Support Agile avec Kanban – quelques trucs et astuces

par Tomas Björkholm

Avant-propos

Il y a un an, j'ai animé un atelier au Scrum Gathering de Stockholm sur le Support Agile. Depuis, j'ai reçu plusieurs demandes pour détailler plus avant ce sujet, ce que je fais ici.

Il s'agit d'un mélange de mes expériences et d'idées fondées sur des discussions stimulantes avec mes collègues Henrik Kniberg et Mattias Skarin ainsi qu'avec David J. Anderson, l'un des pionniers de Kanban. Je vous invite à me contacter, si vous pratiquez ce que je décris ici, pour m'informer du résultat. Vous pouvez me joindre à l'adresse suivante tomas.bjorkholm@crisp.se

Je tiens à remercier Henrik Kniberg de m'avoir laissé utiliser ses images pour les flux. Je tiens également à remercier Reza Farhang de m'avoir fait ses retours suite à la lecture de ce document et Yassal Sundman pour avoir corrigé l'Anglais.

Sommaire

Principes

Puisque les activités de support et de maintenance semblent différentes dans chaque entreprise, respecter ces valeurs est plus important que respecter une recette. Une recette est ce que Tom Poppendieck – un pionnier dans le domaine du développement Lean de logiciel – décrit comme une solution pour résoudre le problème d'un autre. Je vous présente rapidement ici les principes et dans la suite de ce document vous pourrez en découvrir un peu plus. La plupart de ces principes sont basés sur Recipe for Success de David J. Anderson.

Concentrez vos efforts en limitant le travail à faire

Minimisez le passage d'une demande de support à une autre et limitez la possibilité d'en débuter de nouvelles avant de terminer celles déjà commencées. Ceci afin d'éviter de générer des files d'attente dans le processus. Collaborez et travaillez ensemble pour terminer les demandes commencées.

Équilibrez la demande et la capacité de travail

Évitez le stress car il ne vous aidera pas à améliorer votre efficacité. Si vous êtes stressé, vous êtes plus susceptibles de commettre des erreurs, ce qui vous rendra moins productifs. Ne vous engagez pas sur du travail que vous n'aurez pas la capacité de réaliser. Promettez plutôt un comportement et non un résultat.

Donnez de la visibilité et soyez transparent

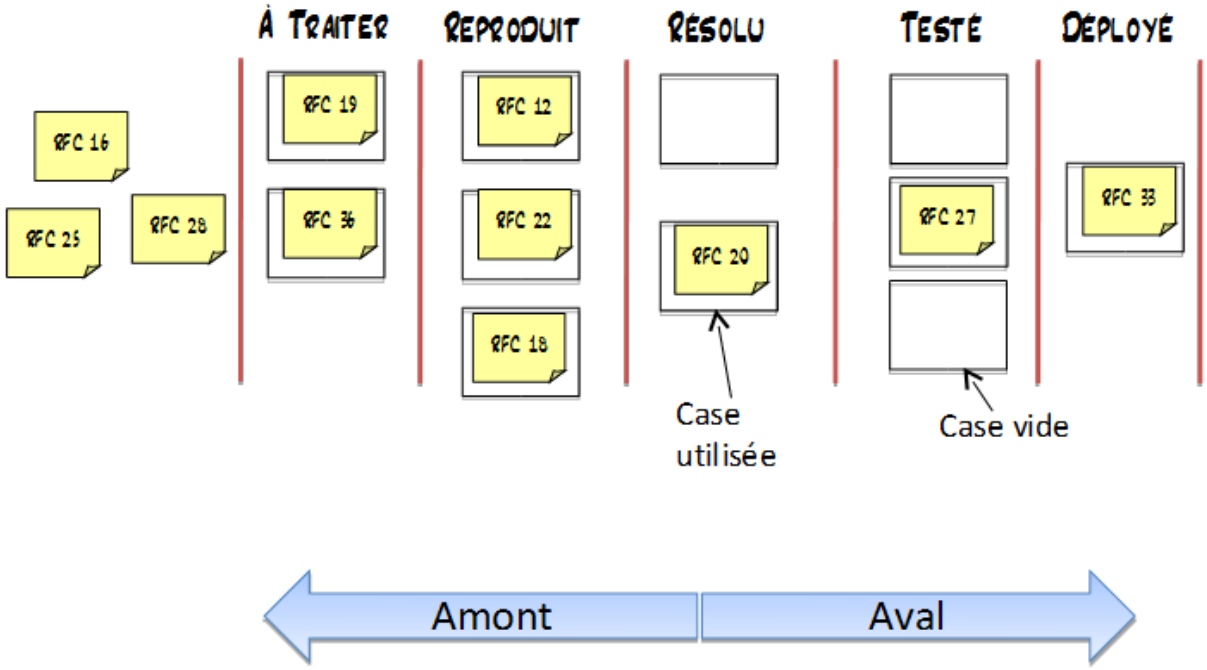
Donnez de la visibilité sur votre travail, vos succès ainsi que vos problèmes. Laissez le visible aux différentes parties prenantes. La principale raison pour visualiser le processus est d'être capable de l'améliorer, laissez les problèmes et les opportunités visibles.

Laissez votre Management assumer les priorités

J'entends trop souvent les équipes de support se plaindrent parce qu'elles ont un problème avec leurs clients (le plus souvent internes) contrariés parce que leur demande n'a pas été affecté d'une priorité suffisamment élevée. Cela détourne l'équipe de son objectif de traiter les demandes en cours. Ma suggestion est de décider de la façon d'établir les priorités – un algorithme – qui soit pleinement assumée par votre Management. Si quelqu'un n'est pas d'accord avec vos priorités alors il devrait se plaindre à votre Management et non à vous. Ceci pour vous permettre de travailler sur la résolution des problèmes au lieu d'argumenter sur quoi résoudre en premier.

