Rapport de stage

Génération de modèles calibrés automatiquement

Projet encadré par Bruno Baynat, Alexandre Brandwajn
et Serge Fdida

Auteur : Thomas Begin

Master II Réseaux Parcours Recherche

Août 2005

Remerciements

Je tiens à remercier Serge Fdida, Alexandre Brandwajn et Bruno Baynat pour leur disponibilité, leurs conseils et leur bienveillance au cours des cinq mois de mon stage au sein de l'équipe Réseaux et Performances du LIP6. Je tiens par ailleurs également à remercier l'ensemble de l'équipe R&P pour leur accueil et leur soutien.

Table des matières

Introduction

Modalités administratives Le sujet de génération de modèles calibrés automatiquement a été proposé par M. Fdida aux étudiants de Master II options Réseaux qui souhaitaient réaliser un stage de fin d'étude à forte coloration recherche dans la thématique des performances. Le stage a démarré mi-avril et se terminera fin août 2005. Durant ces 5 mois j'ai intégré pleinement l'équipe R&P du LIP6 et pris part aux réunions d'équipe.

Cadre de travail Quatre participants ont principalement contribué aux activités menées. En la qualité d'encadrant, Alexandre Brandwajn (professeur et chercheur à *University of California, Santa Cruz*), Bruno Baynat (maître de conférences à l'*UPMC* et chercheur au *LIP6*) et Serge Fdida (professeur à l'*UPMC* et chef de l'équipe R&P du *LIP6*) et moi-même en la qualité de stagiaire. Nous avons eu de nombreuses réunions de travail qui ont permi de recadrer les travaux, d'orienter les travaux futurs et de soumettre à notre critique les avancées réalisées.

Génèse du sujet L'idée de ce sujet est née de la dépréciation progressive des modélisations analytiques en faveur des techniques de résolution par simulations pour pronostiquer les performances d'un système informatique lorsqu'il n'est pas possible d'y faire des mesures. Le socle conceptuel est de tenter d'aborder le processus de modélisation sous un nouvel angle et de voir si les perspectives offertes permettent d'obtenir des résultats exploitables.

Formulation simplifiée du sujet Le sujet vise à échauder une méthodologie systématique de génération de modèles calibrés à l'égard de systèmes peu documentés ou inaptes à être modélisés. Plus précisément, notre ambition consiste à proposer une modélisation à la fois simple et fidèle au comportement étudié d'un système informatique en se basant essentiellement sur un ensemble de mesures observé sur le système considéré.

Présentation des résultats Les avancées réalisées et les résultats obtenus au cours de ce stage sont présentés à travers deux médias.

- Par écrit, dans ce rapport qui contient l'ensemble des résultats obtenus au gré du cheminent de nos avancées et a pour ambition de délivrer une vue à la fois précise et détaillée sur l'ensemble de nos travaux réalisés (et à venir).
- Par oral, lors d'une soutenance le 7 septembre 2005 devant un jury composé de M. Timur Friedman, M. Benoît Donnet, M. Jérémie Leguay, M. Franck Legendre, M. Bruno Baynat et M. Serge Fdida.

Enoncé du plan La première partie de ce rapport est consacrée à des notions générales telles les enjeux liés à l'évaluation des performances, les mécanismes de modélisation et de résolution de modèles. Dans la deuxième partie sont présentées sous un angle comparatif les différenciations entre le cheminement classique de modélisation et l'approche que nous souhaitons développer. Cet objectif s'ensuit naturellement de la troisième partie qui a pour but de présenter

les concepts, les outils, les mécanismes clés nécessaires ainsi que la logique et l'intuition sous-jacentes à nos travaux. La quatrième partie propose un descriptif synthétique des connaissances requises pour chacun des types de modèles impliqués dans nos travaux et explicite également l'adaptation de la logique générique pour chacun de ces modèles. Cette analyse détaillée des modèles précède la partie réservée aux résultats obtenus au gré de nos expériences (rassemblés dans la cinquième partie). La sixième partie approfondit certains détails techniques indispensables pour mettre en oeuvre la méthodologie exposée. Quant à la dernière partie, elle aborde les travaux futurs et les améliorations à apporter à l'existant.

Première partie

Notions générales sur la modélisation des systèmes

Pour plus de détails sur cette partie, nous recommandons la lecture de [1], un article dont le contenu permettra au lecteur d'obtenir des informations plus précises et plus générales sur le sujet abordé dans cette partie.

1 Les modèles

Un modèle a vocation à reproduire le comportement d'un système (existant ou pas) dans le but d'obtenir des résultats relatifs à l'usage du système considéré. Les paramètres de performance et les propriétés structurelles font partie de ces résultats. Une modélisation complète d'un système comprend d'une part la reproduction du système et d'autre part la reproduction de la charge de travail à laquelle est soumis le système.

