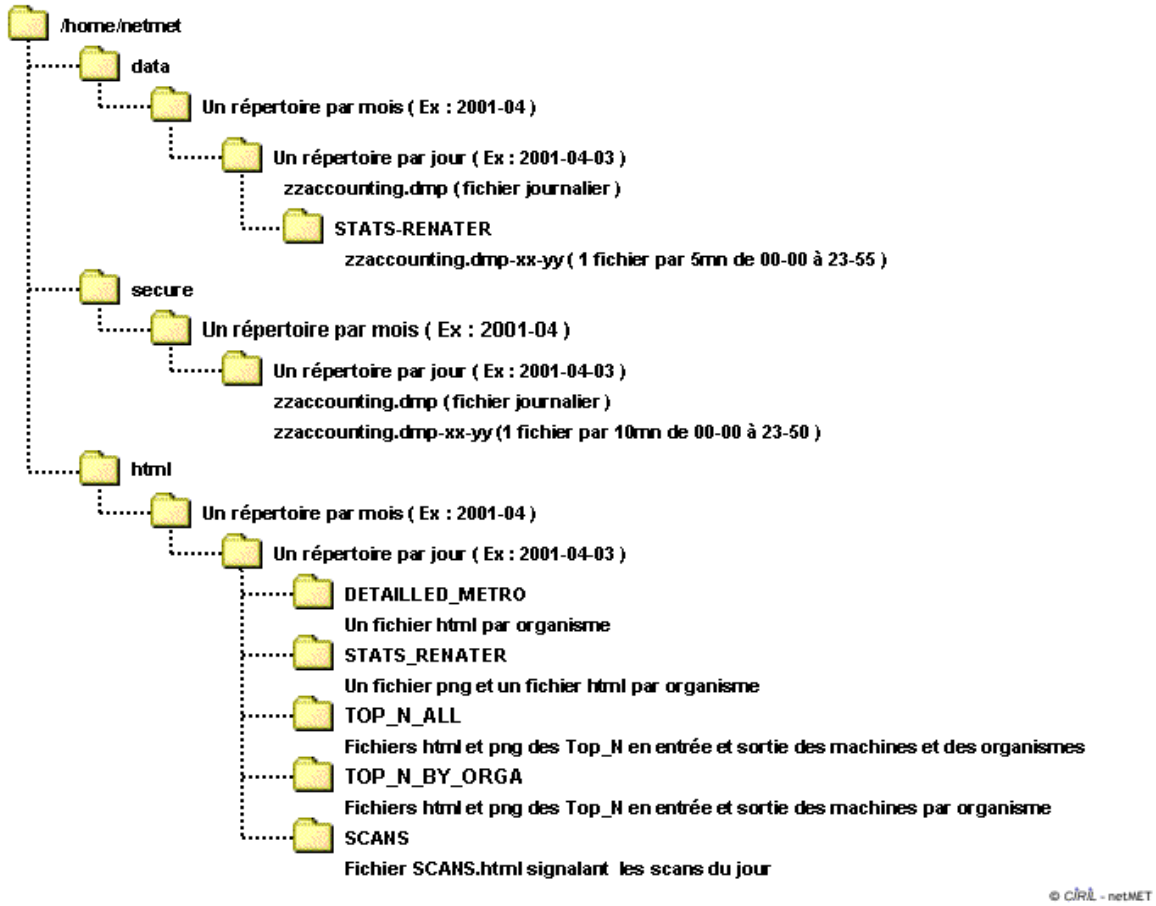


Figure -4- Arborescence des données et résultats

La recherche des scans se déroule en plusieurs étapes, dans un premier temps on reconvertit les informations contenues dans le fichier *dump* correspondant aux 24 heures de collecte sous le format "@src-@réseau_dst srv/prot (Nb-machines)" à l'aide de l'exécutable netMETscn (/home/netmet/netMet). Ensuite, chaque couple @src-@réseau_dst est inspecté afin de savoir si la machine source a scanné (a dépassé le seuil toléré pour) le réseau destination. Si c'est le cas, on stocke les informations disponibles dans la structure de données présentée Figure -5-, sinon le comportement est considéré comme normal et aucune information n'est stockée. Lors de cette première étape, on ne s'intéresse réellement qu'au scan en hauteur, l'autre type de scan étant recherché par la suite.

