Revue Française de Gestion Industrielle
Vol.19, N°4

LA PERFORMANCE EN COUTS ET EN DELAIS LA THEORIE DES FILES D'ATTENTE ET LE BPR

Jean-Louis Peaucelle *

Résumé. - Le Business Process Reengineering (BPR) est un mouvement dont les principes paraissent qualitatifs, bien que les résultats de la réorganisation soient affichés sous forme quantitative, en amélioration des délais et des coûts. Cet article fournit une interprétation des préceptes qualitatifs du BPR formulés par Hammer en comparant les différentes manières de traiter des flux aléatoires de travail dans les processus administratifs. Cette comparaison est menée de manière discursive, et plus formellement avec les équations liant les délais et les coûts dans chaque solution. Les préceptes qualitatifs avancés par Hammer prennent le sens d'une recherche de la solution organisationnelle la meilleure en délai et en coût. Si on ne tient compte ni des effets de productivité, ni des spécialisations des métiers, ni des différences de salaires, le centre d'appels, à un seul niveau, est l'organisation archétypale la meilleure, en termes de coût et de délais.

Mots clés : Centre d'appels, BPR, Processus, Reengineering, reconfiguration, modèle mathématique, file d'attente, Hammer, Beckmann, délai, coût.

1. Succès et échecs du BPR (Business Process Reengineering)

La réorganisation des activités de service selon les principes de Hammer et de Champy (1993) est à la mode dans les entreprises occidentales dans la décennie 1990. Ces auteurs, et d'autres organisateurs, se targuent de succès considérables et fournissent des références convaincantes. Beaucoup d'entreprises ont suivi le mouvement. Les résultats n'ont pas toujours été à la hauteur des espérances. Les recettes incantatoires ne suffisent pas. La question qu'on va se poser ici concerne les principes du BPR. Pourquoi des améliorations considérables de performance peuvent-elles être obtenues ? Selon quels principes les gains peuvent-ils être aussi

* **Professeur des Universités, IAE de Paris, peaucelle.iae@univ-parisl.fr**

considérables qu'on l'annonce ? Par la mise en commun de flux et la polyvalence des agents, telle est la réponse que nous développons dans cet article. Cette argumentation s'appuie sur la théorie des files d'attente et sur la lecture attentive des principes de Hammer.

Le mouvement du BPR a été lancé par Michael Hammer. Dans un article de 1990, il critique les informatisations traditionnelles qui ont copié les procédures manuelles antérieures notamment pour ne pas remettre en cause la répartition des rôles entre les divers services. Il avance sept nouveaux principes pour organiser les services en faisant «table rase» ("obliterate") de l'organisation passée. Hammer ne les justifie pas. H se contente de donner des exemples d'entreprises où ils sont mis en œuvre avec succès. Cet article a comme objectif de reconstruire un raisonnement correspondant à ces sept préceptes classiques du BPR. Dans ce but, commençons par examiner en détail les diverses solutions pour s'organiser face à une demande aléatoire à laquelle on veut répondre vite. La comparaison entre ces solutions est un outil pour comprendre les principes de Hammer. Les sept principes s'éclairent alors comme des préceptes pour passer d'une organisation à une autre, meilleure.

La méthode qu'on va utiliser est comparative. Une réorganisation tente d'améliorer les performances de l'entreprise. Elle n'a de sens qu'en termes de comparaison entre deux modes d'organisation, l'ancien et le nouveau. Dans chaque entreprise, dans chaque processus, les organisations anciennes sont spécifiques et les organisateurs du BPR mettent en place des solutions sur mesure. Cependant, chaque organisation particulière est une combinaison de solutions archétypales. La comparaison de ces solutions archétypales indique dans quel sens il existe des améliorations. Les sept principes de Hammer s'interprètent comme l'évolution des organisations concrètes vers la solution archétypale qui surclasse toutes les autres.

Décrivons d'abord ces solutions archétypales. Leur description correspond à des activités administratives, des activités de service. On parlera de travail sur des dossiers. Pour chaque solution, on comparera la performance, en termes de coût et de délai, avec une autre organisation. Ensuite, chacun des principes de Hammer sera rapproché des écarts entre les solutions archétypales.

2. Les solutions archétypales de l'organisation du travail administratif

Les modes d'organisation qu'on appelle ici « solutions archétypales » répondent au même problème, caractéristique des activités de service : traiter des cas (dossiers) survenant de manière aléatoire. Pour les traiter, est mise en place une organisation, c'est-à-dire des personnes, une répartition du travail entre elles, une spécialisation éventuelle sur certaines tâches (ou une polyvalence), des règles sur le rythme.

On suppose que le travail arrive de manière aléatoire selon une loi constante au cours du temps. Pour atteindre une situation stable face à cet aléa, toutes ces organisations doivent disposer de capacités de traitement supérieures au flux moyen d'arrivée des dossiers. Ces capacités excédentaires sont utilisées au moment où les arrivées sont nombreuses et sont inoccupées quand les arrivées sont rares. Plus il y a de capacités excédentaires, moins on fait attendre les dossiers au moment où ils arrivent en grand nombre. En conséquence, le délai moyen

est plus faible. Mais ces capacités excédentaires sont inoccupées quand il y a peu de cas qui se présentent. Elles sont présentes, prêtes à faire du travail, payées. Donc, pour aboutir à des délais faibles, le coût est élevé.

La performance d'une solution archétypale est définie ici comme le couple « coût par dossier » et « délai moyen d'achèvement du traitement des dossiers ». Ces deux critères sont privilégiés par tous les auteurs de l'école du BPR. D'autres critères sont possibles, comme la qualité (l'absence d'erreurs) ou la flexibilité. Ils ne sont pas pris en compte ici.

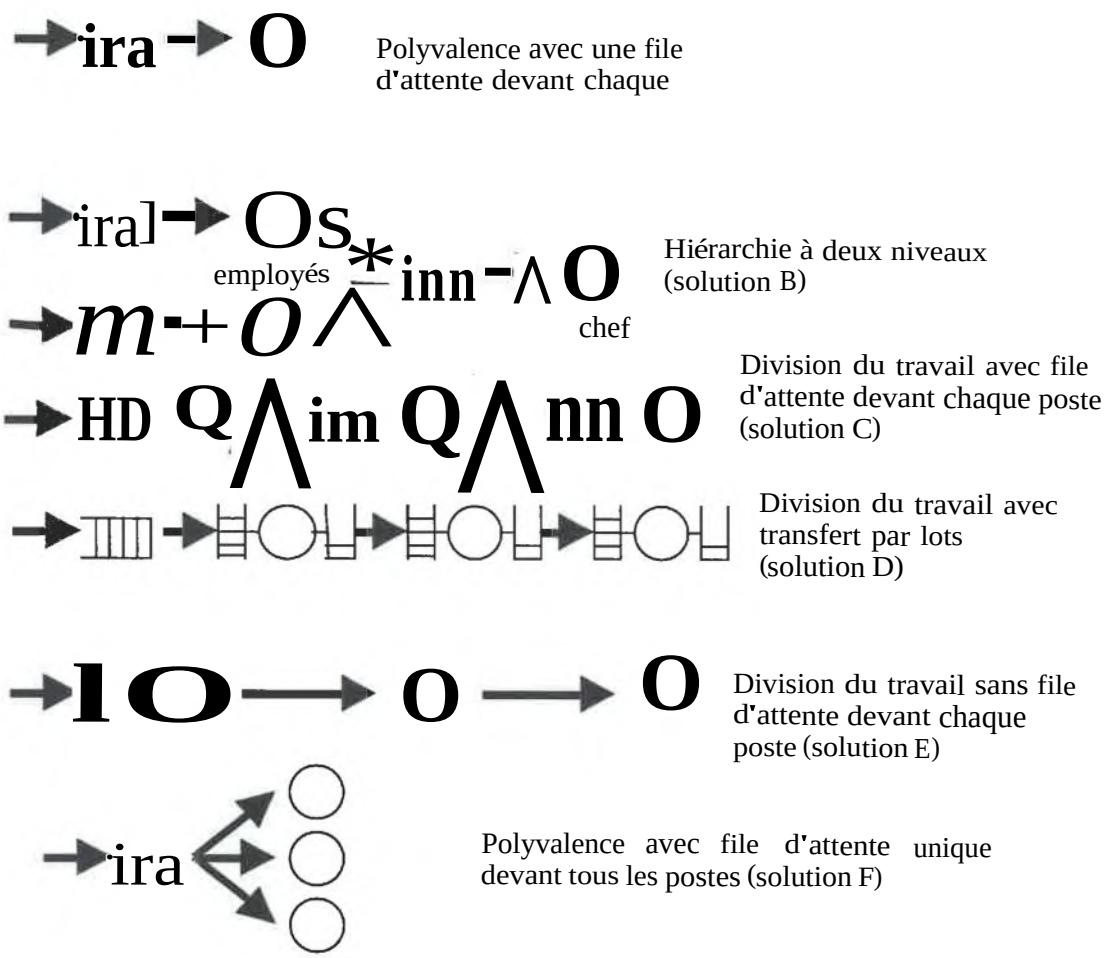

Figure 1 : organisation polyvalente et organisation avec division du travail

Les solutions archétypales diffèrent par la manière de diviser le travail, de le répartir entre personnes. La division du travail est évoquée depuis Adam Smith comme moyen d'augmenter la productivité. Un autre facteur intervient : la manière de s'organiser face à l'arrivée aléatoire du travail. On étudie ici ce seul facteur. La productivité est donc supposée constante, c'est-à-dire que le temps de travail passé sur chaque dossier est constant, qu'il soit réalisé par une personne ou par plusieurs successivement. De plus, on n'étudiera pas l'effet des salaires. Ceux-ci sont supposés constants entre les personnes.