Visualisez le flux de travail[1] sur un tableau

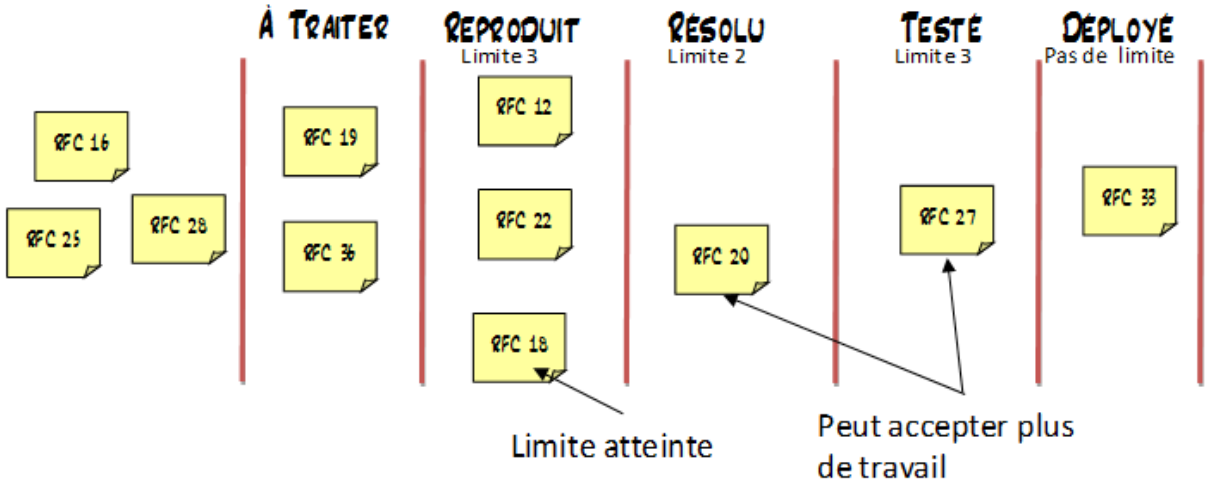
La première chose à faire est de visualiser le flux de travail sur un tableau, le travail à faire ainsi que le travail sur le point de démarrer. Le tableau vous permet de voir ce sur quoi vous travaillez et s'il y a des problèmes avec le flux de travail. Puisque le tableau a un nombre fixe de cases, il vous permet également de limiter la quantité de travail à faire (TAF[2] ~ WIP). J'ai aussi constaté une meilleure productivité juste parce qu'il est facile de voir sur le tableau ce qu'il faut faire ensuite.



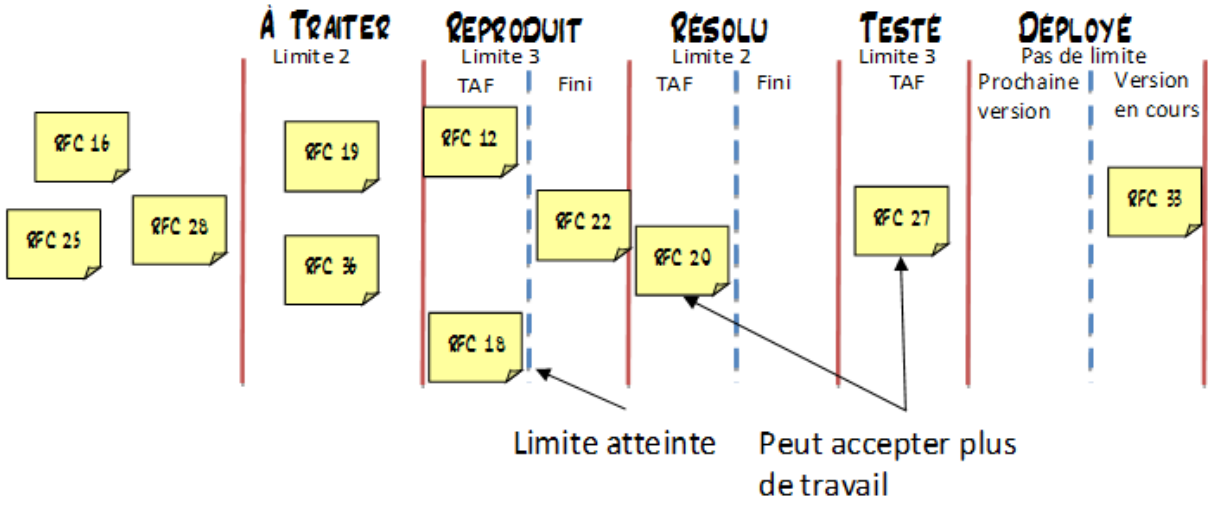
La règle de base du tableau est qu'une équipe ne peut prendre (~pull) une demande de support que s'il existe une case vide. Il se peut que vous ayez fini votre travail, mais il n'y pas de case vide. Cela se produit généralement lorsque l'équipe en aval n'a pas été en mesure de terminer ses demandes et n'est donc pas en mesure de prendre plus de travail venant de votre part. Cela signifie que vous avez un problème dans le flux de travail et au lieu de se contenter de démarrer de nouvelles choses, il vaut mieux que vous aidiez l'équipe avale à traiter le goulot d'étranglement. Une alternative serait d'augmenter votre limite, mais cela signifie que vous créez une file d'attente dans le système, ce qui nuit au temps de cycle. Le temps de cycle est le temps qu'il faut pour commencer à traiter une demande de support jusqu'à sa résolution.

1 NdT. Flux de travail ou Workflow, c'est vous qui voyez...
2 NdT. Traduction de "**W**ork **I**n Progress" par "**T**ravail **A** **F**aire"proposée par Claude Aubry.