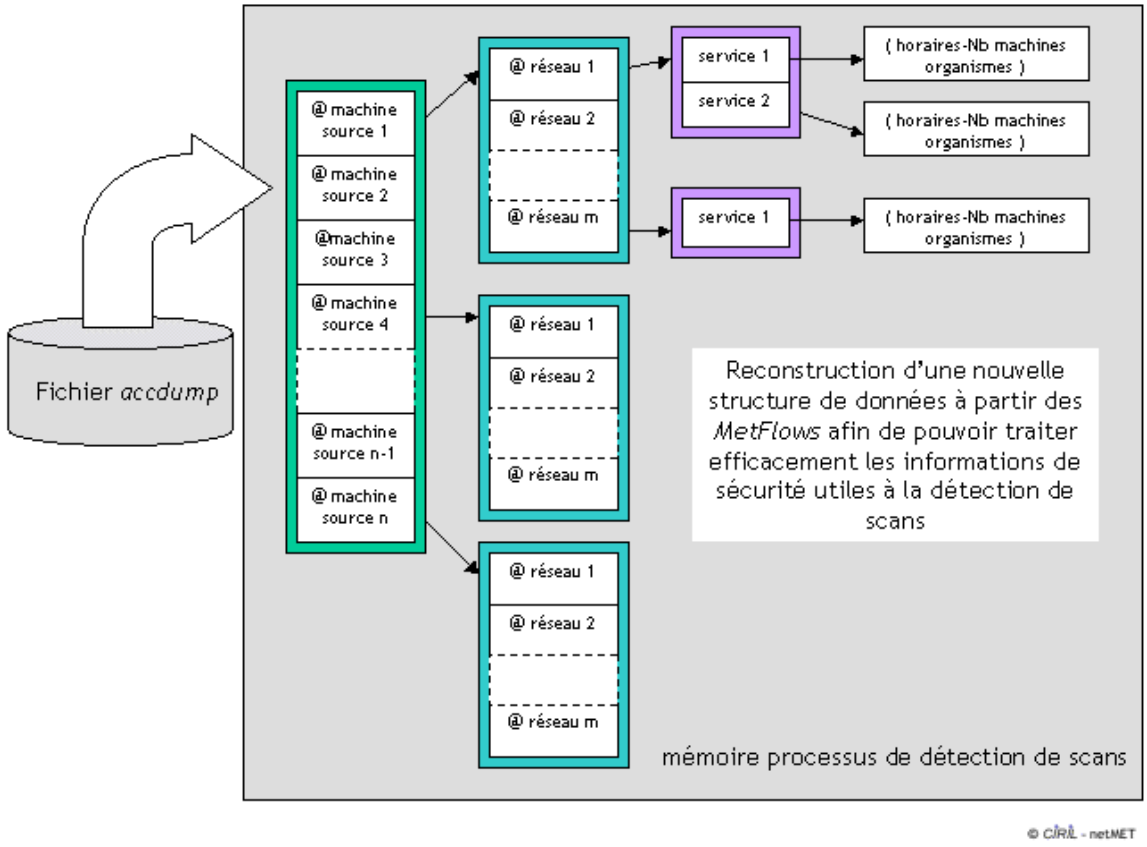


Figure -5- Structure de données utilisée pour la détection de scan en hauteur

La seconde étape consiste à rechercher les horaires des scans et les éventuelles réponses des machines scannées dans les fichiers *dump* produits toutes les 10 minutes. En effet, le fichier *dump* précédemment traité ne permet pas de connaître les horaires des scans car il regroupe les données collectées sur 24 heures sans notion de temps.Pour attribuer des horaires de début et de fin aux différents scans, on identifie des couples @src-@dst tels que @dst appartiennent à un réseau scanné, l'intitulé du fichier *dump* permet alors de déduire l'horaire du scan avec une granularité de 10 minutes (zzaccounting.dmp-**10-20** par exemple).