2 Les paramètres de performances

Un système Σ, soumis à un ensemble de paramètres d'entrées (matérialisés par les clients de la charge de travail) réagit (ou interagit) et délivre en sortie des paramètres, les paramètres de performance. Il est possible que les paramètres de sortie influent sur l'entrée du système ce qui correspond à un asservissement du système. Naturellement les performances d'un système, s'il existe, peuvent être directement mesurées mais il est également possible de les obtenir à l'aide d'un processus de modélisation. Les métriques de performance généralement adoptées peuvent être approximativement perçues comme un sous-ensemble, ou comme une combinaison mathématique des paramètres de sorties d'un système. On définti les clients d'un modèle comme étant les entités auxquelles on s'intéresse. On pourra par exemple étudier le comportement des requêtes dans un serveur, d'un paquet IP dans un réseau, ...

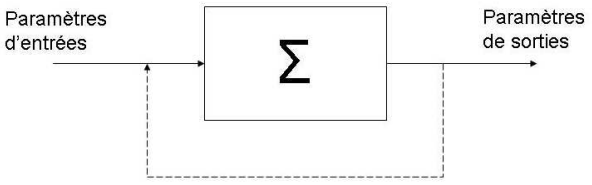


Fig. 1 – Représentation symbolique d'un système

Les grandeurs caractéristiques les plus courantes parmi les nombreuses existantes pour quantifier les performances d'un système sont :

- Le débit qui correspond à la quantité de clients qui s'écoule par unité de temps
- Le temps de séjour des clients dans le système

- Le nombre de clients dans le système
- Les probabilités de rejet (dans les systèmes à perte)
- Le taux d'utilisation d'une ressource dans le système
- ...

Notons que l'exercice d'évaluation des performances vise à estimer les valeurs des métriques telles celles précédemment citées. Typiquement il pourra s'agir de calculer l'espérance (i.e. la valeur moyenne), l'écart-type, la valeur maximale du débit ou bien du temps de réponse.

3 La définition d'une charge de travail

La charge de travail joue un rôle capital dans le processus de modélisation d'un système. Elle rassemble toutes les demandes de traitement destinées au système considéré. Il peut s'agir de programmes, de données, de commandes, de trafics IP (observés à l'échelle des paquets ou bien des flots) selon le secteur dont est issu le système.
Or il est des systèmes pour lesquels la caractérisation de la charge de travail ou bien sa modélisation est très difficile à établir. Ce manque de précision sur la charge de travail peut être très préjudiciable (dans l'évaluation des performances) tant la dépendance entre les performances d'un modèle et sa charge de travail est étroite. En effet les performances d'un système découlent directement de la charge de travail appliquée en entrée du système.
Par conséquent, toute évaluation de performance rigoureuse, que ce soit par les mesures ou par une modélisation, devrait indiquer systématiquement la charge de travail avec laquelle l'évaluation a été menée (malheureusement cela n'est pas toujours possible).

4 Les buts et les usages

L'analyse qualitative et l'analyse quantitative

On distingue deux types d'évaluations possibles aux fonctionnalités et aux méthodes très différentes. L'objet d'une modélisation peut consister à étudier qualitativement ou bien quantitativement un système. S'il s'agit d'une étude qualitative, l'analyse visera à définir les propriétés structurelles et comportementales du système (telles que l'absence de blocage, les invariants du système, le comportement fini ou borné). Tandis que dans une étude quantitative, c'est la recherche des performances du système (à travers les paramètres de performance) qui motive l'analyse. Rigoureusement, cette dernière n'a de sens que si une analyse qualitative a été préalablement menée. Il est par exemple inutile de vouloir obtenir les performances d'un système qui se trouve dans un état de blocage.
Notre projet de génération de modèles calibrés automatiquement s'appuie sur un alliage des approches quantitatives et d'analyse quantitative. Comme nous le verrons ultérieurement, l'analyse qualitative permet de définir et de circonscrire l'ensemble des modèles candidats (en imposant des contraintes structurelles sur les modèles) tandis que l'analyse quantitative aura pour but de les départager en se fiant au comportement des modèles dans leur état stationnaire (par opposition au régime transitoire).