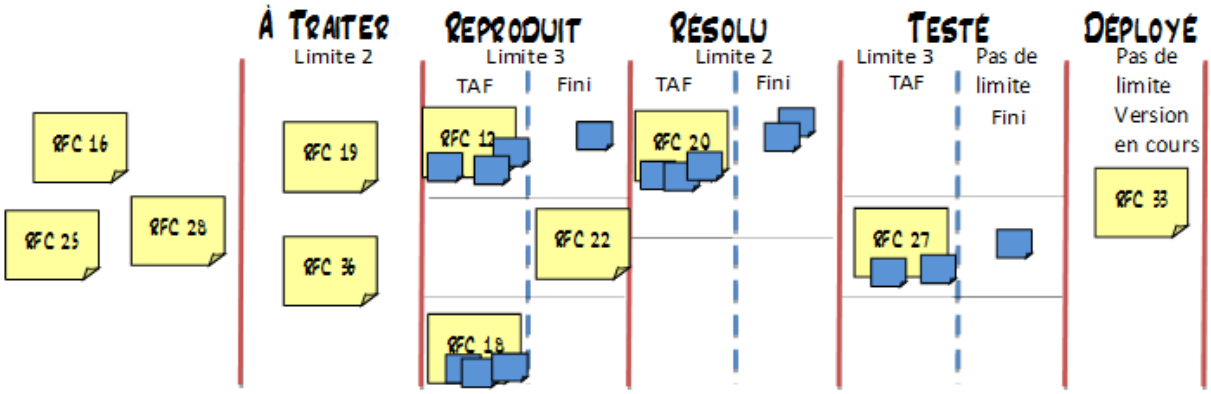
Autres exemples de tableau



Au lieu d'utiliser des cases, vous pouvez simplement inscrire la limite dans chaque colonne.



Pour que ce soit plus facile pour l'équipe avale de voir votre progression, vous pouvez scinder votre colonne en deux avec une colonne pour le TAF et une colonne pour les demandes que vous avez terminées. Votre limite reste valable pour les deux colonnes.



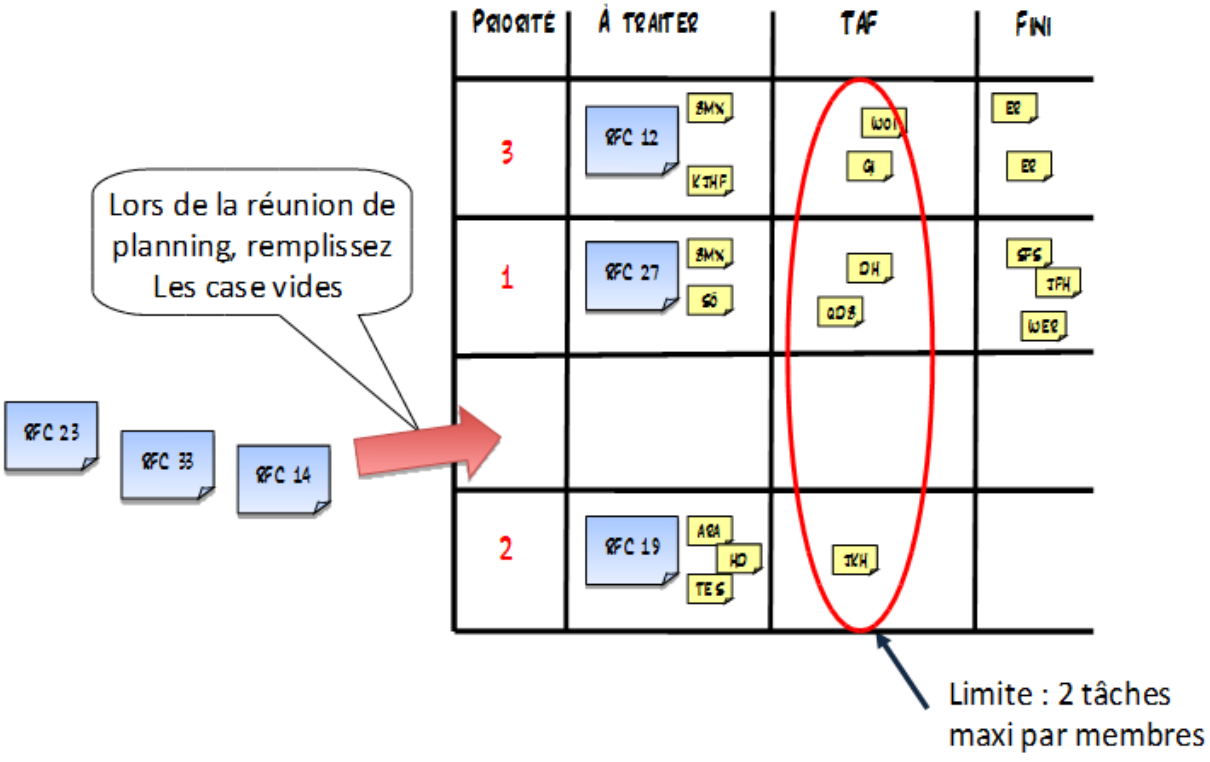
Si c'est plus simple pour vous de décomposer les demandes de support en tâches (activités techniques nécessaires pour résoudre la demande de support), faites-le. Pour qu'il soit facile de voir quelles sont les tâches qui appartiennent à une demande de support, vous pouvez ajouter des ”couloirs de nage” (~swimlanes).

Tableau pour une équipe de support multifonctionnelles

Les tableaux ci-dessus sont découpés en différentes colonnes afin de montrer le travail des différentes équipes. Si les demandes de support sont prises en charge par une seule équipe, sans différenciation nette entre qui fait quoi, on qualifie cette équipe de multifonctionnelles (~cross-functional). Si c'est le cas, ils peuvent utiliser un tableau avec uniquement trois colonnes : "À traiter", "TAF" et "Fini". Cela ressemble beaucoup à un tableau Scrum.

La différence avec Scrum est que le nombre de lignes sur un tableau Scrum est limité au nombre d'éléments du Backlog sur lesquels l'équipe s'engage dans le sprint en cours. Ici, la limite est un nombre prédéfini de demandes de support pouvant être traitées en même temps. Alors que Scrum travaille par itération, ici vous travaillez sur un flux permanent de demandes de support.

Pour une équipe multifonctionnelle, le tableau – avec des demandes de support qui peuvent être décomposées en tâches techniques – peut ressembler à ceci.



Pour minimiser le risque de surcharge ou d'éparpillement, il existe deux limites. Une pour l'équipe, qui se situe au niveau de la demande de support, et une pour chaque membre de l'équipe, au niveau de la tâche. Dans l'exemple ci-dessus, on a au maximum quatre demandes de support en cours, et chaque membre ne peut travailler que sur

deux tâches à la fois. J'ai vu des équipes utilisant des aimants, avec leurs noms dessus, pour marquer le travail en cours. Il s'agit d'une très bonne idée pour montrer qui fait quoi et pour s'assurer que les limites sont respectées.