Pour comptabiliser le nombre de réponse à un scan on recherche des couples @src-@dst tels que, cette fois-ci, @src appartiennent à un réseau scanné. Afin de rester efficace, une nouvelle structure de données est chargée de stocker les réponses des machines d'un réseau scanné. Cette structure sera bien sûr très proche de la structure présentée Figure -5- car les données et les informations manipulées sont pratiquement identiques. D'autre part, dans chaque fichier *dump* de 10 minutes, on recherche les couples @src-@dst ayant demandé plus de n services différents pour pouvoir construire la liste des machines scannées en largeur.

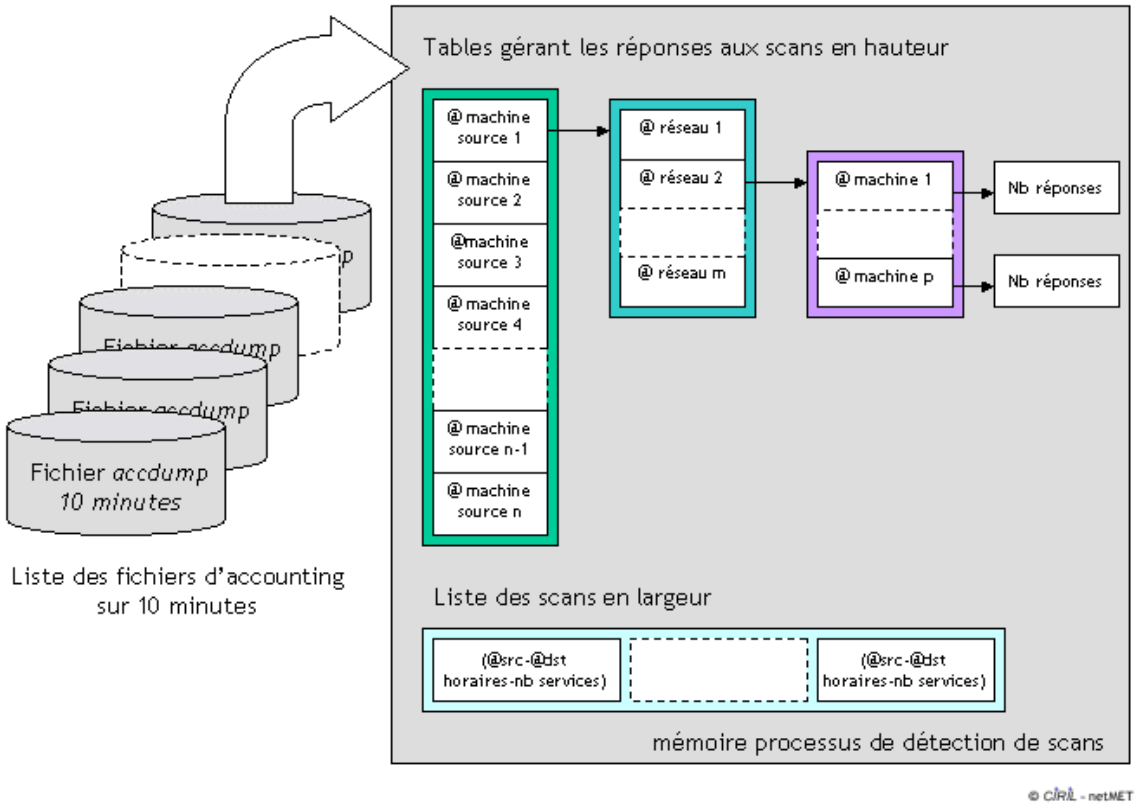


Figure -6- Structures de données utilisées lors du parcours des fichiers d'accounting sur 10 minutes

On peut résumer plus brièvement le fonctionnement du service de détection de scan en considérant que le traitement des données est réalisé par netMETscn et que l'exploitation de ses données est réalisée par un script PERL générant le rapport journalier.