Les six solutions archétypales sont symbolisées dans la figure 1 ; voici leurs principes et leur fonctionnement en termes de délais et de coût.

A - *Poste de travail unique traitant les dossiers de bout en bout*

Pour faire face à une arrivée aléatoire de travail, la solution la plus simple consiste à installer une file d'attente devant un poste de travail unique (souvent appelé « serveur »). Dès que le serveur a fini de travailler sur un cas, il traite le suivant dans la file d'attente. Si la file d'attente est vide, il reste inoccupé. Les cas nouveaux se placent à la fin de la file d'attente. C'est la première solution, celle d'un poste de travail unique. Cette solution est traitée rigoureusement par la théorie mathématique des files d'attente, classique en recherche opérationnelle (voir annexe).

Si le flux de travail est trop important, on le divise avant de l'envoyer sur chaque poste de travail. Il y a autant de files d'attente que de postes de travail. La division du flux de travaux est faite souvent sur une base géographique (territoires de compétence) de manière à ce que les flux soient égaux. En pratique, cette égalité des flux n'est pas toujours parfaitement réalisée.

B - *Hiérarchie avec transfert des dossiers au fil de l'eau*

La deuxième solution archétypale se rapporte à un ensemble de postes de travail recevant indépendamment des dossiers aléatoires (par exemple provenant d'un territoire dont ils sont chargés). Chaque poste de travail est précédé d'une file d'attente qui lui est propre. Il la gère comme dans le cas (A) précédent. Un chef vérifie ensuite chaque dossier. Il stocke les dossiers dans une file d'attente. Les dossiers attendent successivement dans une file d'attente à chaque niveau hiérarchique. Le temps total de travail sur un dossier est réparti en temps de travail du chef et en temps de travail de ses collaborateurs. Ce modèle a été développé par Martin Beckmann (1988) dans le cadre de ses modélisations du fonctionnement de la hiérarchie dans les organisations.

Quand les salaires sont égaux, cette solution à deux niveaux hiérarchiques est équivalente à celle du serveur unique en termes de coût et de délais. C'est étonnant, puisque les dossiers attendent dans deux files différentes. Mais le temps total de traitement est divisé en deux parts. Devant chaque poste de travail, l'attente est plus courte. Comme cette attente est proportionnelle en moyenne au temps de traitement du poste de travail, le total des temps d'attente reste constant. Le temps de traitement aussi (par hypothèse). En conséquence, le modèle hiérarchique est équivalent au modèle de serveur unique.

Quand le chef a un salaire supérieur, le coût est supérieur pour un même délai, sauf à compenser par un salaire moindre des collaborateurs.

C - *Division du travail avec transfert des dossiers au fil de l'eau entre bureaux communs à plusieurs procédures*

Il est aussi possible de créer une division du travail. Le traitement du dossier est partagé entre une séquence de bureaux. Chaque bureau effectue une partie spécialisée du travail global.

Ces bureaux successifs ne sont pas spécialisés dans la procédure considérée. Ils traitent d'autres dossiers, provenant d'autres procédures. Par exemple, l'étape de contrôle juridique est faite dans un service spécialisé. Mais ce service traite des aspects juridiques de tous les problèmes traités dans l'entreprise. Il intervient dans plusieurs procédures différentes. Spécialisé dans le droit, il a une polyvalence sur le genre de dossiers.

Il existe alors une file d'attente devant chaque poste de travail contenant des dossiers venant des différentes procédures. Si chaque bureau est composé d'une seule personne et que les salaires sont égaux, ces procédures enchevêtrées sont équivalentes à un processus de serveur unique. La raison en est la même que pour la hiérarchie (B) de Martin Beckmann. Le temps de traitement à chaque bureau n'est qu'une partie du temps de traitement global. En conséquence la longueur de la file d'attente est plus petite. La somme des temps de traitement est la même que celle devant un poste de travail unique polyvalent. La somme des attentes moyennes aussi.

D - *Division du travail avec constitution de lots transférés en bloc*

Supposons maintenant que les différents bureaux qui se partagent le travail en séquences soient spécialisés sur la procédure. Une solution consiste à lancer le traitement des dossiers selon une période régulière (tous les jours, tous les mois, tous les ans). Devant le premier bureau, les dossiers attendent. Quand le lot est constitué, le premier bureau traite tous les dossiers. Quand il a fini, il transfère tous les dossiers au bureau suivant. Les dossiers sont transférés en bloc de bureau en bureau. Le travail est terminé quand tous les dossiers sont traités.

Le traitement par lots est une très mauvaise solution en coûts et en délais. La raison est bien connue des organisateurs d'usine. La taille du lot de transfert influe fortement sur les délais. En effet, les éléments d'un lot attendent à chaque poste de travail que les autres éléments soient traités. Plus il y a de dossiers dans le lot, plus le délai est long. La taille du lot dépend de la périodicité (à même flux d'arrivée). La périodicité est donc le facteur dominant de l'explication du délai. Cette solution s'améliore avec une accélération de la périodicité. Elle reste toujours moins bonne que le traitement au cas par cas (au fil de l'eau) sur un serveur unique.

Les traitements par lots périodiques, comme il en existe beaucoup dans les processus administratifs (comptabilité, impôts, examens) ont une mauvaise performance en termes de délais et de coût.

E - *Division du travail avec transfert des dossiers au fil de l'eau*

Le travail est réparti comme dans la solution précédente entre plusieurs bureaux successifs dédiés à la procédure, mais les cas sont traités dès qu'ils arrivent (au fil de l'eau). Il existe une file d'attente devant le premier bureau. Le délai de traitement du premier bureau est aléatoire. Si le temps de travail de chaque bureau ultérieur est strictement égal à ce temps mis par le premier bureau, il n'y a pas de file d'attente devant les autres bureaux. En négligeant les délais de transmission entre postes de travail successifs, cette solution (E) est meilleure que celle du serveur unique (A).

Les dossiers n'attendent qu'une fois, devant le premier bureau. Cette attente est réduite parce que le temps de traitement moyen par le premier bureau n'est qu'une fraction du temps de traitement global. Ce délai d'attente est d'autant plus petit qu'on a divisé le travail en un grand nombre d'étapes. La supériorité de ce modèle s'accroît avec la division du travail. C'est un effet d'économies d'échelle, car il faut avoir suffisamment de travail pour pouvoir le diviser entre beaucoup de personnes spécialisées.

Si on suppose que les temps de travail de chaque bureau sont égaux en moyenne mais indépendants, ils ne sont pas égaux pour un cas donné. Alors une file d'attente se crée devant chaque bureau. Cette situation est celle du cas C avec une seule procédure occupant tous les bureaux.

F - *Postes de travail traitant les dossiers de bout en bout avec une file d'attente commune*

La dernière solution archétypale considérée ici reprend le principe de personnes polyvalentes traitant la totalité des dossiers, de bout en bout (A). Mais plusieurs postes se partagent le flux arrivant. Ils peuvent traiter indifféremment tous les dossiers qui surviennent. Il existe une seule file d'attente pour tous. C'est pourquoi on parle souvent de "serveurs en parallèle". Dès qu'une personne est libre, elle prend le premier dossier de la file d'attente. S'il n'y en a pas, elle reste inoccupée. Cette organisation est celle des centres d'appels à un seul niveau.

D est bien connu que la solution F est meilleure que la solution A. Elle est aussi meilleure que la solution E. Elle minimise les occasions pour une personne de rester inoccupée. En effet, dans la procédure séquentielle (E), il existe des moments où la file d'attente est vide. Le premier bureau ne fait rien à ce moment. Cette inoccupation se propage ensuite vers les bureaux suivants. Quand plusieurs dossiers arrivent en même temps, on dépasse la capacité du premier bureau. Il y a une attente dans la file. Simultanément, il existe des personnes inoccupées dans les bureaux (autres que le premier).

Au contraire, quand les serveurs sont polyvalents (F), il n'y a aucun moment où des dossiers attendent alors que des personnes sont inoccupées. Chaque personne traitant tout dossier indifféremment, elle a comme règle de se saisir du premier dossier en attente, s'il y en a, dès qu'elle est inoccupée. Cette caractéristique est la raison pour laquelle, en termes de délai et de coût, cette solution surclasse celle de la procédure séquentielle au fil de l'eau (E) et donc toutes les autres. Cette dominance n'est pas fondée sur une réduction des délais de transmission puisque les dossiers sont traités par une seule personne. Plus le nombre de personnes est grand, plus cette dominance est forte.