L'exploitation d'un modèle

En règle générale, les desseins qui occasionnent un processus de modélisation sont à coloration prédicitive. Ainsi l'exploitation d'un modèle peut être menée pour des raisons :

- De dimensionnement. C'est-à-dire en vue de concevoir des systèmes conformes à un cahier des charges ou bien de calibrer favorablement les paramètres clés d'un système.
- De planification de capacité (« capacity planning »). C'est-à-dire en vue de prédire le comportement d'un système existant en dehors de son point de fonctionnement normal. Concrètement, il s'agit de déterminer les besoins à venir en ressources d'un système en se basant sur le taux d'utilisation actuel des ressources et sur une hypothèse de croissance pour la charge de travail. Par conséquent, l'opération de planification de capacité requiert une méthode de prédiction de la charge prévisionnelle.

Enfin précisons que lorsque le système existe, la façon la plus simple d'évaluer les performances d'un système (dans une zone de fonctionnement normal) est d'employer les méthodes de mesures directes.

5 Les outils de modélisation

Selon la nature de l'analyse (qualitative ou quantitative), les outils à la disposition des chercheurs/ingénieurs varient. On recense très succinctement parmi ces formalismes :

- Les Chaînes de Markov à temps discret et à temps continu et les réseaux de files d'attente regroupés dans la théorie des files d'attente (à but quantitatif)
- Les réseaux de Pétri (à but essentiellement qualitatif)
- Les modèles fluides (à but essentiellement quantitatif)
- ...

Notre travail fait appel à la théorie des files d'attente, théorie largement éprouvée dans de nombreux et variés domaines, pour calculer les performances des modèles à l'étude.

Notons que la grande majorité des modèles issus de la théorie des files d'attente sont de nature stochastique. La notion de caractérisation stochastique des systèmes résulte en fait d'une solution astucieuse des scientifiques pour refléter de façon "moyenne" un comportement qui serait trop compliqué à représenter de façon exacte (il s'agit alors d'une simplification), ou bien pour lequel l'analyste manque d'information. Autrement dit, on rompt volontairement avec l'approche déterministe en introduisant de l'aléatoire dans les modélisations afin de rendre les modèles solubles.

6 Les méthodes de résolution

6.1 Une vue globale des techniques de résolution

La figure 2 situe les méthodes de modélisation parmi l'ensemble des techniques d'évaluation des performances et présente les alternatives possibles de résolution des modèles. Une fois l'étape de modélisation réalisée, deux types

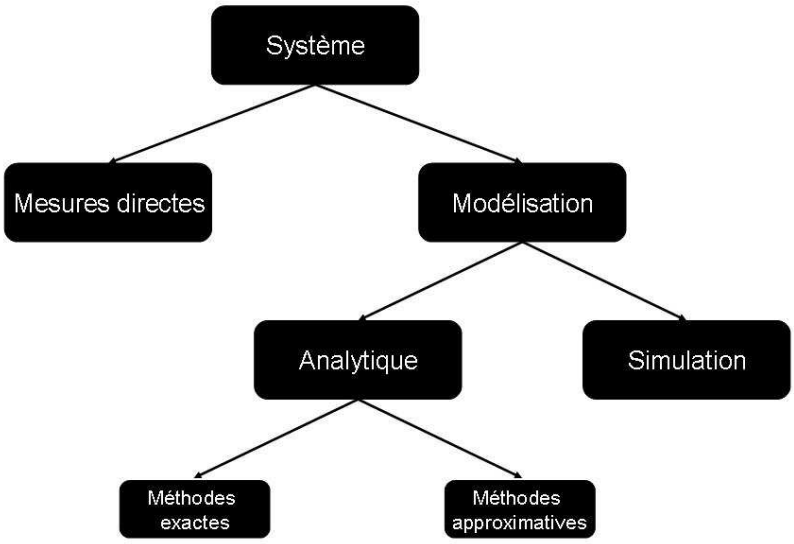


FIG. 2 – Techniques d'évaluation des performances

d'approches sont possibles pour résoudre le modèle et obtenir *in fine* les résultats (dans notre cas, il s'agira des paramètres de performances) escomptés :

- La résolution analytique (quelle soit exacte ou approchée)
- Et la résolution par simulation.

Chacune de ces méthodes comporte ses avantages et ses inconvénients.

Les solutions analytiques bénéficient de temps de résolution très rapides (les résultats peuvent être immédiats car ils sont déterminés à partir d'équations mathématiques issues du formalisme emprunté). Toutefois les modèles doivent être suffisament simples pour demeurer solubles par voie analytique. Ainsi la zone de fonctionnement de ces solutions, bien qu'étendue par les méthodes de résolutions approximatives, demeure assez restreinte. Notons que les simplifications sur les modèles comme les approximations de résolution, constituent des procédés à risques car elles peuvent aboutir à des résultats incohérents (en évinçant des phénomènes ayant une incidence notable sur le comportement à étudier).