Afin de minimiser la surcharge de travail inutile, une demande de support remplace celle qui est tout juste terminée et les demandes ne sont pas déplacées pour refléter les priorités. Par contre, une colonne a été ajoutée pour montrer les priorités.

Comment concilier les activités urgentes avec les activités quotidiennes et préventives

Certains journées de travail des équipes de support sont un mélange d'activités quotidiennes, d'activités préventives et de correctifs urgents. Il est possible de les gérer à partir d'un tableau de ce type.

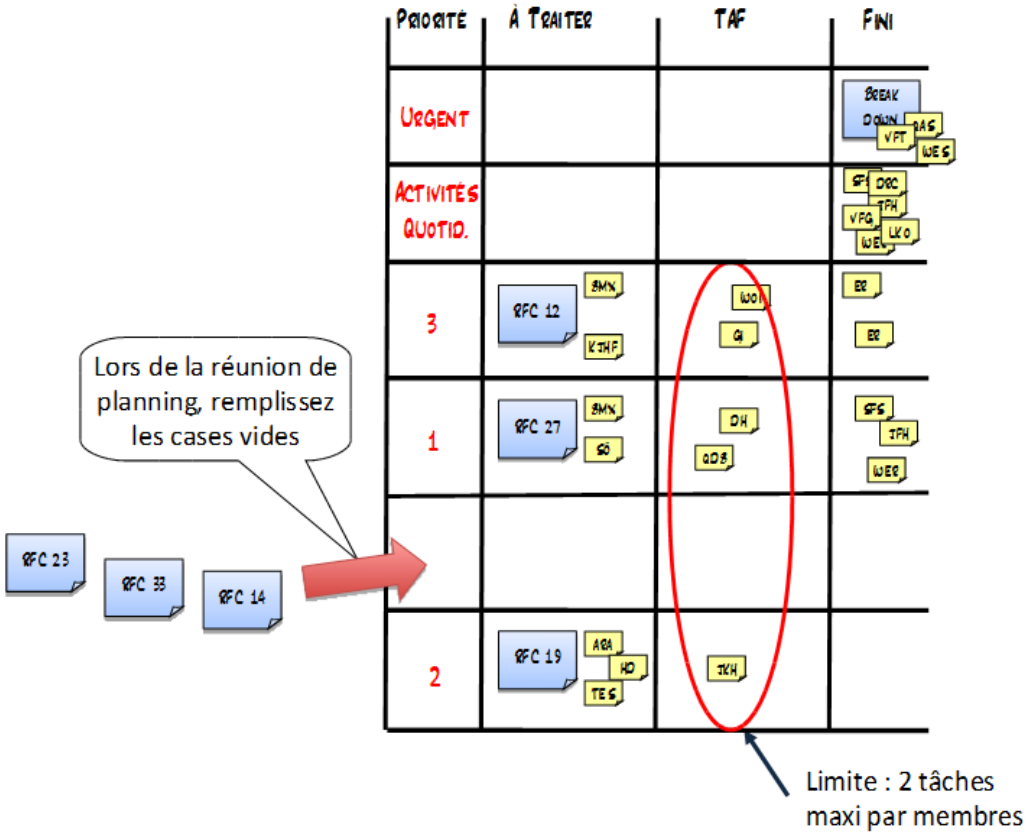


Tableau du flux de travail de l'équipe combinant des activités urgentes, des activités quotidiennes et le préventif

La journée commence par le déplacement des tâches quotidiennes dans la colonne "À traiter". Les tâches quotidiennes sont traitées une par une jusqu'à ce qu'elles soient toutes finies. Après cela, les membres de l'équipe peuvent commencer à travailler sur le préventif, dans cet exemple limité à quatre demandes. Si un problème urgent apparaît, les autres tâches sont interrompues et l'équipe intervient sur l'urgence. Lorsque le problème urgent est résolu, le travail revient à la normale. Une bonne règle pour les problèmes urgents qui ne peuvent pas attendre, c'est de commencer à les traiter dès que vous avez terminé avec la tâche que vous avez commencée. Si le problème est un peu

moins urgent, peut-être peut-il attendre qu'il y ait une ligne vide, c'est-à-dire lorsque vous avez fini de traiter la majorité du préventif. Vous et votre équipe pouvez décider de modifier l'ordre de priorité des activités préventives à tout moment.

Pourquoi respecter les limites

Lorsque je parle de Kanban et des limites de TAF, j'ai l'habitude qu'on me demande pourquoi on ne peut pas transgresser les limites. Ne serait-il pas mieux de continuer à travailler tant qu'il y a quelque chose à faire ?

Pour répondre à cette question, je montre d'abord souvent ce qui se passe si vous dépassez votre limite et continuez à travailler.

Dans ce qui suit, je vous présente une organisation nécessitant cinq équipes pour traiter une demande de support émanant d'un client jusqu'à ce que la solution soit déployée. Les demandes de support prises en compte par le système sont représentées avec de l'eau et le système est représenté avec un tuyau. La largeur des tuyaux représente la capacité de l'équipe concernée.

Dans le premier schéma, le nombre de demandes de support prises en compte reste inférieur à la capacité que chaque équipe peut gérer. Cependant l'équipe 4 (test) travaille à pleine capacité alors que les autres équipes sont en sous-capacité. Les collaborateurs de ces équipes sont inactifs une partie du temps. Comme il n'y a pas de files d'attente dans le système, le débit est rapide.