Dans cette comparaison, on n'a supposé aucun effet de productivité découlant de la division du travail. Le temps de travail global est seulement coupé en temps de travail de chaque bureau. Par ailleurs, on ne suppose pas d'effet de salaire. Quand les salaires sont différents (cas traité par Beckmann), le salaire élevé du chef augmente les coûts.

Les solutions E et F, qui sont les meilleures, ne peuvent être mises en place sans un flux de travail suffisant pour le répartir entre plusieurs personnes. Plus ces personnes sont nombreuses, meilleures sont ces solutions. C'est un effet d'économies d'échelle.

Ces arguments sont construits sur l'identification des attentes et leur concomitance avec une inoccupation de personnes. Ils permettent un classement entre les solutions archétypales. Avec des hypothèses restrictives, un raisonnement quantitatif précis est possible en utilisant les résultats de la théorie des files d'attente. Chaque solution archétypale est caractérisée par une courbe de coût en fonction du délai (voir figures 3 et 4). L'annexe de cet article donne les équations de chaque cas. On y retrouve, naturellement, le même classement entre les solutions archétypales. L'avantage de cette approche quantitative est de pouvoir caractériser l'ampleur des gains ainsi obtenus. Pour rendre concrets ces calculs, prenons des exemples.

3. Illustration des solutions archétypales et discussion des hypothèses

En fonction des équations indiquées en annexe, il est possible de calculer les délais et les coûts dans des cas particuliers. Considérons un flux de travail de 45 dossiers à traiter par heure. Supposons que les personnes passent 4 minutes en moyenne pour traiter un dossier, soit une capacité de travail de 15 dossiers par heure. Voici les six organisations possibles pour 6 personnes, équivalent temps plein ou à temps plein, payées au même salaire. Le coût sera constant, double du coût correspondant au temps passé en traitement des dossiers, car la capacité de traitement est de 90 dossiers (6 x 15) par heure. Calculons le délai d'attente moyen WT dans chacune des six solutions archétypales.

A - *Poste de travail unique traitant les dossiers de bout en bout.*

Le flux arrivant est identifié par région. Chacune des six régions apporte 7,5 dossiers par heure à une personne qui les traite de bout en bout.

WT = 8 minutes (moitié par attente et moitié par traitement)

B - *Hiérarchie avec transfert des dossiers au fil de l'eau.*

Le flux arrivant est identifié par région. Mais on a un découpage en cinq régions dont chacune apporte 9 dossiers par heure à une personne qui les traite puis les passe à un contrôleur unique (chef) qui traite le flux de 45 dossiers par heure. Supposons que le chef ait le même salaire que ses collaborateurs. Le travail est réparti de telle manière que le chef traite 90 dossiers à l'heure et ses collaborateurs 18 dossiers. Chacun a donc une capacité de travail qui est le double du flux réel.

WT = 8 minutes (200 secondes en attente devant les subordonnés, 200 secondes de traitement par les subordonnés, 40 secondes en attente devant le chef et 40 secondes en traitement par le chef)

Pour illustrer l'effet des écarts de salaires, supposons que le chef soit payé deux fois plus et que ses collaborateurs soient réduits à quatre avec le même salaire (quatre régions). Le coût total est le même. Alors, il faut que le chef soit capable de traiter 68 dossiers par heure et que ses quatre collaborateurs aient une capacité de traiter 19 dossiers par heure chacun. Le délai de traitement est plus long :

WT = 10,2 minutes (1,7 minutes en attente devant le chef et 0,9 minute de traitement par le chef, 4,5 minutes en attente devant chaque collaborateur et 3,1 minutes de traitement par eux)

C - *Division du travail avec transfert des dossiers au fil de Veau entre bureaux communs à plusieurs procédures.*

Plusieurs (n >= 6) bureaux interviennent successivement sur les dossiers. Les dossiers attendent dans une file d'attente à chaque bureau. Chaque bureau traite les dossiers en 4/n minutes en moyenne. Chaque bureau est supposé composé d'une seule personne ayant une charge de travail de 50% (pour avoir les mêmes coûts). Si l'enchevêtrement des procédures restitue l'aléa supposé pour une file d'attente avec un seul serveur, chaque dossier attend en moyenne 4/n minutes devant chaque poste de travail.

WT = n (4/n + 4/n) minutes = 8 minutes

D - *Division du travail avec constitution de lots transférés en bloc.*

Toutes les heures, le premier bureau reçoit 45 dossiers. Chaque dossier attend en moyenne une demi-heure avant d'être traité. Quand le premier bureau a reçu tous les dossiers, il commence à travailler dessus. Une personne travaillant à raison de 40 secondes par dossier y consacre une demi-heure. Ensuite, les cinq autres bureaux traitent chacun le lot des 45 dossiers en une demi-heure.

WT = 3,5 heures (dont 4 minutes de traitement)

Si les dossiers étaient rassemblés en lots journaliers de 360 dossiers :

WT = 3,5 jours (dont 4 minutes de traitement)

Ces chiffres sont des moyennes. Les bureaux sont inoccupés, en moyenne, la moitié du temps. En fait, on accepterait dans cette solution une charge de travail moyenne beaucoup plus proche de la capacité et les délais doubleraient.

E - *Division du travail avec transfert des dossiers au fil de l'eau.*

Les dossiers sont traités successivement par les six personnes qui y consacrent en moyenne 40 secondes chacune. Les dossiers attendent 40 secondes devant le premier poste. Ils n'attendent pas devant les autres postes parce que ceux-ci ont exactement le même temps de traitement que le premier poste.

WT = 4,7 minutes (dont 4 minutes de traitement)

Si les temps de travail de chaque bureau étaient indépendants (au sens des probabilités), il y aurait des files d'attente entre bureaux. On retrouverait le cas C avec une file d'attente devant chaque bureau.

F - *Postes de travail traitant les dossiers de bout en bout avec une file d'attente commune.*

Les six postes de travail traitent les dossiers de bout en bout (4 minutes par dossier). Ils se partagent une file d'attente commune où arrivent 45 dossiers par heure.

WT = 4,13 minutes (dont 4 minutes de traitement)

Selon la solution, le délai, à coût constant, varie de plusieurs jours à quelques minutes. Les performances sont considérablement améliorées par la mutualisation de la file d'attente (F) comme dans les centres d'appels à un seul niveau.

Dans le tableau 1, on donne un autre exemple. Pour les mêmes six personnes, travaillant avec le même rythme, le flux moyen de travail est de 84,1 dossiers par heure, proche de leur capacité (90 par heure). Le coût par dossier est amélioré de 46,5 %. Les délais sont considérablement accrus. Ils restent raisonnables pour la solution F. Le changement, à délai constant de 8 minutes environ, de la solution A vers la solution F peut faire gagner près de 45% en coût.

	X = 45, (i = 15 dossiers par heure, n = 6 personnes	X = 84,1 , p = 15 dossiers par heure, n = 6 personnes
solution archétypale	**délai moyen de traitement WT**	**coût moindre de 46,5% délai moyen de traitement WT**
A, B et C	**8 minutes**	**61 minutes**
D (lots de 1 heure)	**3,5 heures**	**6,1 heures**
E	**4,7 minutes**	**13,5 minutes**
F	**4,13 minutes**	**8,25 minutes**

Tableau 1 : Deux exemples de délais dans les six solutions archétypales.

Revenons sur les hypothèses faites lors de ces comparaisons. On a supposé nuis les délais de transmission. Si ces délais ne sont pas nuis, les solutions avec une division du travail (B, C, D et E) sont encore moins bonnes. La domination de l'organisation en centres d'appels avec une file d'attente unique (F) devient encore plus forte. Lever cette hypothèse ne modifie pas la conclusion.

Deux solutions présentent une division du travail avec traitement au fil de l'eau. Dans la solution C, il y a une attente devant chaque poste de travail. Les postes de travail sont indépendants. Dans la solution E, les temps de travail sur un cas sont rigoureusement égaux d'un poste à un autre. Ils dépendent, par exemple, de la complexité des dossiers. Il n'y a alors qu'une seule file d'attente. On peut imaginer des solutions intermédiaires créant des files d'attente non nulles mais plus réduites que pour la solution C. Ces solutions auraient une performance intermédiaire. La dominance de la solution F n'en est pas modifiée.

On a supposé que les tâches divisées demandaient des compétences que toutes les personnes pouvaient acquérir. La polyvalence était possible. Corollaire de cette hypothèse, les salaires sont considérés comme identiques. Or la division du travail trouve tout son sens dans les cas de compétences très spécifiques, exercées par des personnes à haut salaire qui laissent

les parties simples du travail à des personnes peu payées. L'organisation en centres d'appel (F) est alors soit impossible soit irréaliste. Il faudrait y placer des personnes trop payées dont la haute compétence ne serait pas employée tout le temps. Dans ces cas, les organisateurs conservent la division du travail et ils préconisent une Gestion Electronique de Documents pour réduire les délais de transmission. Les files d'attente sont alors négligeables.

