

# HUITIEME COLLOQUE SUR LE TRAITEMENT DU SIGNAL ET SES APPLICATIONS

NICE du 1<sup>er</sup> au 5 JUIN 1981

PROBLEMES DE DEBORDEMENTS DE DECODEURS SEQUENTIELS A PILE

DAVID HACCOUN

Ecole Polytechnique de Montréal C.P. 6079 A, Montréal Canada H3C 3A7

## RESUME

Dans les canaux de communication sans mémoire, les systèmes utilisant le codage convolutionnel avec décodage probabiliste sont parmi les plus intéressants tant du point de vue de leur performance d'erreur que du point de vue de leur réalisation et implantation matérielle. Les deux principales techniques de décodage probabiliste des codes convolutionnels sont le décodage séquentiel et le décodage de Viterbi.

L'effort de décodage en nombre de calculs effectués par bit décodé est constant, mais en général élevé pour le décodeur de Viterbi, alors qu'en décodage séquentiel cet effort est en moyenne très faible mais aussi très variable, avec une distribution de type Pareto. Cette variabilité de calcul est l'inconvénient principal du décodeur séquentiel et nécessite l'emploi d'un tampon d'entrée afin d'y stocker les données reçues et qui attendent d'être décodées. La variabilité de l'effort de calcul a une influence directe sur la file d'attente au tampon d'entrée dont le débordement constitue un événement d'erreur catastrophique.

Basés sur l'algorithme de Zigangirov-Jelinek, plusieurs variantes d'algorithmes de décodage séquentiel qui réduisent cette variabilité de calcul sans dégrader les performances d'erreur sont présentées. Ces algorithmes utilisent une pile et un tampon d'entrée, et explorent simultanément plusieurs chemins dans l'arbre. L'impact de la réduction de la variabilité de calcul sur le comportement de la file d'attente du tampon d'entrée et sur l'utilisation de la pile est examiné.

Les résultats de simulation montrent que pour tous les algorithmes la taille de la pile nécessaire au décodage suit une distribution de type Pareto; en conséquence, pour une probabilité d'effacement acceptable ( $10^{-2}$  -  $10^{-3}$ ) il apparaît préférable de faire déborder la pile avant le tampon. Les résultats obtenus avec des codes de longueur de contrainte  $K=24$  montrent qu'en diminuant la taille de la pile du décodeur, la taille du tampon d'entrée peut alors être considérablement réduite sans augmentation appréciable de la probabilité d'effacement (blocs non complètement décodés).

Enfin, une procédure de retransmission utilisant un décodeur séquentiel ayant une pile de petite taille est présentée. Dans cette technique les blocs difficiles à décoder provoquent le débordement de la pile et sont retransmis évitant ainsi de congestionner le tampon d'entrée. Des résultats de simulation montrant les compromis entre les tailles de la pile et du tampon, et la vitesse de décodage sont analysés et comparés à ceux du décodage séquentiel habituel.

## SUMMARY

In discrete memoryless channels, systems using convolutional coding with probabilistic decoding are the most promising from both an error performance and a practical implementation point of view. The two main probabilistic decoding techniques for convolutional codes are Viterbi decoding and Sequential decoding. The decoding effort in number of computations per decoded bit is constant but is large in general for Viterbi decoding, whereas for sequential decoding it is typically very small but highly variable, with a Pareto distribution. This computational variability is the principal drawback of sequential decoding and requires the use of an input buffer in order to store the received data waiting to be decoded. The computational variability has a direct influence on the waiting line behavior in the input buffer whose overflow may be a catastrophic communications breakdown.

Based upon the Zigangirov-Jelinek stack algorithm, several variants of sequential decoding are presented. These variants also use a stack and an input buffer and explore several paths simultaneously. The ensuing computational variability is shown to be reduced and the impact of this variability reduction on the stack and input buffer is examined.

Simulation results show that necessary the stack size is Pareto distributed and it appears preferable to have overflows occurring in the stack rather than in the input buffer. It is shown that under certain conditions a buffer of size two blocks will never overflow.

Finally containing the overflows in the stack, a retransmission procedure for the overflowed blocks is presented. Simulation results about effective rates, error probability, buffer and stack sizes, speed factors, etc. will be compared to those of the usual sequential decoder.