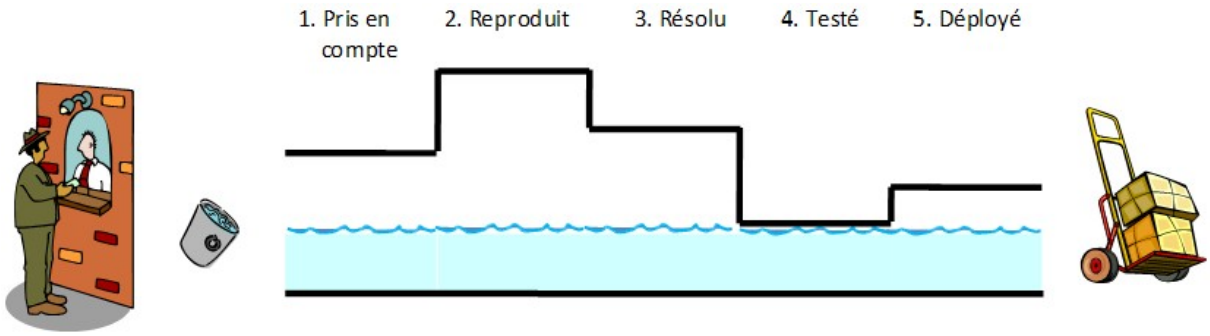


Schéma 1 : le flux de travail reste inférieur au goulot d'étranglement.

Si le nombre de demandes prises en compte est augmentée jusqu'au maximum de la capacité de la première équipe (pris en compte et priorisé), alors on a un goulot d'étranglement sur l'équipe 4 en surcapacité et une file d'attente est générée. Le résultat en sortie reste le même puisque limité par le goulot d'étranglement.

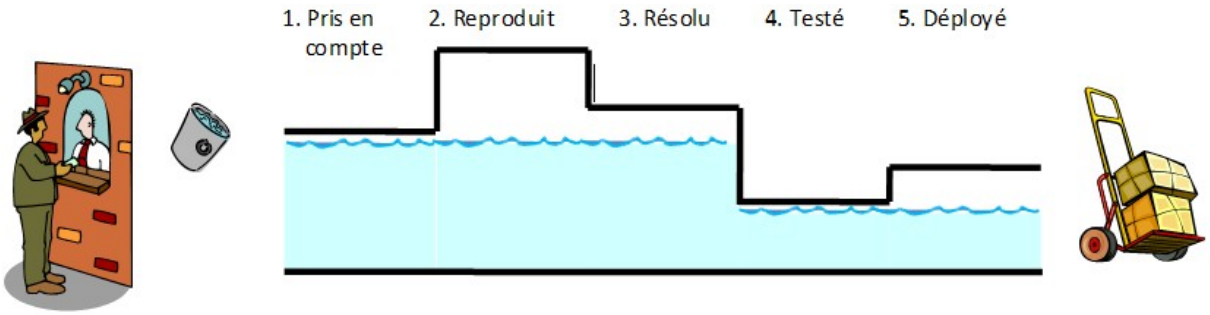


Schéma 2 : le flux de travail est supérieur à la capacité du goulot d'étranglement.

Si vous continuez à prendre des demandes de support jusqu'au maximum de la capacité de la première équipe, vous allez générer des files d'attente partout dans le système. Le résultat sera toujours le même, mais le débit sera plus lent.

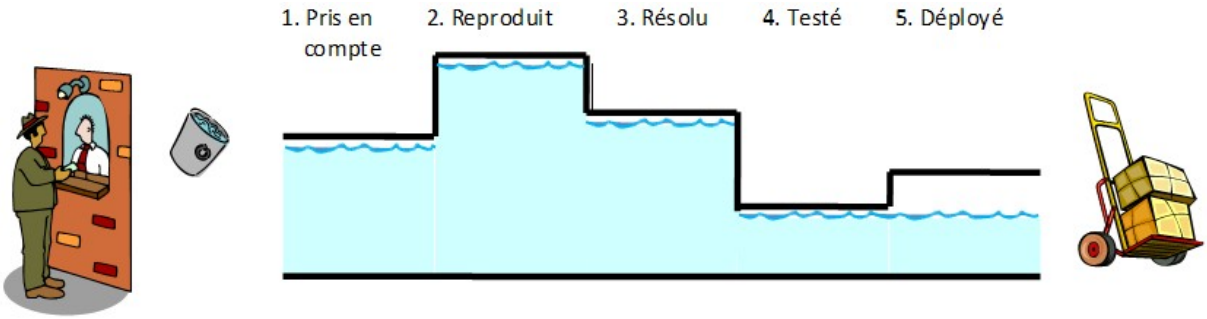


Schéma 3 : les demandes de support sont prises en compte tant qu'il y a de la capacité dans la première équipe et ceci sans s'inquiéter de savoir si l'équipe suivante est prête ou non à commencer à travailler sur une demande de support.

Conclusion : si vous continuez à travailler à votre propre capacité sans tenir compte de la capacité des autres équipes, vous allez probablement générer des files d'attente pour toutes les équipes en amont du goulot d'étranglement. Toutes les équipes avec une file d'attente en entrée vont penser qu'elles ont un problème de capacité alors que vous êtes réellement en train de masquer le vrai problème. Le point important de cet exemple est que votre client aurait été plus heureux si toutes les équipes avaient respecté leur limite. Même si cela signifie que certaines personnes ne font rien pendant une partie importante de leur journée de travail. Le résultat aurait été le même mais le débit aurait été plus élevé.

Le problème avec un débit faible est que les clients pourraient vous quitter parce que vous êtes trop lent à réagir. Peut-être même que les demandes de support sont obsolètes au moment d'être déployées. Peut-être même qu'il aurait été préférable que ces demandes de support n'aient jamais été prises en compte par votre système, puisque vous investissez à perte.

Si vous respectez vos limites, vous verrez facilement où se situe les goulots d'étranglement dans le système. Lorsque le problème est visible, il est plus facile de le résoudre et aussi de vérifier que cela a effectivement eu un effet positif.