Peut-être est-il bon de revenir sur une hypothèse évoquée rapidement au début de cet article. On a considéré les travaux du tertiaire répétitifs, ceux pour lesquels des procédures fixent l'organisation du travail. Il existe bien d'autres formes de travail arrivant aléatoirement dans les entreprises de service. Le cas du travail pour lequel il n'existe pas de procédures mais où un ensemble déterminé de personnes interviennent, par des réunions ou par des échanges désynchronisés sur le réseau a déjà été présenté dans la revue *SIM* (Peaucelle, 1998).

La comparaison des solutions archétypales fournit une conclusion simple : il faut mettre en place, quand cela est possible, des structures d'organisation avec serveurs polyvalents, des centres d'appels avec un seul niveau (F). Les préceptes du BPR, exposés par Hammer (1990), sont-ils cohérents avec cette conclusion si simple ? Cela mérite d'être examiné en détail.

4. Interprétation des principes du BPR

Les sept principes du BPR, tels que Hammer les a avancés en 1990, peuvent s'interpréter au regard des solutions archétypales. Il est temps de les examiner soigneusement. On les passe en revue ci-dessous.

4.1 Polyvalence des personnes de bout en bout dans le processus ("to have one person to perform all the steps in a process")

La même personne réalise toutes les étapes du processus. Elle suit le dossier de bout en bout. Les tâches traditionnellement séparées sont intégrées dans un même poste.

Ce principe va complètement à l'encontre de toutes les habitudes de division du travail. Dans la production industrielle, on a depuis très longtemps spécialisé les personnes, comme Adam Smith l'avait remarqué. Les raisons de cette spécialisation étaient :

- l'amélioration de l'habileté sur une tâche réduite (à condition qu'elle ne soit pas contrebalancée par l'ennui),
- la faible compétence de la main-d'œuvre (avec une division du travail très poussée, on peut former rapidement les personnes),
- les aptitudes inégalement réparties chez les travailleurs (des personnes sans formation, donc moins payées, exercent les postes les plus simples),
- l'adaptation aux machines qui sont les auxiliaires du travail industriel pour des étapes bien précises (les machines permettent une augmentation de productivité).

On sait que la division du travail trouve probablement son origine dans la différenciation des salaires de la chaîne de production (Peaucelle, 1999 b). Les ouvriers les plus compétents peuvent travailler de manière polyvalente, ils préfèrent se spécialiser sur les postes les mieux payés. Ils occupent dans la chaîne de valeur les points où leur rémunération est la plus forte.

Certaines de ces raisons continuent d'être vraies dans les entreprises de service en cette fin du 20° Siècle. Les caractéristiques de la population des pays industrialisés et les technologies informatiques rendent ces raisons moins pressantes :

- les employés ont une compétence de base assurée par le système éducatif et on les sélectionne, au moment de l'embauche, en fonction de cette compétence,
- la formation permanente permet d'accroître rapidement la compétence et l'adaptabilité à de nouveaux modes de travail,
- les machines informatiques ont une polyvalence qui permet aujourd'hui d'accéder à toute la variété des programmes et des ressources depuis un poste de travail standard (alors que les dossiers et archives papier n'étaient accessibles que dans un service unique spécialisé sur une tâche),
- la compétence est épaulée et contrôlée par les programmes informatiques (par exemple les systèmes d'assistance peuvent de manière experte apporter la connaissance d'une réglementation appliquée très rarement).

La diversité des compétences reste un argument d'actualité qui peut justifier une place à part pour les personnes disposant d'une compétence particulière (en général avec des diplômes) afin de les utiliser au mieux compte tenu de leur salaire plus élevé. Elle reste un principe de différenciation des tâches. Par exemple, restent spécialisées les tâches de signature par le chef hiérarchique, seul habilité à engager juridiquement l'entreprise.

Si les arguments de la division du travail cessent partiellement d'être vrais, il faut considérer les arguments inverses. Herzberg (1959) plaidait dès la fin des années 1950 pour élargir les tâches et les enrichir, c'est-à-dire regrouper les tâches divisées au long de la séquence des opérations et intégrer dans l'exécution quelques tâches de contrôle ou quelques tâches fonctionnelles (petite maintenance par exemple). Mais son objectif premier était la motivation et ce n'est qu'au travers de celle-ci qu'il visait la productivité.

La division du travail a un inconvénient bien connu. Il faut assurer le déplacement des matières (ou des dossiers dans les services) d'un poste de travail à un autre. Avec la chaîne, Ford mécanise ce déplacement. Dans les services, on a mis en place des systèmes automatiques de transport des dossiers papier, plus ou moins perfectionnés. En allant plus loin, le workflow (Gestion Electronique de Documents) transforme tous les dossiers en image informatique et les réseaux assurent leur communication d'un poste de travail à un autre. Ces systèmes automatisent le passage d'un poste de travail à un autre, donc réduisent les délais et suppriment les postes des personnes (en général peu formées) chargées de faire circuler les dossiers.

Pourtant le premier principe de Hammer ne fait pas allusion à ses systèmes qui conservent la division du travail en limitant seulement ses inconvénients. En concevant un poste qui assure toutes les tâches du processus, on change radicalement de modèle d'organisation, et les coûts et les délais associés sont meilleurs. Bien entendu ce n'est pas toujours possible.

En faveur du regroupement de tâches, on pourrait avancer d'autres arguments. Dans une procédure séquentielle, chaque bureau prend connaissance du dossier. Ce temps de lecture se répète. En supprimant la division du travail, on supprime cette duplication de la lecture. Il est donc possible que le temps total passé par la personne polyvalente pour traiter le dossier de bout en bout soit inférieur au total des temps éclatés. Les comparaisons faites ci-avant ne prenaient pas en compte cet effet, pas plus que celui, inverse, d'accroissement de productivité par division du travail.

Dans tous les exemples donnés par les chercheurs ou les consultants, on note la diminution du nombre d'étapes dans les procédures. Par exemple Hammer (1993) montre que la compagnie de téléphone Bell Atlantic a raccourci le délai de raccordement des nouveaux abonnés de 15 jours à 5 heures en constituant des équipes de 5 à 6 personnes à compétence transfonctionnelle dans un seul lieu, pour toute la chaîne du travail effectuée auparavant par 13 personnes en séquence, installées dans des lieux différents.

Le premier principe de Hammer consiste à abandonner une organisation avec division du travail (solution (E)) pour une organisation de personnes polyvalentes partageant la même file d'attente (F), un centre d'appels à un seul niveau (voir figure 2).

Faire traiter les cas de bout en bout par la même personne suppose des outils informatiques puissants pour qu'elle dispose, sur son poste de travail, de tous les éléments du dossier qui est traité à ce moment, quel que soit ce dossier. Ce ne serait pas possible avec des archives papier. L'informatique aide à rendre possible la solution F.

4.2 Fournir les informations et des aides informatiques partout ("computer-based data and expertise are more readily available")

Le deuxième principe signalé par Hammer (1990) n'est qu'une condition pour que le premier soit réalisable. Il s'agit de fournir toute l'information à la personne qui traite le dossier. L'accès aux bases de données, la construction de systèmes experts spécialisés, la fourniture de documentation en ligne sont des moyens pour rendre compétentes des personnes, sans formation ni mémoire considérable sur chaque cas qui se présente.

Ces aides informatiques au travail permettent aux opérationnels de faire tout seuls, sans attente, sans plus d'erreurs, ce qu'ils devaient demander aux services fonctionnels spécialisés (comptabilité, achats, maintenance). On passe de la solution (C) à la solution (F) (voir figure 2). Les procédures ne peuvent être abandonnées au profit de tâches globalisées que si ces traitements de l'information sont accessibles aux opérationnels. Si la mise en place de ces outils d'aide n'est que progressive, les transformations organisationnelles suivent ce rythme. Les services informatiques sont ainsi des partenaires obligés de toutes les actions de BPR.

La perspective de l'informatisation associée au BPR n'est donc pas de remplacer les personnes dans leurs tâches mais de déplacer le lieu de traitement. Du service spécialisé, il migre vers les personnes polyvalentes qui traitent les cas de bout en bout. C'est en ce lieu que sont fournies les informations, au moment où les gens en ont besoin. L'informatique communique et suit le rythme de travail des personnes.