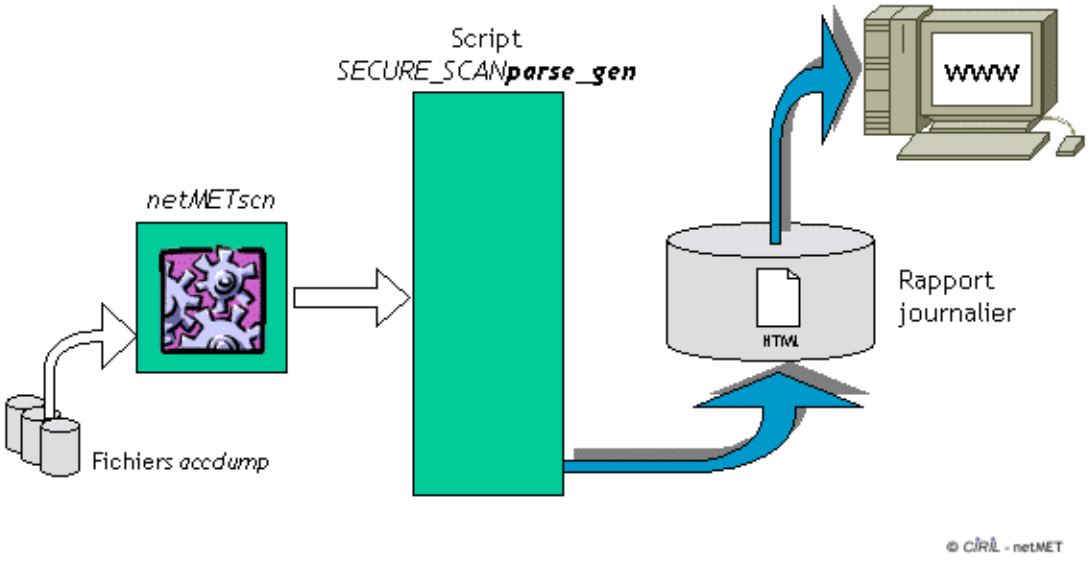
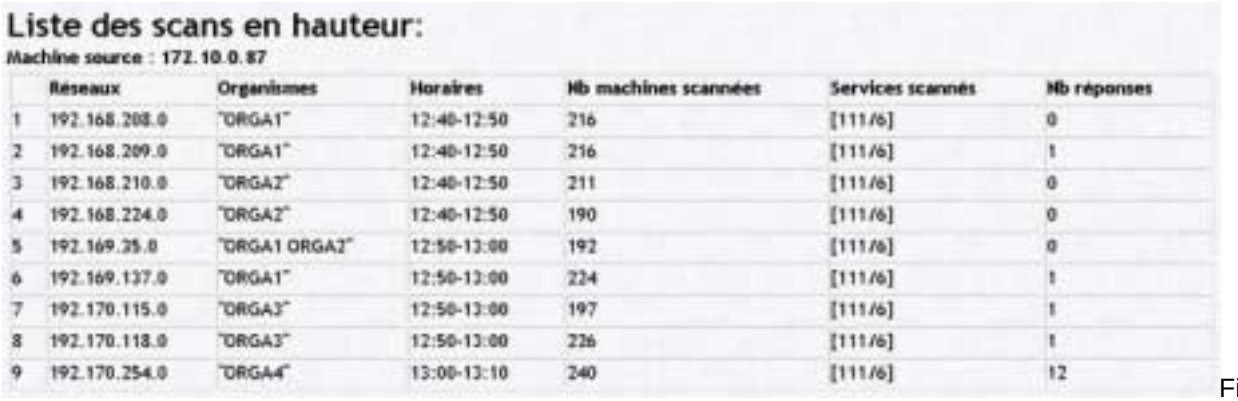


Figure -7- Fonctionnement global du service de détection de scans

Le rapport produit se découpe en 2 parties, la première partie concerne tous les scans en hauteur détectés et la seconde les scans en largeur.

- dans le rapport sur les scans en hauteur est édité, pour chaque machine ayant un comportement suspect, un tableau contenant l'adresse IP des réseaux scannés, les organismes concernés, le nombre de machines scannées, les horaires, les services demandés et le nombre de machines du réseau "ayant répondu au scan".