---

Cette recherche a été subventionnée en partie par le conseil des Recherches en Sciences Naturelles et en Génie, Canada N° A9336; et par le programme Formation de Chercheurs et Action Concertée, Québec, N° FCAC EQ-210.



## I. INTRODUCTION

L'usage de plus en plus répandu de techniques de transmissions numériques dans les télécommunications terrestres et par satellite conduit à l'utilisation grandissante de procédures de correction d'erreur par codage de canal qui sont puissantes, fiables et pratiques. Aussi un problème important consiste à développer des techniques de codage et décodage délivrant de faibles probabilités d'erreur avec des décodeurs de complexité acceptable. Dans les canaux de communication sans mémoire, les systèmes utilisant le codage convolutionnel avec décodage probabiliste sont parmi les plus intéressants tant du point de vue de leur performance d'erreur que du point de vue de leur réalisation et implantation matérielle. Le décodage probabiliste comprend un ensemble de techniques où le message décodé est obtenu par des procédures probabilistes plutôt que par des opérations algébriques fixes, et où les codes utilisés n'ont pas, en principe, à satisfaire à une structure algébrique particulière comme pour les codes en blocs. Les deux principales techniques de décodage probabiliste des codes convolutionnels sont le décodage séquentiel [1] et le décodage de Viterbi [2]. Chacune de ces techniques consiste à trouver un chemin particulier (le message transmis) dans un graphe orienté où on assigne aux branches des "métriques" ou valeurs de vraisemblance entre les données reçues du canal de transmission et les données qui auraient pu être transmises. L'objectif général du décodeur est donc de déterminer le chemin ayant la métrique totale maximum, et ce, avec un minimum d'effort et un maximum de fiabilité. Ce chemin trouvé par le décodeur est la séquence décodée. Les techniques de décodage séquentiel et de décodage de Viterbi sont nettement différentes et trouvent des domaines d'application distincts.

Le décodage de Viterbi avec quantification douce du canal peut fournir un gain de codage supérieur à 5dB à une probabilité d'erreur de  $10^{-5}$ , alors que le décodage séquentiel avec quantification douce permet d'atteindre des gains de codage de quelque 7 à 8 dB. L'applicabilité de ces techniques de décodage est particulièrement intéressante dans les liaisons par satellite où chaque décibel d'énergie à fournir est extrêmement coûteux [3].

Après avoir introduit quelques algorithmes de décodage séquentiel à pile qui permettent de diminuer la variabilité de l'effort de calcul, on étudiera le comportement de leur pile et leur impact sur la dynamique de la file d'attente au tampon d'entrée. Enfin on présentera une procédure d'utilisation de décodeurs séquentiels à pile qui tendent à réduire au minimum la taille du tampon d'entrée.

On supposera connues les structures en arbre et en treillis des codes convolutionnels. Considérant seulement des codes convolutionnels binaires de taux de codage  $R=1/V$  et de longueur de contrainte  $K$  finie, à chaque bit d'information il y correspond 2 branches d'un arbre portant chacune  $V$  symboles codés. L'extrémité de chaque branche est un noeud caractérisé par un état du codeur, et le nombre d'états distincts est égal à  $2^{K-1}$ . Un chemin dans l'arbre est spécifié par la séquence d'information qui est entrée dans le codeur et deux chemins reconvergent si leurs  $(K-1)$  derniers bits d'information sont identiques. L'arbre d'encodage contient donc une énorme redondance qui peut être éliminée en ne gardant qu'un seul chemin au-delà de chaque noeud de reconvergence. L'arbre devient alors un treillis ayant  $2^{K-1}$  états et pour une séquence d'information de longueur  $L$  bits, les chemins dans l'arbre ou le treillis ont donc une longueur maximale égale à  $L$  branches.