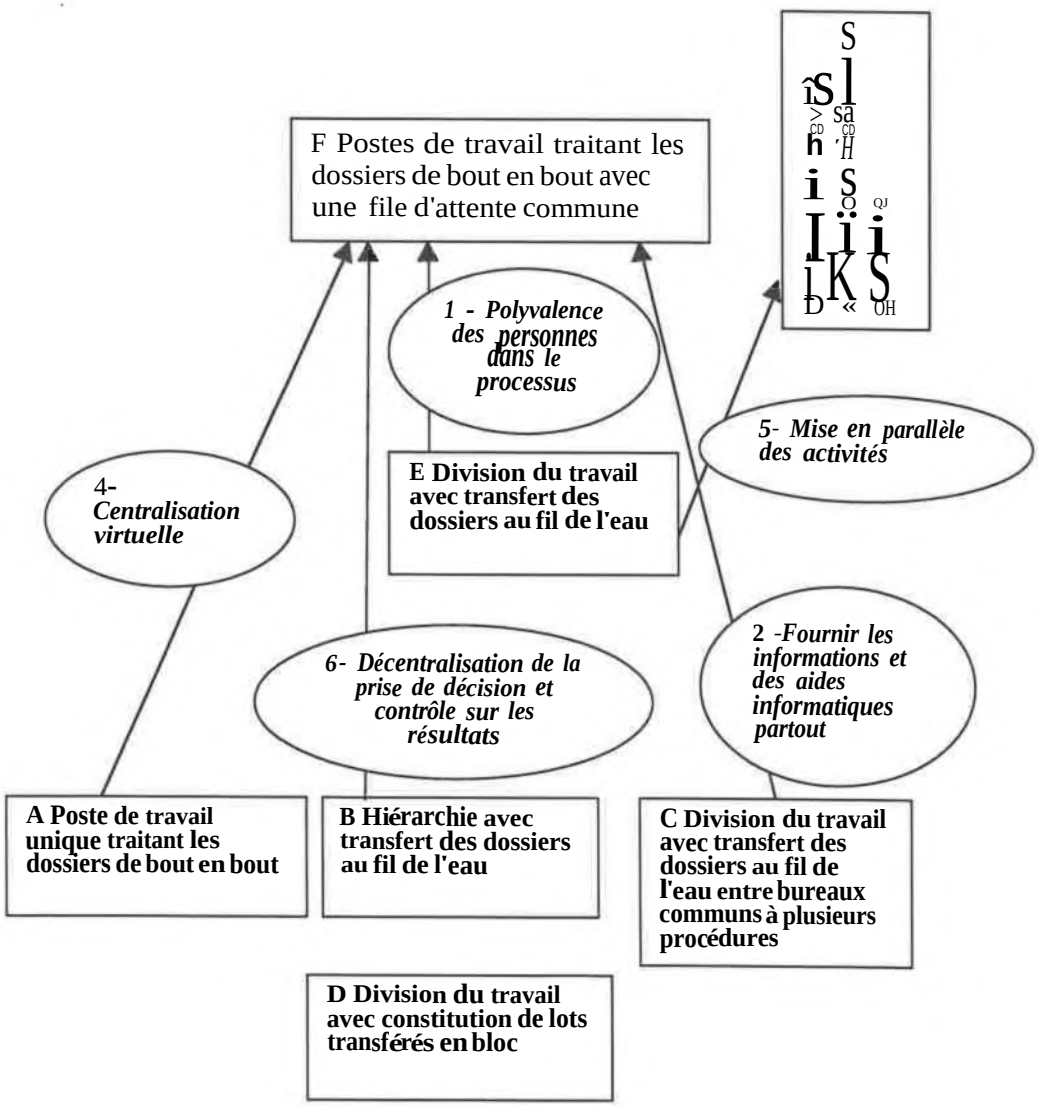

Figure 2 : interprétation des principes du BPR selon Hammer (1990)

Lucas, Bemdt et Truman (1996) donnent un exemple de reconfiguration du processus de traitement des valeurs sur support papier déposées par les clients aux agences de Merrill Lynch. Ces valeurs sont envoyées par la poste au centre unique de traitement. Une part du temps était

consommé en allers-retours entre le centre et les agences pour des anomalies (documents annexes non fournis, valeurs n'ayant pas cours légal...). Un système expert, accessible dans les agences, permet de consulter toutes les valeurs en circulation et de décrire tous les cas où des documents annexes sont nécessaires (certificat de décès en cas d'héritage par exemple). On réduit alors à la fois le temps passé par les agents (donc le coût) et le délai pour le client (à cause des allers-retours). Des solutions de workflow ont aussi été mises en œuvre. Les valeurs sont scannées (ou microfilmées si elles sont de format trop important). Une part du traitement est faite après reconnaissance optique du document s'appuyant sur une base de données des "valeurs papier" en circulation.

Cette migration des tâches spécialisées vers les agents polyvalents correspond à un abandon du modèle (C) de bureaux spécialisés communs à plusieurs procédures. On a vu qu'effectivement il n'est pas le meilleur (s'il n'y a pas d'effet de productivité). Il faut cependant vérifier que les tâches spécialisées sont bien accomplies par la personne polyvalente aidée de son informatique. Ceci dépend sans doute de la complexité des cas.

4.3 *Intégration du système d'information au monde réel ("subsume information-processing work into the real work that produces the information")*

Traditionnellement, les activités administratives, d' « écritures », sont effectuées par des employés spécialisés. Il s'agit de la distinction bien connue entre les cols blancs et les cols bleus. Les personnels des ateliers ne sont pas amenés à tenir des crayons et des papiers. Quand ils le font, on sait que la qualité de leur écriture n'est pas fameuse.

L'informatique renouvelle cette approche. On peut, par des terminaux, spécialisés ou non, faire réaliser la saisie des informations par les personnes qui sont le plus directement en contact avec le « monde réel », celui sur lequel portent les informations. Avec cette intégration, quand elle est possible, on supprime les postes administratifs par informatisation.

La saisie des données par les productifs, par les clients, par les demandeurs exige des systèmes informatiques spécialisés, sécurisés, conviviaux. Ici encore la réalisation informatique est une partie essentielle du BPR.

En intégrant système d'information et monde réel, on supprime des étapes dans une procédure, par automatisation. Au bout de ce principe, on supprime tout traitement manuel de l'information. Après la saisie par le client, toutes les tâches sont prises en charge par la machine.

Cette perspective de totale automatisation est loin d'être neuve. Dans le cas de Bell Atlantic, rapporté par Hammer (1993) et cité plus haut, on veut aller vers du zéro délai d'installation de service téléphonique. Le client saisira lui-même sa demande (par Internet) et ce sont des programmes qui déclencheront les actions sur les centraux téléphoniques pour que le système soit mis à disposition. C'est possible si l'infrastructure, les lignes notamment, est déjà en place. La partie administrative, signature de contrat, acceptation du tarif, sera prise en charge après la mise à disposition du service.

4.4 *Centralisation virtuelle ("treat geographically dispersed resources as though were centralized")*

Le quatrième principe concerne les grandes entreprises qui disposent de localisations géographiques sur un territoire : agences, bureaux, boutiques... Si on les regroupe, on bénéficie des économies d'échelle qu'il y a entre le modèle de serveur unique répété (A) et la mise en commun de la file d'attente (F). On évite ce gâchis d'un serveur inoccupé à un endroit alors qu'il existe des dossiers en attente dans un autre endroit.

L'avantage supplémentaire de la centralisation vient de l'ajustement plus facile aux flux. Les bureaux répartis n'ont jamais une charge complètement égale, même si, au moment de leur création, leurs charges sont similaires. Les évolutions créent des distorsions entre bureaux locaux. La mise en commun de la charge de travail évite l'existence de surcharges locales, avec des charges très faibles en d'autres lieux.

Ces économies d'échelle face aux aléas de l'arrivée des dossiers sont les mêmes que celles qui sont entraînées par la centralisation des stocks. Le stock centralisé évite les situations de rupture dans un dépôt local avec des marchandises dans d'autres dépôts.

Bien sûr, il faut que ce regroupement soit possible, c'est-à-dire que le service soit offert localement malgré le regroupement géographique. Les technologies de la communication rendent l'accès à distance possible. Les exemples sont bien connus avec les centres d'appels qu'on retrouve ici comme un résultat du BPR.

Mais la technologie va plus loin. Il n'est pas nécessaire de déplacer physiquement les personnes en un seul lieu. Puisque leur travail est délocalisé par rapport à la demande, il peut l'être par rapport à l'équipe des personnes substituables. La délocalisation du travail rend possible la constitution d'un service unique sans déplacer les personnes. La centralisation géographique peut être virtuelle. La charge de travail est mutualisée entre tous les serveurs. Dès qu'un dossier attendrait sur son site local, il est proposé aux agences éloignées géographiquement qui sont libres à ce moment. Les centres locaux s'échangent la charge de travail. C'est le regroupement en équipes virtuelles préconisé par Hammer. Sur la figure 2, il est représenté par le passage de solutions (A) répétées avec autant de files d'attente à la solution (F) d'une file d'attente commune.

Le regroupement permet aussi d'ajuster les effectifs à la saisonnalité de la charge de travail : variation dans la journée, dans la semaine, dans l'année. Quand les effectifs sont limités, voire réduits à une seule personne, il n'y a pas de modulation possible. Quand les effectifs sont importants, cet ajustement limite le surcoût découlant d'effectifs pléthoriques à certains moments et les délais d'attente des dossiers lors des périodes de pointe.