Liste des scans en hauteur:

Machine source : 172.10.0.87

	Réseaux	Organismes	Horaires	Nb machines scannées	Services scannés	Nb réponses
1	192.168.208.0	"ORGA1"	12:40-12:50	216	[111/6]	0
2	192.168.209.0	"ORGA1"	12:40-12:50	216	[111/6]	1
3	192.168.210.0	"ORGA2"	12:40-12:50	211	[111/6]	0
4	192.168.224.0	"ORGA2"	12:40-12:50	190	[111/6]	0
5	192.169.35.0	"ORGA1 ORGA2"	12:50-13:00	192	[111/6]	0
6	192.169.137.0	"ORGA1"	12:50-13:00	224	[111/6]	1
7	192.170.115.0	"ORGA3"	12:50-13:00	197	[111/6]	1
8	192.170.118.0	"ORGA3"	12:50-13:00	226	[111/6]	1
9	192.170.254.0	"ORGA4"	13:00-13:10	240	[111/6]	12

Figure -8- Exemple de rapport sur les scans en hauteur

Dans l'exemple donné, il apparaît nettement que la machine 172.10.0.87 a scanné 9 réseaux de classe C entre 12h40 et 13h10 sur le service 111/6. Le réseau 192.169.35.0 (ligne 5) appartenant à plusieurs organismes a été scanné sur 192 machines et aucune d'entre-elles n'a répondu à la machine source du scan.

- le rapport sur les scans en largeur est constitué d'un tableau. Chaque ligne du tableau contient la machine source du scan supposé, la machine cible, le nombre de services scannés et l'horaire du scan.

Liste des scans en largeur:

	Machine source	Machine destination	Horaires	Nb services
1	toto.orga1.fr (192.168.200.254)	61.103.90.22	10:00	365
2	61.103.90.22	toto.orga1.fr (192.168.200.254)	10:00	347
3	tata.orga2.fr (192.168.144.16)	suspect1.mal.intentionne.fr (122.203.45.21)	12:30	312
4	suspect2.mal.intentionne.fr (122.213.85.150)	www.machine1.fr (192.168.156.53)	13:40	441
5	www.machine1.fr (192.168.156.53)	suspect2.mal.intentionne.fr (122.213.85.150)	13:40	441
6	192.168.51.233	202.58.198.58	14:50	6521

Figure -9- Exemple de rapport sur les scans en largeur

Le rapport permet facilement de voir que la machine 192.168.51.223 a scanné la machine 202.58.198.58 sur 6521 services vers 14h50 (ligne 6).

Il est facile de "recouper" les différentes informations obtenues dans les deux rapports. En effet, un jeu de couleur permet de repérer facilement dans la liste les scans en largeur, les machines étant déjà concernées par un scan en hauteur. Lorsque la victime d'un scan en largeur est la source d'un scan en hauteur, la ligne lui correspondant apparaît en vert. Lorsqu'au contraire, la source d'un scan en largeur est également la source d'un scan en hauteur, les informations la concernant apparaissent en rouge.

Interprétation des résultats

Les renseignements fournis dans le rapport sont des données "brutes" obtenus par métrologie. Les informations de métrologie n'étant pas forcément adaptées à la détection d'intrusion/de scan, ces renseignements sont plus à considérer comme des alertes sur des comportement étranges et suspects que comme systématiquement des tentatives d'intrusion. Ainsi les gros serveurs peuvent se retrouver comme machine source de scan vers l'Internet alors que c'est un fonctionnement normal de la machine et pas une tentative de piratage.

Dans l'exemple donné figure -9-, la machine toto.orga1.fr (un serveur de orga1) scanne et est scannée par la machine 61.103.90.22. Cette situation n'est en réalité pas une tentative d'intrusion mais le fonctionnement normal du serveur celui-ci répondant aux demandes de son client. De même, pour les machines suspect2.mal.intentionne.fr et www.machine1.fr, le serveur www.machine1.fr ne fait que répondre au client 122.213.85.150. Dans le rapport de scans en largeur que nous analysons ici, seule la dernière ligne est vraiment suspecte et peut-être considérer comme un scan.