## 2. DECODAGE SEQUENTIEL

Un décodeur séquentiel utilise la structure en arbre du code et n'explore un chemin à la fois, que la

partie de l'arbre qui paraît être la plus vraisemblable sans explorer l'arbre entier. C'est donc une procédure sous-optimale. Dans un canal sans mémoire, la fonction de vraisemblance utilisée, appelée aussi "métrique", est cumulative le long des branches d'un même chemin. Elle tend à croître le long du chemin correct et tend à décroître le long de tous les chemins incorrects. L'objectif du décodeur est donc d'explorer l'arbre d'encodage le long du chemin ayant la métrique la plus élevée parmi tous les chemins explorés. Par contre un décodeur de Viterbi utilise la structure en treillis du code et examine tous les chemins qui auraient pu être transmis. En conséquence, l'effort de décodage en nombre de calculs effectués par bit décodé est constant, mais en général élevé pour le décodeur de Viterbi, alors qu'en décodage séquentiel cet effort est en moyenne très faible mais aussi très variable, avec une distribution de type Pareto, c'est-à-dire:

$$P(C \geq N) \approx k N^{-a}, N \gg 1 \quad (1)$$

où  $k$  et  $a$  sont des constantes qui ne dépendent que du taux de codage et du canal. L'expression (1) indique que la distribution du nombre de calculs suit une décroissance algébrique et non pas exponentielle. La Figure 1 montre quelques distributions obtenues par simulation à l'ordinateur de décodage séquentiel.

Cette variabilité de calcul est l'inconvénient principal du décodage séquentiel alors que le grand nombre de calculs est l'inconvénient du décodeur de Viterbi. La variabilité de l'effort de calcul nécessite l'utilisation d'un tampon à l'entrée du décodeur afin d'y stocker les branches reçues en attente d'être décodées. Le débordement de ce tampon constitue un événement d'erreur catastrophique entraînant un grand nombre de bits en erreur. Il est donc très important de réduire la probabilité de débordement de ce tampon.

L'algorithme de décodage séquentiel dont il est question dans cet article est l'algorithme à pile de ZIGANGIROV et JELINEK (Z-J) [1]. Cet algorithme utilise une pile pour stocker toutes les caractéristiques des chemins explorés, et un tampon d'entrée pour stocker les séquences reçues en attente d'être décodées. Un schéma de principe du décodeur est montré à la Figure 2. La pile est une liste ordonnée où sont stockés les chemins explorés par ordre décroissant de leur métrique. Le sommet de la pile contient le chemin ayant la métrique maximum courante; ce chemin est donc celui qui sera prolongé. L'algorithme a donc pour objet de déterminer à chaque étape le sommet de la pile et d'en faire le prolongement. Il se compose des 3 étapes suivantes:

- 1) Calcul des métriques des deux chemins issus du sommet de la pile et insertion dans la pile de ces deux chemins.
- 2) Elimination du sommet qui vient d'être prolongé.
- 3) Détermination du nouveau sommet. Si c'est le noeud terminal, stop. Sinon retour à 1).

Lorsque l'algorithme arrête, le sommet de la pile est le noeud terminal du chemin décodé, qui est alors facilement récupéré.

Bien que très simple, cet algorithme n'est pas pratique car le temps nécessaire à la mise en ordre exacte de la pile est beaucoup trop élevé. Cette difficulté est contournée en effectuant une mise en ordre approximative: les noeuds explorés sont insérés aléatoirement dans des sous-piles de la façon suivante:

Un noeud  $U$  de métrique  $\Gamma(U)$  est inséré au hasard dans la sous-pile  $Q$  si

$$QH < \Gamma(U) \leq (Q+1)H \quad (2)$$

où  $H$  est une valeur arbitraire.

Avec cette modification la recherche du sommet de

la pile est réduite à la recherche de la sous-pile maximum non vide, et le chemin prolongé est choisi au hasard dans cette sous-pile.

### 3. ALGORITHMES A PILE GENERALISES

Dans le but d'unifier les techniques de décodage séquentiel et de Viterbi, une classe d'algorithmes généralisés à pile a été développée [4,5]. Dans cette classe d'algorithmes les aspects désirables des deux techniques de décodage sont systématiquement exploités, et chacune de ces techniques ne devient qu'un cas particulier. La variabilité et l'effort de décodage peuvent être réduits sans perte de performance d'erreur; ils peuvent même s'adapter automatiquement aux besoins courants imposés par le bruit dans le canal. Les cas particuliers d'algorithmes généralisés qui sont considérés ici sont:

#### i) Algorithme M-chemins

Faire l'extension simultanée des M-chemins les plus vraisemblables.