Décidez de la limite

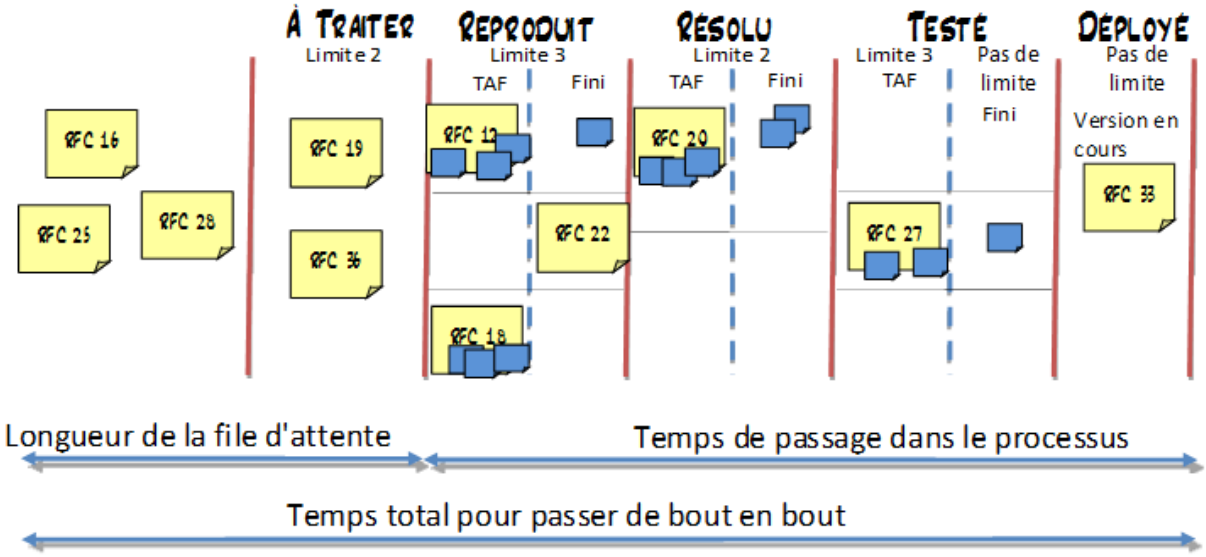
Pour trouver la bonne limite, il vaut mieux essayer et mesurer ensuite le résultat. Si vous voulez améliorer le temps de cycle, vous pouvez essayer de baisser la limite. Il est évident que cela prend plus de temps pour terminer le processus si vous avez dix cases dans le système plutôt que cinq. Une limite trop basse peut entraîner des coûts plus élevés puisque les gens ne font rien en attendant que l'équipe amont ou avale soit prête. Il arrive que les gens aident d'autres équipes même s'ils ne sont pas experts dans le domaine et donc pas très rapides. Des limites faibles signifient que vous avez un débit élevé, mais vous n'êtes pas protégé contre les variations de taille des demandes de support. Les goulots d'étranglement apparaîtront plus souvent.

Métriques

Pour prendre les bonnes décisions et les suivre pour voir si elles donnent de bons résultats, vous avez besoin de mesurer certains indicateurs de performance importants. Les indicateurs que je préconise sont le délai de livraison[3] (~lead time), la vélocité, la qualité et le flux de travail. Pour en savoir plus sur les décisions prises à partir de ces métriques, lisez le chapitre "Variation et taille de la file d'attente".

Délai de livraison

Il existe deux moyens importants pour mesurer le délai de livraison. Le premier est le temps total nécessaire en moyenne pour faire passer la demande de support à travers tout le système, à partir du moment où elle entre jusqu'à ce qu'elle soit terminée. Le second est le temps consommé par une personne sur le sujet. Le second moyen est intéressant pour calculer le retour sur investissement (~ROI) car vous commencez à investir lorsque vous commencez à travailler sur la demande de support.



3 NdT. ***Lead time*** = Délai de livraison. Point de vue Client : donc entre le moment de sa demande et le moment où la solution est déployée. ***Cycle time*** = Temps de cycle. Point de vue Système : donc entre le moment de la prise en compte de la demande et le moment où la solution est prête à être déployée.

Vélocité

Connaissez-vous votre production/niveau de résolution au bout de, disons, une semaine ? Vous pouvez compter les demandes de support ou si vous voulez, vous pouvez donner une valeur de 3 pour des demandes de complexité forte, de 2 pour une complexité moyenne et de 1 pour une complexité faible. La somme totale est votre vélocité.

Qualité

Quelle partie de votre travail ajoute de la valeur ? Les activités positives sont celles qui rendent le client heureux ainsi que le préventif. Les activités négatives sont les demandes urgentes ainsi qu'aider un client de nouveau mécontent parce que le dernier correctif ne l'a pas aidé. Pour obtenir des métriques de bonne qualité, vous pouvez diviser la quantité de valeur ajoutée (activités positives) par la quantité de travail totale (activités positives + activités négatives).

Flux de travail

Pour savoir si votre capacité est correcte, vous pouvez mesurer le nombre de fois où votre flux de travail est bloqué. Cela arrive quand il n'y a pas d'éléments terminés en amont ou que l'équipe avale est bloquée.

Comment décider de la suite

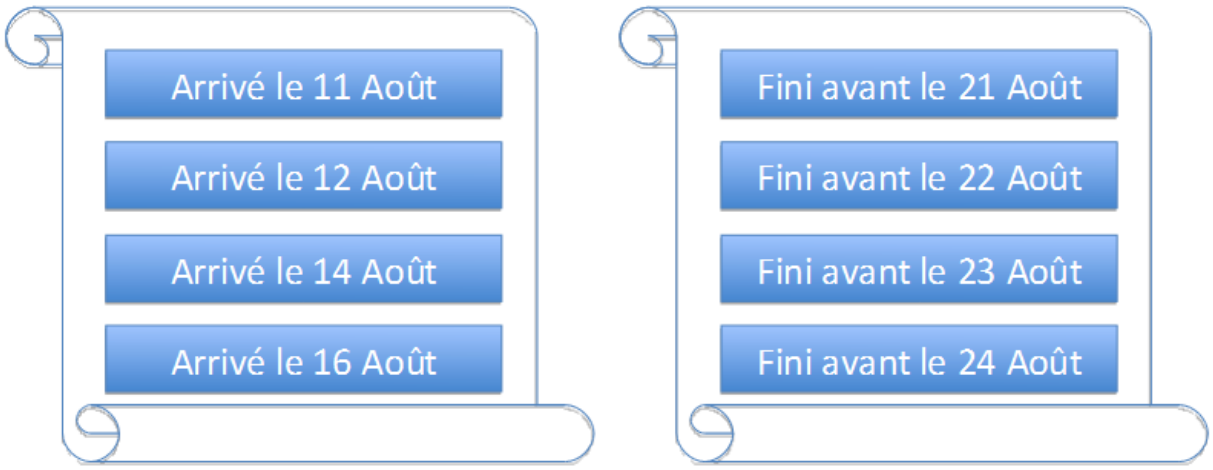
Il existe plusieurs moyens pour décider quelle demande de support traiter en priorité par le système quand il y a des cases vides dans la première colonne. Suivent quelques suggestions. Choisissez ou combinez celles qui conviennent le mieux à votre situation.