D'autre part, la colonne "Nb de réponse" dans le rapport sur les scans en hauteur représente le nombre de machines différentes du réseau scanné ayant répondu à la machine source de scan potentiel, c'est-à-dire ayant renvoyer un flow vers la machine source. Il faut noter que la comptabilisation des réponses ne se base que sur la recherche du couple @src-@dst et sans prendre en compte le service: c'est à dire qu'on ne peut pas être sûr que la réponse correspond bien à une réponse du service sollicité. Cette colonne ne représente en aucun cas le nombre de machines éventuellement piratées (avec intrusion). ainsi un serveur scanné répondant que le service demandé est inaccessible sera comptabilisée dans les réponses. Les figures -10- et -11- illustrent ce phénomène et ses répercutions sur les tables d'accounting.

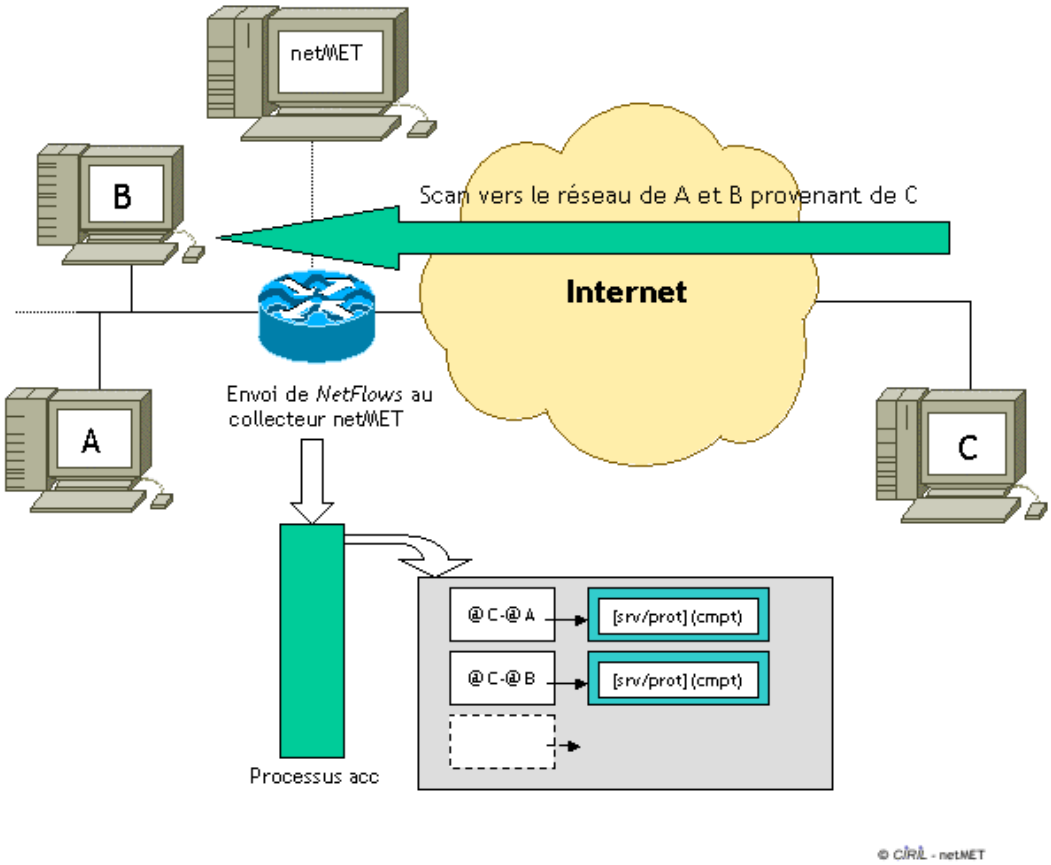


Figure -10- Table d'accounting obtenue après un scan en hauteur

Dans cette exemple simpliste, la machine C scanne un réseau de classe C dont nous n'avons représenté que deux machines A et B. Le routeur envoie tout les NetFlows à la machine netMet connectée sur une de ses interfaces, il transmet donc les flows correspondant au scan provenant de la machine C. Les processus d'accounting dédiés à la sécurité les comptabilisent dans leurs tables, le scan est alors détectable et toutes les informations importantes sont

disponibles dans les différentes tables d'accounting. La seconde étape de la détection consiste simplement à compter les communications en provenance du réseau scanné à destination de la machine C.