#### ii) Algorithme N sous-piles

Faire l'extension simultanée de tous les chemins contenus dans les N sous-piles maximum. Ce nombre de chemins est variable et a tendance à s'adapter au niveau de bruit dans le canal.

#### iii) Algorithme Adaptatif

Observer la croissance de la métrique maximum. Si la métrique croît prolonger l chemin. Si elle chute d'une quantité D, prolonger simultanément un nombre M(D) de chemins, où M(D) est une fonction non décroissante de D.

#### iv) Algorithme Prédicteur

La chute D de la métrique maximum est déterminée à l'avance en pénétrant l'arbre le long du chemin le plus vraisemblable. Utilise la valeur D pour faire varier N ou M(D) des algorithmes ii) ou iii). Dans toutes ces variantes, la réduction de la variabilité de calcul est obtenue au prix d'un effort moyen plus grand. Il s'agit donc d'évaluer l'impact de ces algorithmes sur les comportements de la pile et de la file d'attente du décodeur. Etant donné la difficulté d'analyse, cette évaluation fut réalisée par simulation à l'ordinateur.

## 4. FONCTIONS DE REPARTITION DES CONTENUS DE LA PILE ET DU TAMPON

### 4.1 Taille de la Pile

L'information étant transmise en blocs de 500 bits, le contenu de la pile est observé à la fin du décodage de chaque bloc. Les paramètres de la simulation sont: Taux de codage  $R=0.5$

Longueur de contrainte  $K=24$

$E_b/N_0 = 4.64$  dB ( $R/R_{comp} = 0.99$ ) et

$E_b/N_0 = 5.60$  dB ( $R/R_{comp} = 0.85$ ).

Les résultats obtenus pour  $R/R_{comp} = 0.99$  sont donnés au Tableau 1 et une fonction de répartition typique est montrée à la Figure 3. Cette figure indique une distribution Pareto pour le contenu de la pile. Cette distribution est générale à tous les algorithmes étudiés.

Des simulations intensives ont montré que les algorithmes multichemins présentent tous un exposant Pareto supérieur à celui de (Z-J), avec cependant un nombre de calculs  $\bar{C}$  supérieur. A la lumière de ces résultats il apparaît que pour des piles de taille relativement modeste et une probabilité de débordement  $P(s \geq G)$  de l'ordre de  $10^{-2}$  à  $10^{-3}$  il est préférable de choisir un algorithme tel le (Z-J) dont le nombre

Tableau 1: Contenu de la Pile

$R/R_{comp}=0.99$ ; 200 blocs; Taille= $10^4$

Paramètre \ ALG.	Z-J	M=2	N=5	N=7	M(D)= 1+√2D
# Erreurs	0	0	0	0	0
# Débordements	2	2	2	2	2
Exp. Pareto	1.37	1.80	1.80	2.63	2.10
$\bar{C}$	1.67	2.45	2.19	2.88	2.20
p Taille moyenne	1698	2570	2291	2951	2291
I $P(s \geq G) = 10^{-1}$	2344	3020	2951	3981	2884
L $P(s \geq G) = 10^{-2}$	10,000	10,000	8,912	10,000	8,912
E $P(s \geq G) = 10^{-3}$	53,703	35,481	31,623	23,988	26,915

moyen de calculs est relativement faible. Par contre lorsque  $R/R_{comp}$  tend vers 1, la préférence devrait aller à l'un des algorithmes multichemin proposés.

### 4.2 Tampon d'entrée

Les débordements de la pile affectent la queue de la fonction de répartition et constituent donc des événements relativement peu probables. Cependant ces événements peuvent entraîner un accroissement considérable de la file d'attente au tampon d'entrée du décodeur, voire même provoquer un débordement de ce tampon. La dynamique de la file d'attente est donc dépendante de celle de la pile. Un raisonnement simple permet de voir que sous certaines conditions un tampon de taille maximum 2 longueurs de bloc ne débordera jamais. En effet soit U le gain de vitesse du décodeur, S la taille de la pile et L la longueur des blocs. Si le premier bit d'un bloc fait déborder la pile, alors le nombre de branches qui doivent être stockées dans le tampon d'entrée est égal à  $S/2U$ . On peut alors montrer [G] que si  $S/2U > L$  la file d'attente dans le tampon peut augmenter indéfiniment, alors que si  $S/2U < L$  cette file d'attente ne peut être supérieure à 2L. Sous cette condition, naturellement la distribution de la file d'attente ne sera pas de type asymptotiquement Pareto. Tel que montré à la Figure 4, la queue de la distribution tend à décroître verticalement. Des résultats de simulation avec  $L=523$ ,  $S=10,000$  et  $U=10$  et  $U=3$  sont donnés au tableau 2 pour illustrer deux cas possibles.