FIFO

Le moyen le plus facile de prioriser est de systématiquement choisir la demande de support la plus ancienne en utilisant l'algorithme premier-entré-premier-sorti (~FIFO[4]).

SLA

Si vous avez un contrat de niveau de service (~SLA[5]) avec des temps de réponse garantis, vous devez prioriser les demandes de support en fonction de leur date d'échéance.



Deux exemples de files d'attente : celle de gauche priorisée selon l'algorithme FIFO et celle de droite priorisée selon les engagements sur les temps de réponse.

Picorer le dessus du panier[6]

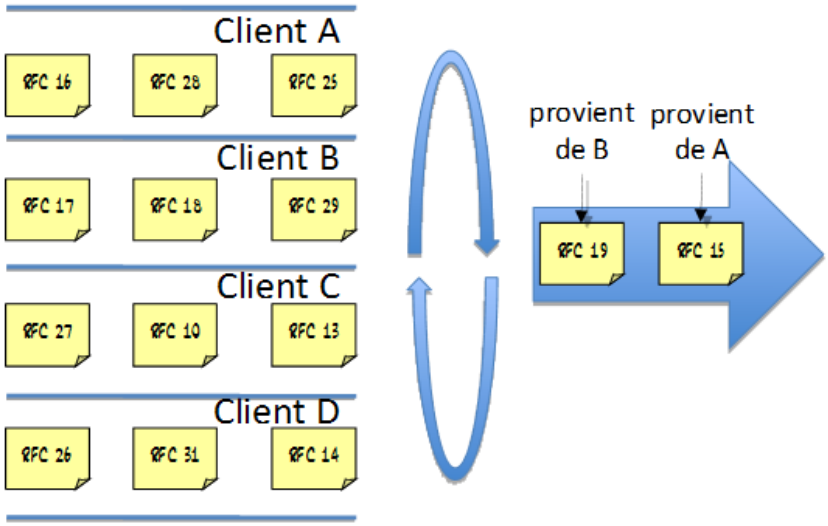
Afin de rapidement réduire le nombre de demandes de support dans la file d'attente vous pouvez choisir les plus faciles à traiter en premier. Cela signifie que le nombre de demandes traitées sera élevé, mais il y aura quelques demandes difficiles qui vont traîner un moment dans votre système. Ces dernières vont diminuer votre délai de livraison moyen (~average lead time).

Tourniquet[7]

Avez-vous un nombre limité de clients ? par exemple, gérez-vous le support d'un certain nombre de bureaux de vente répartis dans différents pays ou régions ? Si c'est le cas, ils peuvent avoir leur propre file d'attente et avoir la responsabilité de la prioriser. Vous

4 NdT. First In First Out
5 NdT. Service Level Agreement.
6 NdT. Cherry Picking.
7 NdT. Round Robin.

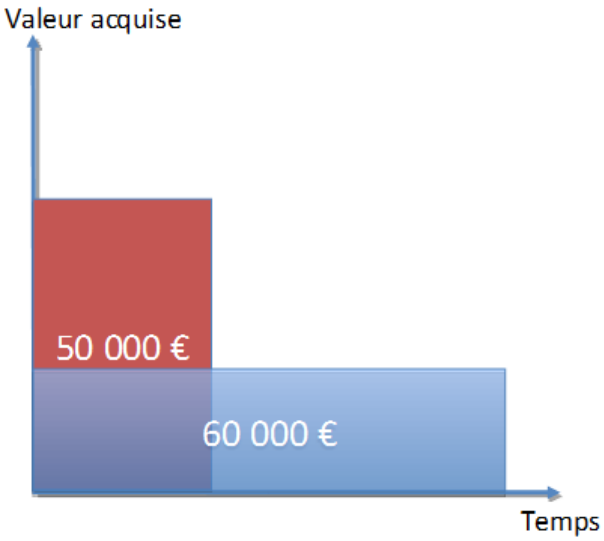
pouvez alors alterner entre les files d'attente, une par une, et choisir la demande avec la plus haute priorité. Ceci est connu comme l'algorithme du Tourniquet.



Exemple de files d'attente priorisées par les clients :
chaque demande est sélectionnée selon l'algorithme du tourniquet

Valeur

Un algorithme très intéressant à utiliser est l'optimisaton par la valeur. La demande de plus grande valeur en premier. Mais il est également intéressant de savoir comment la valeur est répartie dans le temps. Certaines demandes pourraient avoir une valeur élevée pendant un court laps de temps et devrait donc avoir une priorité plus élevée qu'une demande de valeur égale mais répartie sur une période plus longue. Il est également parfois préférable de prioriser les demandes dont la valeur augmente sans cesse et qui disposent d'une date de fin.



Prioriser une demande présentant une grande valeur ajoutée à court terme
plutôt qu'une demande présentant une grande valeur ajoutée sur du plus long terme.

Piloté par le Management

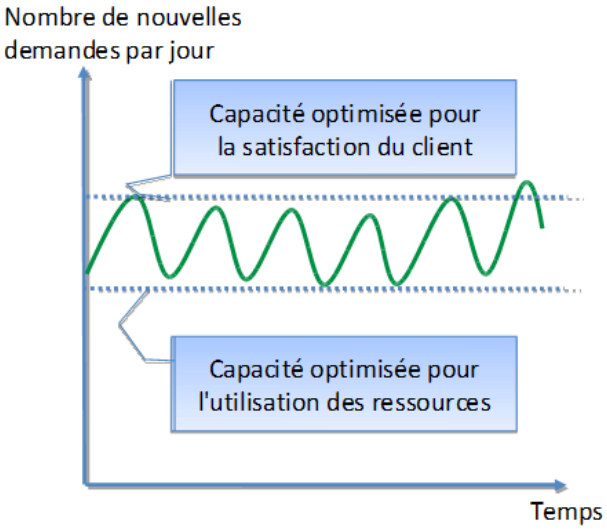
Si un manager vous aide à décider l'ordre des demandes à traiter en suivant, il est bon d'avoir une file d'attente pour ces demandes afin de savoir laquelle prendre la prochaine fois qu'une case est vide. Ceci n'est bien sûr pas nécessaire si la file d'attente des demandes est déjà priorisée.