Wareham et Neergaard (1998) rapportent le cas de la réorganisation de la compagnie d'assurance mutuelle Alka au Danemark. La réorganisation a eu comme principe directeur le « guichet unique ». Chaque client doit avoir un seul point de contact avec Alka, directement par téléphone, pour gérer son contrat (ventes) et un autre pour gérer son sinistre (déclaration, remboursement...). Antérieurement il devait s'adresser successivement par courrier aux divers services spécialisés. A la place de 26 agences

locales, ce sont 6 agences régionales puis trois qui assurent les contacts commerciaux. Les auteurs n'indiquent pas les effectifs consacrés à la vente. Ils disent que 56 postes de travail ont été supprimés (sur un effectif total de 375 personnes). Et la compétence (polyvalence) du personnel en place a été largement augmentée. Les procédures, derrière le guichet unique, continuent d'exister. Elles ont été simplifiées avec un maximum de 69 étapes réduit à 30 pour le commercial et, pour les sinistres, on passe de 193 étapes au maximum à 44 seulement. Le délai moyen passe de 32 jours à 6 jours pour régler un sinistre.

4.5 Mise en parallèle des activités ("link parallel activities instead of integrating their results")

La structure traditionnelle des procédures est la séquence, la succession d'étapes. On copie ainsi la succession des postes de travail de la chaîne. Mais même sur la chaîne des usines, il arrive que deux opérations soient simultanées, par exemple quand elles portent sur des sous-ensembles différents avant leur montage. Pour le traitement d'un dossier, certaines étapes peuvent se dérouler simultanément, d'autant plus aisément que le dossier est dématérialisé. Il est clair que les délais de chaque étape ne s'ajoutent plus. On gagne en délais.

Ce principe de simultanéité des étapes est celui qu'on trouve en production pour le montage des outils et le réglage des machines en « temps caché » (pour réduire les temps d'arrêt entre deux séries de production). Pour gagner du temps, on commence une étape avant que la précédente soit terminée.

Ce principe conduit à une solution qui n'a pas été examinée en tant que solution archétypale. Cette solution est meilleure que la succession des étapes (E). Elle peut être meilleure que (F) ou moins bonne selon les cas. Sur la figure 2, on l'a représentée avec des performances qui peuvent être meilleures ou moins bonnes que (F).

4.6 Décentralisation de la prise de décision et contrôle sur les résultats ("put the decision point where the work is performed and build control into process")

Dans son sixième principe, Hammer parle de la division verticale du travail, entre les exécutants et la hiérarchie. Le traitement d'un cas par les opérationnels fait parfois intervenir des niveaux plus élevés, pour des décisions, des autorisations, des signatures officielles. Cet appel à la hiérarchie est tout à fait similaire à la division fonctionnelle du travail que Hammer veut éviter (deuxième principe).

C'est par des outils informatiques qu'on déplaçait le lieu de réalisation de ces tâches spécialisées. La même solution est proposée pour ces tâches réalisées par la hiérarchie. Le rôle de la hiérarchie est joué par les outils informatiques qui encadrent les actions et enregistrent ce qui a été fait. Les personnes sont alors responsables de leurs actes. La hiérarchie jouera son rôle de contrôle, en intervenant *a posteriori* au lieu d'intervenir *a priori*.

Hammer préconise donc une décentralisation de la décision et du contrôle au point le plus bas, le plus proche du processus du monde réel. La hiérarchie peut ainsi presque complètement disparaître. Ce sont des coûts qui disparaissent, en même temps que les délais sont diminués.

La hiérarchie apparaît dans la solution archétypale (B) de Beckmann. On a vu qu'elle est surclassée par la solution (F) de prise en charge globale d'un dossier par les mêmes personnes, mutualisant leur file d'attente. Le sixième principe de Hammer s'interprète comme un passage à la solution (F) (voir figure 2).

4.7 *Saisie unique à la source ("capture information once and at the source")*

Le septième principe de Hammer concerne la saisie unique à la source. Il s'agit encore de la suppression d'étapes dans le processus de traitement informationnel. Ce principe a été énoncé depuis longtemps pour la conception des systèmes d'information. Il a pour but d'automatiser une partie du processus informationnel, donc de supprimer du traitement administratif manuel, comme le principe 3 dont au fond il ne diffère pas.

5. La cohérence des principes de Hammer

Quatre des sept principes de Hammer s'interprètent précisément comme un changement vers la solution la meilleure en termes de coût et de délai, celle de serveurs suivant de bout en bout le traitement des dossiers et partageant une file d'attente commune (F), c'est-à-dire les centres d'appels à un seul niveau avec des agents polyvalents. Le principe 5, préconisant une mise en parallèle des activités, est de même nature. Au lieu de regrouper les tâches sur une seule personne, on gagne encore en délai si on fait exécuter plusieurs morceaux de la tâche globale simultanément. Les principes 3 et 7 (intégration du SI au monde réel et saisie unique à la source) sont au contraire animés d'un esprit très classique d'automatisation. Là où il y a un travail humain, parfois la machine peut l'accomplir. Dans ces conditions les délais et les coûts sont ceux de l'informatique.

Cette analyse laisse percevoir deux aspects complémentaires de la démarche du BPR. D'une part, on réorganise vers une exécution des tâches globalement, de bout en bout (ou on parallélise les tâches découpées). D'autre part, on informatise. Cette informatisation possède son propre but de substituer la machine à aux personnes dans le traitement des dossiers (automatisation), mais elle est aussi la condition pour rendre possible la réorganisation du travail humain. L'un et l'autre but vont de pair. Les exemples décrits par Hammer ne sont pas suffisamment précis pour distinguer ces deux effets.

Le changement déclenché par le BPR est dit "radical". Le sens de cette expression est rarement précisé. Ce caractère radical vient probablement d'un retour sur la division du travail, d'une remise en cause d'un principe couramment admis et qui est contradictoire avec les objectifs de performance en coût et en délai, face à une arrivée aléatoire du travail.

Cette analyse confirme ce que disent les spécialistes sur le rôle des technologies de l'information dans une démarche de BPR. L'informatique est un passage obligé de la réorganisation. Mais l'essentiel réside dans la réorganisation. Comme disent Davenport et Stoddard (1994), considérer l'informatique comme le moteur du BPR est un mythe. En revanche, elle peut être un inhibiteur qui empêche le changement à cause de retards ou d'impossibilités à

construire les nouveaux systèmes. Si le projet de reconfiguration est géré par les informaticiens, il risque sans doute de ne pas saisir la complexité du changement organisationnel. L'échec potentiel dépend alors d'un pilotage insuffisant.

Il arrive aussi que les changements mis en œuvre ne génèrent pas de gain important. Leurs pilotes n'ont pas de ressources à négocier avec les personnes qui font le travail, sous forme de réduction d'horaires ou d'augmentation des salaires. Ceux-ci ressentent la réorganisation comme une intensification des cadences. Ce n'est probablement pas une implication plus grande des directions générales qui résoudra les conflits potentiels.

La réorganisation doit identifier toutes les particularités du travail qui empêchent d'évoluer vers la solution (F). Parmi ces spécificités, notons la relation non formalisée introduite par les agents sur leur territoire. Beaucoup d'éléments non connus du système d'information peuvent contribuer au succès de la tâche. Banaliser le traitement des dossiers locaux dans un centre unique peut alors dégrader la qualité du travail, qualité perçue par les clients.

Parfois aussi les procédures par lots ne peuvent pas être abandonnées. C'est le cas quand le traitement de chaque dossier interfère avec celui des autres dossiers (classement, répartition de ressource limitée, quotas, concours, budget, postes...) ou lorsqu'on veut traiter la totalité des éléments d'un ensemble (période comptable, territoire géographique).

Les écarts de salaires continuent parfois d'être un bon argument de la division du travail. Les spécialistes bien payés sont plus productifs en étant assistés de personnes dont le salaire est moindre. L'intégration des tâches dans un poste de travail unique est alors peu rentable.

En replaçant chaque principe du BPR dans le mécanisme qui fonde son intérêt, on facilite sans doute la compréhension des actions dans chaque processus. Compte tenu de la multiplicité des principes, il est bien possible qu'ils soient mal interprétés et appliqués hors de propos. Par exemple, il serait sans doute inutile de mettre l'accent uniquement sur les principes 3 et 5 centrés sur l'automatisation alors qu'ils ont déjà montré leurs limites dans les vagues d'informatisation précédentes.

6. Conclusion

Les principes du BPR, exposés par Hammer, sont cohérents avec la recherche d'une organisation performante en termes à la fois de coûts et de délais. Avec les résultats, bien connus, de la théorie des files, on a pu montrer comment l'organisation des serveurs en parallèles est meilleure que d'autres organisations, notamment la hiérarchie et la division du travail. C'est cette organisation qui est préconisée par Hammer. Cette organisation est aujourd'hui rendue possible, grâce à l'informatique, dans les centres d'appels. Les centres d'appels se développent très rapidement aux Etat-Unis et en Europe à cause de leurs performances supérieures pour rendre le service. On a ainsi pu montrer les raisons théoriques de faits constatés depuis quelques années par les praticiens.

Cette rencontre entre deux voies de réflexion sur les processus de travail est très intéressante. La démarche pragmatique et empirique de Hammer parait contradictoire avec nos schémas mentaux qui depuis des siècles privilégient la division du travail. Flip se justifie parce que la performance qui est recherchée aujourd'hui n'est pas seulement celle des coûts réduits mais qu'elle est aussi celle de la réactivité. En divisant le travail en séquences, on se prive des avantages du parallélisme en termes de délais. La théorie des files d'attente connaît depuis longtemps l'intérêt de ce parallélisme.