Pour la réalisation matérielle de décodeurs séquentiels un tampon ne contenant que 2 blocs est une solution fort intéressante. Les résultats obtenus montrent qu'on peut donc échanger la taille du tampon et le gain de vitesse pour atteindre un compromis adéquat pour une probabilité de débordement donnée.

## 5. PROCEDURE DE RETRANSMISSION

Dans certaines liaisons telles les communications entre ordinateurs, une probabilité d'erreur inférieure à  $10^{-10}$  est requise. Dans ces conditions même le décodage séquentiel s'avère insuffisant et une procédure de retransmission devient nécessaire. Donc, en principe les blocs provoquant un débordement de la pile ou du tampon devraient être retransmis. Le choix d'un tampon de grande taille entraîne nécessairement l'uti-



Tableau 2: Comportement de la file d'attente

U	ALG.	Z-J	M=2	N=2	N=5	M(D)=1+1/2D
	Paramètre					
3	Moyenne	87	170	134	488	138
	Pr(W>D)=10 <sup>-2</sup>	1122	1122	1259	1259	1259
	Pr(W>D)=10 <sup>-3</sup>	1413	1259	1413	1585	1413
10	Moyenne	5	5	4	4	4
	Pr(W>D)=10 <sup>-2</sup>	126	126	126	160	125
	Pr(W>D)=10 <sup>-3</sup>	355	355	356	282	356

lisation d'une grande pile, alors qu'un débordement de la pile met un terme à l'accumulation des branches dans le tampon et force le décodeur à décoder un nouveau bloc. Si le bruit accompagnant ce nouveau bloc et les blocs suivants est normal, alors ces blocs seront facilement décodés, éliminant ainsi le retard accumulé. Il apparaît donc avantageux de confiner les débordements dans la pile plutôt que dans le tampon. De plus comme la distribution du contenu de la pile est de type Pareto, un grand accroissement de la taille de la pile n'a que peu d'effets sur sa probabilité de débordement. Par conséquent il s'avère avantageux d'utiliser une pile de taille modeste afin d'y détecter très tôt les blocs difficiles à décoder, et d'en demander leur retransmission. Ces blocs étant expurgés assez tôt, le tampon d'entrée devient peu susceptible à des débordements et donc sa taille peut être réduite sans grande conséquence. Comme il a été montré plus haut, en choisissant le gain de vitesse  $U$  tel que  $U > (S/2L) \geq \bar{C}$ , un tampon de taille  $(2L)$  branches ne débordera jamais. La taille minimum de la pile  $S$  pourra donc être déterminée en fonction de  $L$ .

Un autre avantage à retransmettre les blocs qui débordent provient du fait que les blocs nécessitant un très grand effort de calcul sont souvent décodés en erreur. En les retransmettant, ces erreurs peuvent donc être évitées. Par conséquent une telle procédure de retransmission permet de réduire au minimum la taille du tampon, et de réduire la probabilité d'erreurs non détectées, au prix d'une faible détérioration du taux de codage effectif. Ainsi le décodeur séquentiel servirait à la correction et à la détection des erreurs.

La technique de retransmission a été simulée sur ordinateur (seuls les blocs causant un débordement sont retransmis) en faisant varier le gain de vitesse  $U$  de 2 à 10 et la taille de la pile  $S$  de 1000 à 10,000 places. Le nombre de blocs à transmettre est égal à 200, avec des valeurs de  $E_b/N_0$  égales à 4.64 dB et 4.89 dB dans le canal direct ( $R/R_{Comp} = 0.99$  et  $0.95$ ). Le canal de retour est considéré <sup>comme</sup> sans bruit. A la fin de la simulation la taille maximum de la file d'attente est observée, ainsi que le nombre de blocs retransmis et le nombre d'erreurs. Il est alors observé que la fonction de repartition de la pile tend encore à être Pareto alors que celle de la file d'attente dans le tampon chute verticalement au-delà d'une certaine valeur. La Figure 5 montre l'échange possible entre le gain de vitesse, la taille de la pile et celle du tampon. Ainsi, on peut voir le peu d'influence de la taille de la pile sur celle du tampon lorsque le gain de vitesse est très élevé. De plus, tel

que prévu on remarque que dans ces conditions un tampon de longueur 1000 branches ne débordera jamais. Le choix final des paramètres à choisir dans une application particulière devient un problème d'ingénierie et dépendra de cette application et des contraintes du problème.