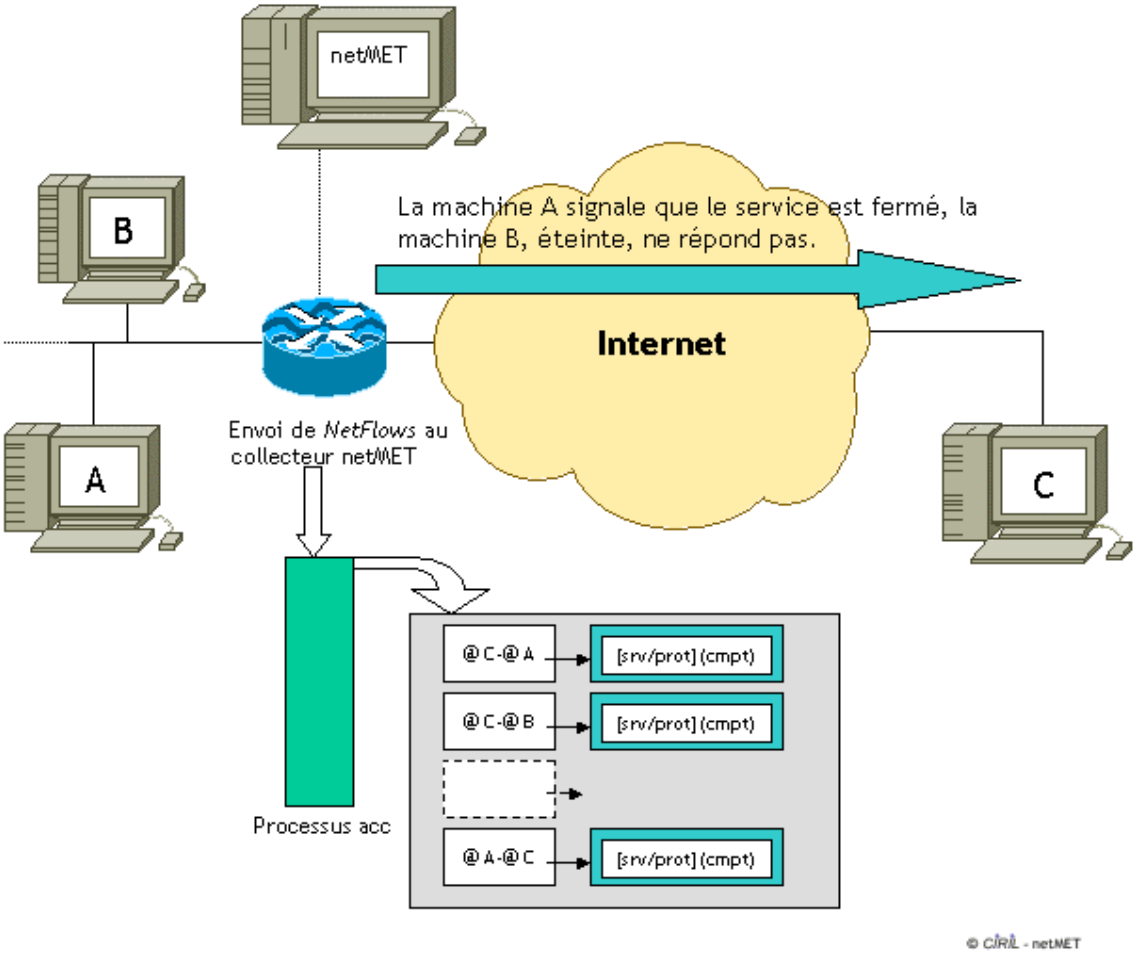


Figure -11- Table d'accounting obtenue après l'émission des réponses au scan

Dans notre exemple, nous supposerons que seul la machine A a répondu au scan en provenance de C. La serveur A signale a son client que le service demandé n'existe pas ou ne lui est pas accessible (ftp par exemple), le routeur transmet alors un flow aux processus d'accounting qui le comptabilisent. Une fois que la machine C a reçu la réponse de A, le "pirate" a des informations supplémentaires sur le réseau-cible. Ainsi, celui-ci sait maintenant qu'il existe bien une machine à l'adresse IP xxx.xxx.xxx.xxx et que cette machine fournie tels ou tels services. Il est donc important de connaître la quantité de machines d'un réseau répondant à un scan pour pouvoir évaluer la sécurité et "l'imperméabilité" de celui-ci. Du point de vue de netMET, la recherche de ces renseignements est très simple car il suffit de parcourir les tables d'accounting en recherchant des couples @rsx-@C ainsi, dans notre exemple, on détecte facilement qu'une seule machine a répondu au scan.