Lisser la charge de travail

Si une demande de support exige beaucoup de travail, disons de la part de l'équipe qui reproduit les anomalies, elle arrêtera le flux et laissera les équipes avales sans travail. Si la limite pour "Reproduit" est supérieure à 1, des demandes faciles et rapidement terminées peuvent être traitées en parallèle pour assurer du travail aux équipes avales.

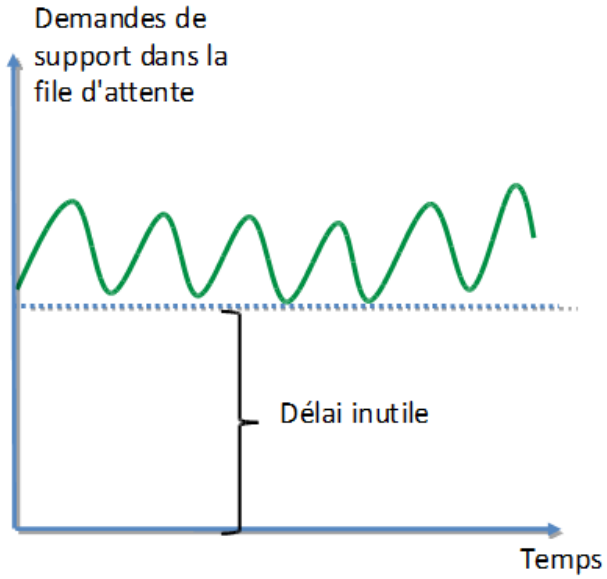
Variation et taille de la file d'attente

Sauf si vous êtes prêt à payer pour une capacité inutilisée, il y aura forcément une file d'attente en entrée de votre processus de support. Ce qui est naturel lorsque l'on traite des entrées générées par des besoins de support.

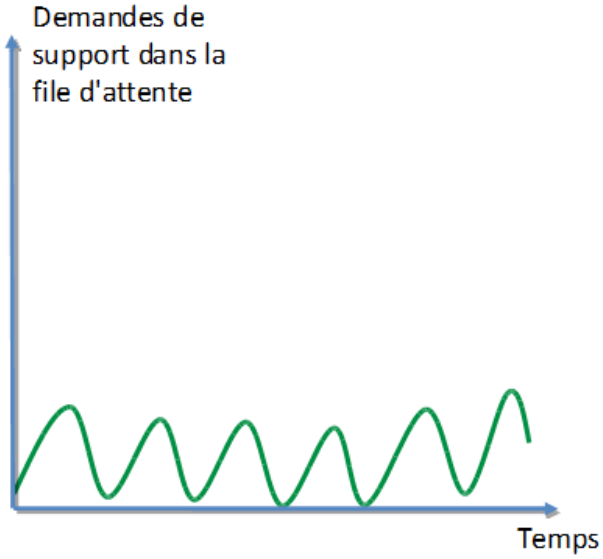


Le nombre de nouvelles demandes de support varie dans le temps.

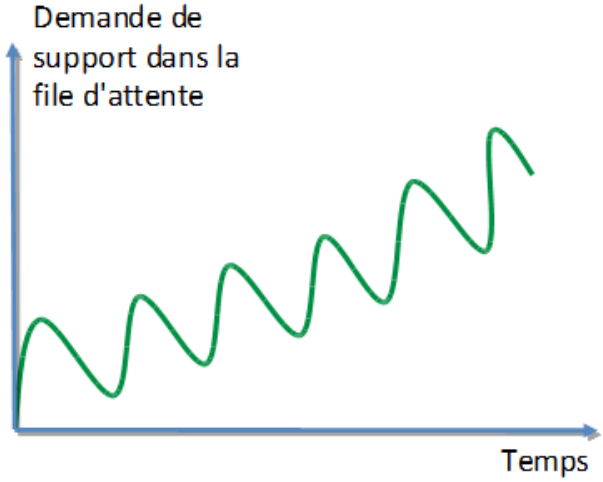
Si vous souhaitez optimiser l'utilisation des ressources alors vous devriez dimensionner votre équipe de telle façon que la file d'attente des demandes de support n'atteignent jamais 0 plus de quelques minutes. Dans le même temps, ne perdez pas de temps à essayer d'avoir une file d'attente qui n'atteignent jamais 1. La distance entre le minimum et 0 n'ajoute pas de valeur. Elle n'ajoute que de l'agacement.



Si vous souhaitez optimiser la satisfaction du client, alors vous devriez dimensionner votre équipe pour qu'elle ait toujours la capacité de commencer immédiatement à traiter les demandes de support qui arrivent. Autrement dit, la file d'attente est toujours à 0.



La capacité qui vous convient se situera probablement quelque part entre les deux. Vous pouvez dimensionner votre équipe pour que vos clients privilégiés (VIP) obtiennent une réponse immédiate pour 50% de leurs demandes, tandis que les autres clients obtiennent une réponse immédiate pour 5% de leurs demandes.



En gardant la trace du nombre de demandes dans la file d'attente, vous pouvez déterminer si votre équipe est correctement dimensionnée. Dans le cas ci-dessus, il est évident que la capacité est trop faible.

Amélioration continue

Un bonne boucle de feedback est plus importante qu'un bon départ. Mes trucs et astuces présentés ici ne décrivent pas un processus de support parfait. Vous devez personnaliser le processus pour répondre aux besoins de votre organisation. Étant donné que les circonstances changent, aucun processus ne reste parfait au cours du temps donc vous devez l'améliorer constamment. Ma recommandation est d'avoir une réunion, au minimum mensuelle, pour parler des problèmes et des améliorations.

Bonne chance !

/ Tomas Björkholm