## 6. CONCLUSIONS

Un certain nombre d'algorithmes de décodage séquentiel à pile ont été simulés, et la probabilité de débordement de la pile et du tampon analysée. La fonction de répartition de la pile est toujours de type Pareto, dont l'exposant croît avec le nombre moyen de calculs effectués. Par conséquent le choix de l'algorithme à utiliser dépendra de la taille de la pile disponible. La taille de la pile limitant l'effort de calcul par bloc, un gain de vitesse suffisant peut limiter la taille du tampon à 2 blocs sans débordement. Enfin, en contrôlant les débordements dans la pile et en utilisant une procédure de retransmission, la probabilité de débordement du tampon et la probabilité d'erreur peuvent être sensiblement améliorées au coût d'une faible réduction du taux de codage.

## 7. REFERENCES

- [1] F. Jelinek, "A fast Sequential Decoding Algorithm using a Stack", IBM Journal of Research and Development, Vol. 13, Nov. 1969.
- [2] A.J. Viterbi, "Convolutional Codes and their Performance in Communication Systems", IEEE Trans, on Com. Tech., Vol. COM-19, Oct. 1971.
- [3] V. Bhargava, D. Haccoun, R. Matyas, P. Nuspl, "Digital Communications by Satellite", John Wiley, New York, Oct. 1981.
- [4] D. Haccoun, M. Ferguson, "Generalized Stack Algorithm for Decoding Convolutional Codes", IEEE Trans. on Inf. Theory, Vol. IT.21, Nov. 1975.
- [5] A. Janelle, D. Haccoun, "Décodage adaptatif des codes convolutionnels", Rapport Technique EP78-R-18, Ecole Polytechnique de Montréal, 1978.
- [6] M. Dufour, "Décodage séquentiel: simulation et analyse de la file d'attente et de la pile", Thèse de maîtrise en ingénierie, Ecole Polytechnique de Montréal, Sept. 1980.

PROBLEMES DE DEBORDEMENTS DE DECODEURS SEQUENTIELS A PILE

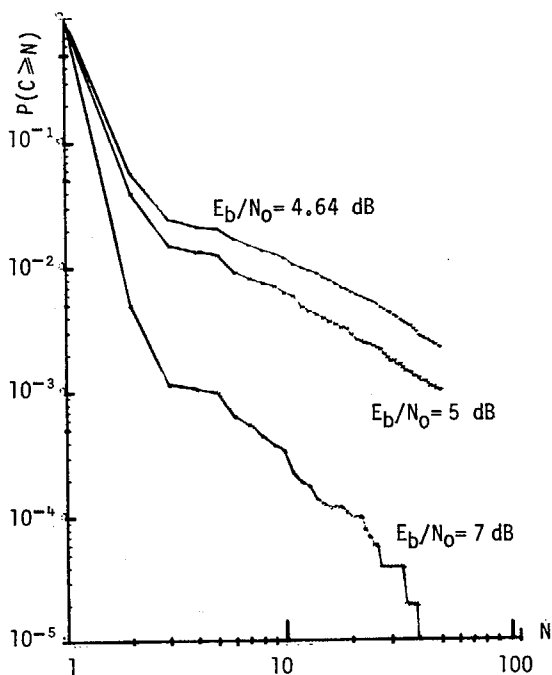


Figure 1: Distribution du nombre de calculs.