Les processus administratifs n'obéissent pas toujours aux lois mathématiques simples de la théorie des files d'attente. Leur réorganisation pour améliorer coûts et délais nécessite de simuler les flux de travail. Les outils de simulation utilisés pour la conception des ateliers trouvent ici un emploi tout à fait justifié.

7. Bibliographie

Beckmann M. J., 1988, *Tinbergen Lectures on Organization Theory*, Springer Verlag, Berlin, 252p.

Churchman C.W., Ackoff R.L., Amoff E.L., 1961, *Eléments de recherche opérationnelle*, Traduit et adapté en français par Jean Lavault, Pierre Rosenstiehl, Patrice Bertier, Bernard Roy, Jacques de Guenin et Francis Piquemal, Dunod.

Davenport T.E., Stoddard D.B., 1994, "Reengineering: business change of mythic proportions?", MIS *Quarterly*, June, 121-127.

Davenport T.H., Beers M.C., 1995, "Managing information about processes", *Journal of Management Information Systems*, Summer, Vol. 12, N°1, 57-80.

Hammer M., 1990, "Reengineering work : don't automate, obliterate", *Harvard Business Review*, July-August, 104-114.

Hammer M., Champy J., 1993, *Reengineering the Corporation*, Harper Business, NY.

Hammer M., Stanton S.A., 1995, *The Reengineering Revolution*, Harper Collins, NY.

Hammer M., 1996, *Beyond Reengineering*, Harper Collins, NY.

Herzberg F., Mausner B., Snyderman B., 1959, *The motivation to work*, John Wiley.

Kawalek J. P., 1994, "Interpreting business process re-engineering on organization work-flow", *Journal of Information Technology*, N°9, 276-287.

Kettinger W. J., Grover V., 1995, "Toward a Theory of Business Process Change Management", *Journal of management Information Systems*, Summer, Vol. 12, N°1, 9-30., June, 223-240.

Loebbecke C, Jelassi T., 1998, "Business process redesign at CompuNet, standardizing top-quality service through IT", *Journal of Strategic Information Systems*, Vol. 6, 339-359.

Lucas H.CJr, Bemdt D.J., Truman G., 1996, "A reengineering framework for evaluating a financial imaging system", *Communications of ACM*, Vol. 39, N°5, 86-96.

Peaucelle J-L, 1998, "Fixer une réunion ou travailler ensemble sur le réseau : comparaison des délais d'achèvement", *Systèmes d'information et de Management*, Vol. 3, N° 3, 29-47.

Peaucelle J-L, 1999, "La division du travail : Adam Smith et les encyclopédistes observant la fabrication des épingles en Normandie », *Gérer et Comprendre*, N° 57, 36-51

Smith A., 1776, *Inquiry into the nature and causes of the Wealth of Nations*, London, 3 volumes.

Sillince JAA., Harindranath G., 1998, "Integration of requirements determination and business process re-engineering: a case study of an ambulatory care and diagnostic (ACAD) centre", *European Journal of Information Systems*, N° 7, 115-122.

Wareham J., Neergaard P., 1998, "Still guilty of technical determinism? Reengineering in the insurance sector", 6° *European Conference on Information Systems*, Aix, France, 4-6 June, 1291-1304.

8. Annexe : comparaison mathématique des solutions archétypales

Le raisonnement de comparaison des solutions archétypales s'appuie sur la théorie des files d'attente afin de calculer les délais et les coûts dans chaque cas (comme dans Kawalek, 1994). L'intérêt est alors de quantifier les écarts entre les solutions. Les formules qui sont données dans cette annexe s'appuient sur des hypothèses simplificatrices afin de pouvoir mener les calculs. Dès que ces hypothèses ne sont plus vérifiées, la bonne solution consiste à mener des simulations pour faire les mesures.

8.1 Notation et hypothèses générales cohérentes avec la théorie classique des files d'attente

- dossiers survenant de manière aléatoire selon un processus de Poisson (l'intervalle entre arrivées suit une loi exponentielle)
- λ = taux d'arrivée des dossiers, nombre par unité de temps
- p = capacité de travail du poste de travail, nombre par unité de temps
- $1/p$ = temps moyen passé à travailler sur chaque dossier
- On suppose que la capacité de travail p est supérieure à k le flux de travail arrivant.
- w = salaire par unité de temps des personnes au poste de travail
- WT = temps d'attente moyen d'un dossier + temps de traitement
- WT = temps moyen d'achèvement du travail
- C = coût par dossier, coût total divisé par le nombre de dossiers
- x = délai normalisé = $p\,WT - 1$, le délai normalisé est nul s'il n'y a aucune attente devant le poste de travail.
- y = coût normalisé = $C\,p\lambda / w - 1$, le coût normalisé est nul si les personnes sont occupées durant la totalité du temps de travail.

Les variables x et y normalisées servent à relier coût et délai indépendamment des autres variables telles que C, w, Â et µ. Le coût (y) et le délai (x) évoluent en sens inverse. Dans un plan "coût x délai", chaque solution archétypale est caractérisée par une courbe spécifique entre coût et délai normalisés. La performance d'une solution organisationnelle est caractérisée par la position de cette courbe. Plus la courbe est proche des axes, plus on atteint un délai faible avec un coût réduit et meilleure est la solution (voir figure 4).

8.2 Poste de travail unique traitant les dossiers de bout en bout

Classiquement, la recherche opérationnelle donne les résultats suivants :

$$WT = 1/(\mu - X) \qquad [1]$$

$$C = w / k \qquad [2]$$

Avec les variables normalisées de coût et de délai x et y, ces deux relations deviennent :

$$y = 1 / x \qquad [3]$$

8.3 Hiérarchie avec transfert des dossiers au fil de l'eau

Ce modèle a été élaboré par Martin Beckmann (1988) dans le cadre de ses études sur la modélisation du fonctionnement des hiérarchies. Beckmann calcule le délai de traitement d'une telle hiérarchie. Il s'intéresse de plus à la conception de la hiérarchie. Pour des salaires donnés et des rythmes de travail connus aux divers niveaux, il existe un nombre optimum de subordonnés. Cet optimum équilibre les charges pour minimiser le coût et le délai. A cet optimum, si les chefs ont des salaires supérieurs, leur taux d'occupation est plus élevé.

Cette modélisation est fondée sur le fait que le processus de sortie de la file d'attente simple est le même que le processus d'entrée, c'est-à-dire exponentiel. De plus, les flux exponentiels réunis donnent un flux exponentiel.

Ce modèle B rentre dans la catégorie plus générale des réseaux de Jackson ouverts (Jackson, 1957, Baynat, 2000). Lorsque les salaires des divers niveaux hiérarchiques sont égaux et que le nombre de subordonnés est équilibré par rapport aux rythmes de travail, le modèle B est équivalent à une file d'attente simple (modèle A), où les temps de travail seraient la somme des temps. La relation entre le coût normalisé et le délai d'attente normalisé est la relation [3].

8.4 Division du travail avec transfert des dossiers au fil de l'eau entre bureaux communs à plusieurs procédures

Cette solution C entre dans le cadre des réseaux de Jackson. Ces réseaux permettraient en plus de représenter des bouclages, des retours en arrière. On suppose les temps de transfert nuis entre les postes de travail.

La théorie montre que le temps d'attente moyen est la somme des temps d'attente aux différents postes, pondérés par le nombre moyen de passages au poste.

Si les postes ont le même salaire et le même rythme de travail (équilibrage de la chaîne), la solution C est équivalente à une chaîne unique faisant face au même flux avec un temps de traitement des dossiers égal à la somme des temps de traitement. La relation entre le coût normalisé et le délai normalisé est donnée par la relation [3].

L'équivalence de performance des solutions A, B et C repose sur les hypothèses de transmission instantanée, d'égalité des salaires et sur l'absence d'effet de productivité. La division du travail ne réduit pas le temps total passé à traiter un dossier.

8.5 Division du travail avec constitution de lots transférés en bloc

Les personnes sont dédiées à la procédure. Elles se transmettent le travail par lots de dossiers, constitués par un rythme régulier au premier poste de travail. Par exemple, le lot est constitué de tous les dossiers reçus pendant un mois. A chaque étape, on traite tous les dossiers reçus au cours d'une période. Chaque poste de travail (ou bureau) traite tous les dossiers du lot avant de transmettre le lot à l'étape suivante. On suppose que les bureaux sont affectés à temps plein à la procédure. On suppose qu'aucun travail n'est reporté d'une période à une autre.