Ainsi; sur l'exemple de la figure -8-, il est clair que 12 machines du réseau 192.170.254.0 n'ont pas été "piratées", mais elles ont simplement répondu négativement à la demande de service de la machine pirate. La connaissance de ce genre d'information est intéressante car elle permet d'évaluer le nombre de machines d'un réseau accessibles depuis l'Internet.

Remarques

- Il est possible de voir apparaître à la place des horaires, dans le rapport de scans, la mention "indef-indef". Ceci vient du fait que la concaténation (au sens netMET) des 144 fichiers *dump* sur 10 minutes ne correspond pas tout à fait au fichier *dump* de 24 heures. En effet, le processus *netMETacc* dédié à la sécurité sur 10 minutes produit régulièrement un fichier *dump* et se re-exécute. Durant ce court laps de temps, un certain nombre de flows ne sont pas pris en compte et par conséquent n'apparaissent pas dans les fichiers *dump* produits (ces flows ne sont pas définitivement perdu car comptabilisés par le processus *netMETacc* réalisant l'accounting sur 24 heures). Ainsi, dans le cadre de la détection de scans, il se peut que l'on détecte un scan en hauteur (recherche dans la table d'accounting correspondant à 24 heures) et que l'on ne trouve pas l'horaire correspondant car les flows en question n'ont pas été pris en compte dans les tables d'accounting des tranches de 10 minutes. Dans ce cas, l'horaire n'est pas attribuée d'où la mention "indef-indef" dans la colonne correspondante.

- Il est également possible que pour un réseau scanné donné, il n'existe pas d'organisme correspondant. Il y a deux raisons expliquant ce phénomène, soit le fichier netMet/etc/organism.def n'est pas à jour et que le réseau scanné n'a pas été attribué dans ce fichier, soit le réseau scanné est un réseau extérieur (côté Internet) et il est dans ce cas normal de ne pas connaître l'organisme correspondant.

- Comme il a été dit précédemment, la recherche du nombre de machines d'un réseau ayant répondu à un scan ne se fait que sur la recherche de couples @src-@dst particuliers. De ce fait, il n'est pas possible, pour une série de scans en hauteur de même source, de même destination mais de ports différents, de connaitre le nombre de réponses obtenues pour chaque scan distinctement.

- Les informations fournies sur les différents scans dans le rapport sont des informations que nous considérons comme minimum et essentielles, mais nous n'exploitons pas complétement les données collectées et gérées dans les différentes tables des Figures -5- et -6-. En effet, les structures utilisées permettent en particulier de compter le nombre de réponses données à un scan machine par machine. Nous avons choisi de ne pas publier ces renseignements dans le compte-rendu journalier pour des raisons de lisibilité mais il est possible d'adapter le script de génération du rapport soi-même. Enfin, nous nous sommes limités à la détection de scans sur une journée mais il est par ailleurs facile d'utiliser les renseignement fournis afin d'alimenter une base de données permettant de détecter des scans sur plusieurs jours ou bien de construire une liste des pirates les plus réguliers.

- Aux vues des quantités d'informations manipulées (24 heures de collecte), les temps de création des différentes structures de données peuvent être conséquents en particulier lors de scans massifs importants. Cependant, nous pensons que ce temps reste très raisonnable par rapport à la quantité de renseignements fournis par le service de détection de scans et au nombre de flows inspectés.

Toutes remarques ou questions concernant cette documentation ou le service de détection de scans sont les bienvenues.

Documentation *netMET*		version 1.1 (1.1_2.4beta)	
par	:	Cyril PROCH Alexandre SIMON (maj)	Cyril.Proch@ciril.fr
crée le	:	2001-07-20	
Mise à jour le	:	2001-11-12 17:46	