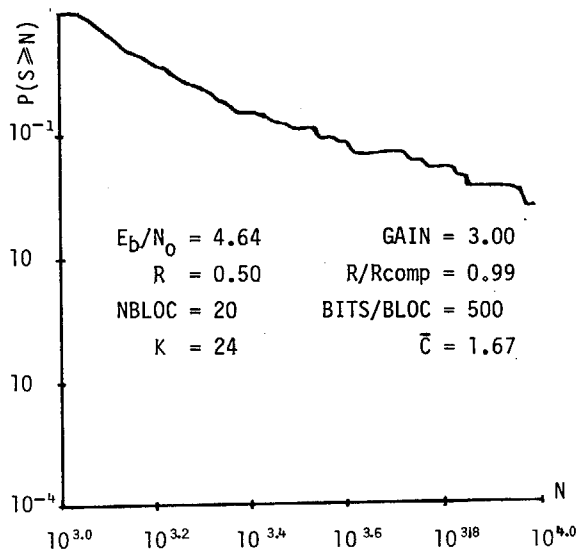


Figure 3: Distribution de la probabilité de débordement de la pile (décodeur Z-J).

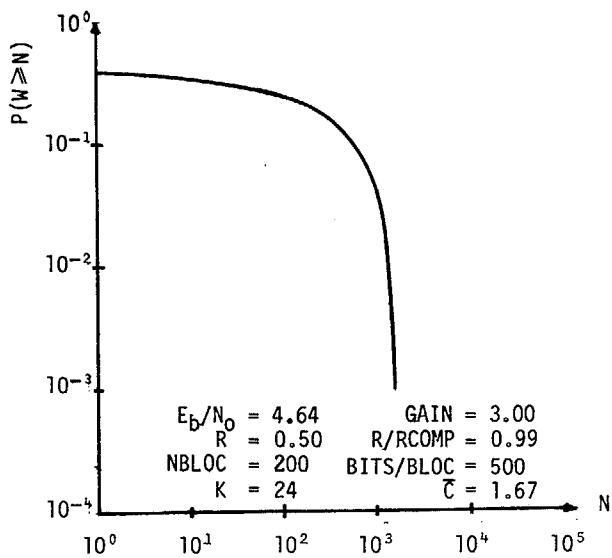


Figure 4: Distribution de la probabilité de débordement du tampon (décodeur Z-J).

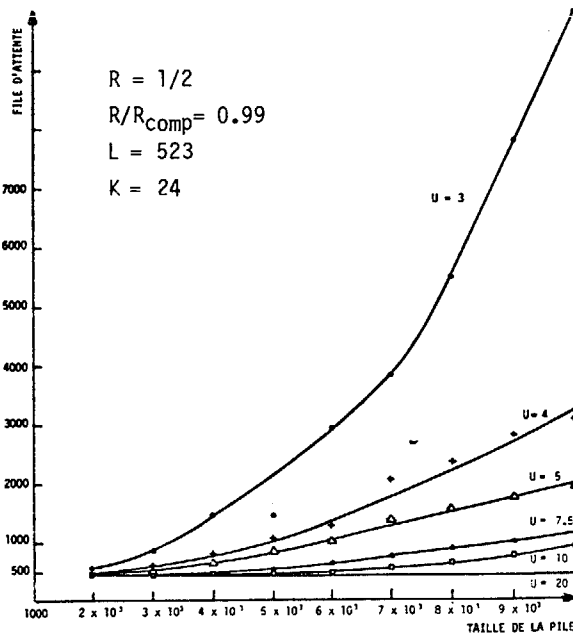


Figure 5: File d'attente en fonction de la taille de la pile pour plusieurs gains de vitesse, décodeur séquentiel avec retransmission.

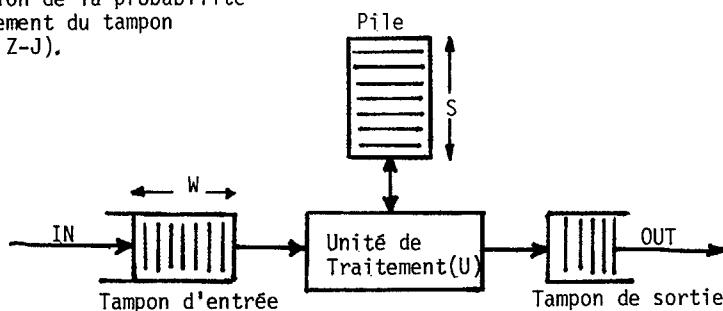


Figure 2: Schema bloc d'un décodeur séquentiel à pile.