Notations :

T périodicité (délai entre deux traitements)

N = X T nombre de dossiers à traiter par période de durée T

n bureaux i installés en séquence

s_i effectifs du bureau i (nombre d'agents)

s = S, s_i effectifs totaux des bureaux

Le délai d'achèvement du travail est la somme du délai de constitution du lot, du temps de traitement dans chaque bureau et du temps de transfert entre bureaux. Le délai de constitution du lot vaut "T/2" si les dossiers arrivent régulièrement, indépendamment des arrêtés de période et si on traite à chaque période tous les dossiers en attente, sans report à la période suivante. Le temps de traitement dans chaque bureau dépend du nombre de cas que traite chaque personne. Les "N" dossiers de la période sont répartis entre les "s_i" personnes du bureau. Le temps de transfert entre bureaux est supposé nul.

$$WT = T / 2 + 2_j \text{ Arrondi.sup}(N/s_i) / v, \qquad [4]$$

En notant "Arrondi.sup(N/s_j)" l'arrondi à l'unité supérieure de N/s_j.

Supposons que le nombre de dossiers par employé N/s, soit important, que le travail ait été réparti également entre les bureaux, que chaque bureau emploie une seule personne, que le temps de transfert entre bureaux soit nul et que les salaires soient égaux, il vient :

$$\forall i \quad \text{Arrondi.sup}(N/s_j \sim N/s_j, \wedge = n\ p, s_i = 1, v_i = w$$

$$WT = T / 2 + XT / p \qquad [5]$$

La durée d'achèvement du travail est liée principalement à la périodicité (T).

Le coût de traitement d'un dossier est :

$$C = n\ w / X \qquad [6]$$

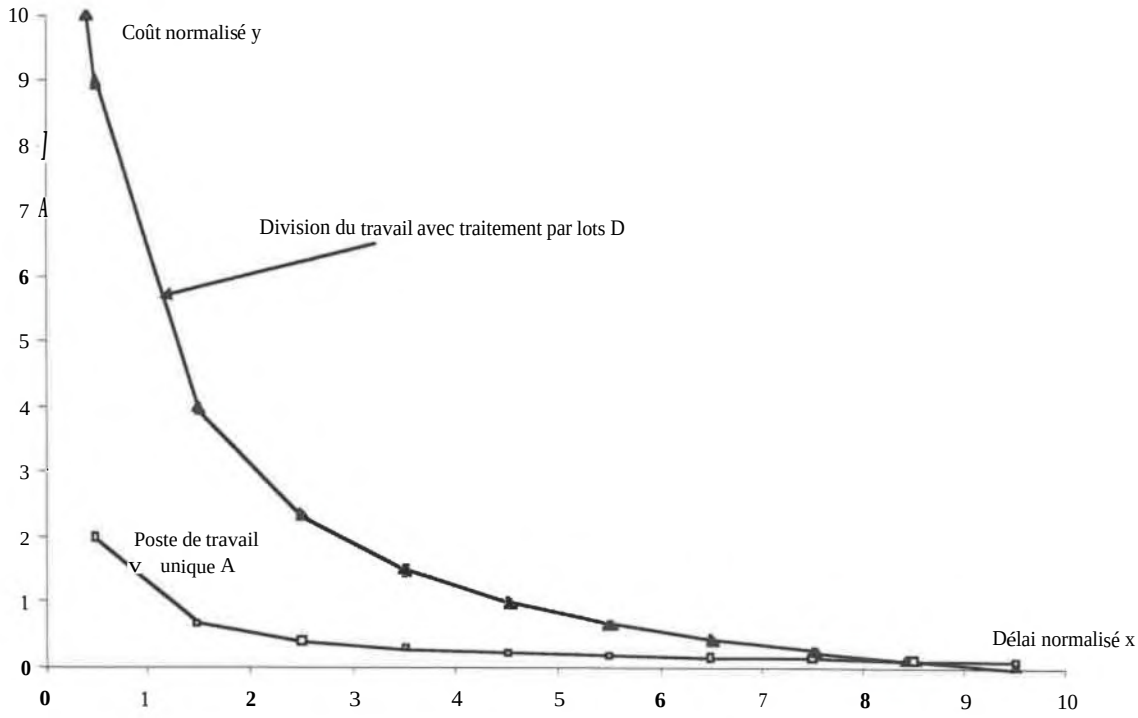

Figure 3 : délai et coût normalisés dans une division du travail avec traitement par lots (n=10, μ=1, T=1).

La figure 3 montre un exemple avec n = 10 bureaux d'une personne, une capacité de traitement p=1 de 1 cas par unité de temps, et une périodicité T = 1 d'une unité de temps. Pour un même délai, les coûts sont très supérieurs à ceux d'un serveur unique (A).

8.6 Division du travail avec transfert des dossiers au fil de l'eau

Les dossiers arrivent au hasard. Ils sont traités dès leur arrivée par une succession de "s" personnes affectées à temps plein à la procédure. Si les temps de travail sur un dossier sont identiques de poste à poste, il n'y a qu'une seule file d'attente, devant le premier poste. Les

dossiers sont traités immédiatement par les autres postes, si le temps de transfert est nul. Le délai d'achèvement du traitement des dossiers est donc le temps d'attente (hors traitement) devant un poste de travail de capacité "sp" auquel s'ajoute le temps de traitement "1/μ" par tous les postes de travail :

$$WT = X / s_{ji}(\varsigma n\text{-})_i) + 1/\eta \qquad [7]$$

$$C = s\,w\,/\,X \qquad [8]$$

Avec les variables normalisées la relation entre le coût et le délai s'écrit :

$$xy = 1\,/\,s \qquad [9]$$

La courbe de coût en fonction du délai est toujours inférieure à la courbe xy=1 du cas A (voir figure 4). La solution E est meilleure que la solution A. Cette dominance est d'autant plus forte que "s" est grand.

NB : Si les temps de travail des divers postes sont indépendants en termes de probabilités, il y a une file d'attente devant chaque poste. On retrouve la solution C.

9. Postes de travail traitant les dossiers de bout en bout avec une file d'attente commune (centre d'appels à un seul niveau)

Le travail arrivant est placé dans une file d'attente unique. Les "serveurs", au nombre de "s", les traitent dès qu'ils sont libres. Les dossiers ne sont donc pas préaffectés à un serveur. Le calcul du temps moyen d'achèvement WT tient compte du cas où les serveurs ne sont pas tous occupés (un nouveau dossier passe un temps 1/p dans le système) et du cas où ils sont tous occupés (le dossier attend alors dans une file). La formule est celle des files d'attente avec plusieurs serveurs (voir Churchman, 1961). Elle permet de tracer la figure 4.

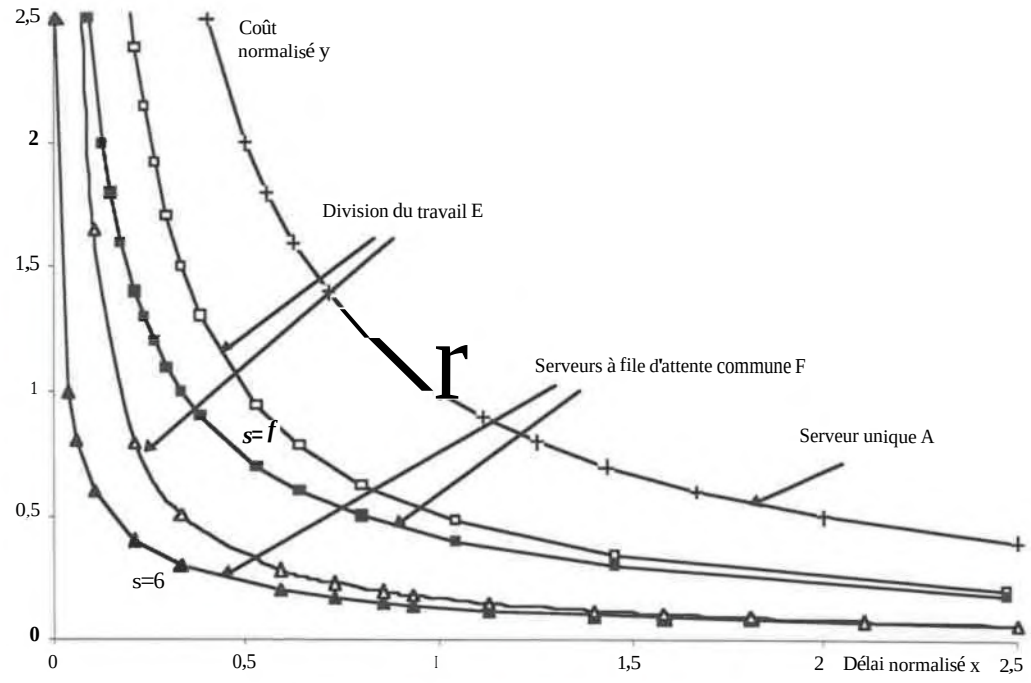

Figure 4 : relation entre le coût normalisé et le délai normalisé. Solution A (1 serveur, relation [3]), solution E (division du travail entre 2 et 6 personnes, relation) et solution F (2 et 6 serveurs).

On constate sur la figure 4 que la courbe de coût en fonction du délai pour la solution (F) est meilleure que celle de la division du travail avec bureaux spécialisés échangeant les dossiers au fil de l'eau (E), elle-même meilleure que la solution d'un seul serveur (A). Ces solutions sont d'autant meilleures que le flux de travail est suffisant pour affecter plus de personnes à la tâche (effet d'économie d'échelle).