A l'inverse les simulations permettent théoriquement de résoudre toutes les modélisations, y compris celles très détaillées, mais leur exploitation peut se révéler très longue et incertaine : une simulation peut durer très longtemps (une simulation peut durer jusqu'à plusieurs semaines). De plus si le modèle n'est pas ergodique (une simulation ne représente alors qu'une trajectoire possible du modèle parmi tant d'autres), l'expérience simulatoire devra être répétée plusieurs fois afin d'obtenir des résultats significatifs et exploitables. Toutefois les simulations représentent un outil de résolution des modèles très apprécié car il constitue une technique « simple » à mettre en oeuvre et efficace, du

moins pour obtenir des résultats bruts et numériques sur les performances d'un système.

Enfin un atout supplémentaire, et non des moindres, des méthodes de résolution analytique, tient au fait qu'elles permettent d'obtenir une perception intrusive dans la structure et dans le comportement du système grâce aux formulations littérales des expressions mathématiques qui régissent le système. Cette propriété peut s'avérer très bénéfique pour la compréhension d'un système car elle permet, entre autres, d'identifier rapidement et clairement les paramètres clés d'un système.

6.2 La résolution analytique des modèles

[2] explicite les principales méthodes de résolution analytique exacte destinées aux files d'attente et aux réseaux de files d'attente.
En très résumé des hypothèses fortes sur les processus d'inter-arrivée des clients et sur les distributions des lois de service des modèles (à savoir, on les suppose de type PH) permettent de rendre solubles analytiquement (en théorie du moins) les files d'attente. La résolution s'opère grâce à l'utilisation des Chaînes de Markov à Temps Continu. Pour les autres types de files d'attente, les procédés de résolution ne sont pas systématiques, voir inconnus dans de nombreux cas.
Pour les réseaux de files d'attente, les hypothèses permettant de résoudre les modèles d'une manière systématique sont encore plus restrictives. La démarche courante consiste à transformer le modèle initial en un réseau de Jackson ou en un réseau BCMP dont les propriétés particulières permettent l'utilisation de la méthode à forme produit.

Deuxième partie

Innovations inhérentes à la méthodologie employée

1 L'approche *classique*

Les systèmes informatiques peuvent généralement être décrits comme un assemblage subtil de ressources matérielles et logicielles couplé à un ensemble de demandes ou de tâches qui rivalisent pour accéder à ces ressources. Dans les ordinateurs personnels et les serveurs centraux, les ressources matérielles représentent typiquement la mémoire principale, le processeur, les disques et les terminaux. Les ressources logicielles pourront quant à elles correspondre aux accès bloquant en écriture à un fichier. Dans les réseaux informatiques, les ressources matérielles font référence aux commutateurs, aux buffers, aux liens physiques ... tandis qu'une ressource logicielle correspondra par exemple aux contraintes d'émission imposées par une fenêtre de contrôle d'émission sur un flot (typiquement un protocole comme TCP). Cette interaction entre les ressources et les demandes fait des systèmes informatiques un terrain favorable pour la modélisation.

1.1 Les pré-requis

Dans l'approche classique, la modélisation d'un système requiert une connaissance assez fine sur le système. En effet les unités logiques/physiques du système, dont dépend principalement le comportement à reproduire, doivent être représentées dans le modèle (le degré de granularité adopté étant fonction de la précision souhaitée). L'approche est qualifiée de constructive.

1.2 Les caractéristiques d'une approche constructive

Le fondement logique d'une approche constructive est de tirer profit de la connaissance du fonctionnement interne du système pour l'intégrer (de façon agrégée éventuellement) dans un modèle. Cette caractéristique justifie les prérequis évoqués précédemment.

1.3 Le processus de modélisation

Le processus de modélisation vise à transposer un système dans un formalisme adapté (mathématique le plus souvent) qui capture les traits essentiels relatifs au comportement visé du système étudié. L'évaluation des performances d'un système informatique par une technique de modélisation s'opère en 3 étapes.

1. La première étape consiste à modéliser le système et la charge de travail dans un formalisme adapté à l'étude que l'on veut en faire.
2. La seconde étape est consacrée à résoudre le modèle obtenu. La résolution peut faire appel à une technique analytique ou à une technique de simulation.

3. La troisième étape est celle de validation. Elle a pour but de vérifier si les performances obtenues par le modèle sont cohérentes (par exemple en confrontant les performances du modèles aux mesures observées). Ainsi il s'agit d'une remise en cause du modèle qui, dans le cas défavorable, aboutit à un rebouclage du processus de modélisation.

Le rebouclage de ce processus oblige souvent l'analyste à redéfinir les valeurs de ses paramètres. Cet exercice de calibrage minutieux des paramètres structurels, à l'enjeu crucial, s'avère en pratique être un exercice difficile et subtil. Au terme

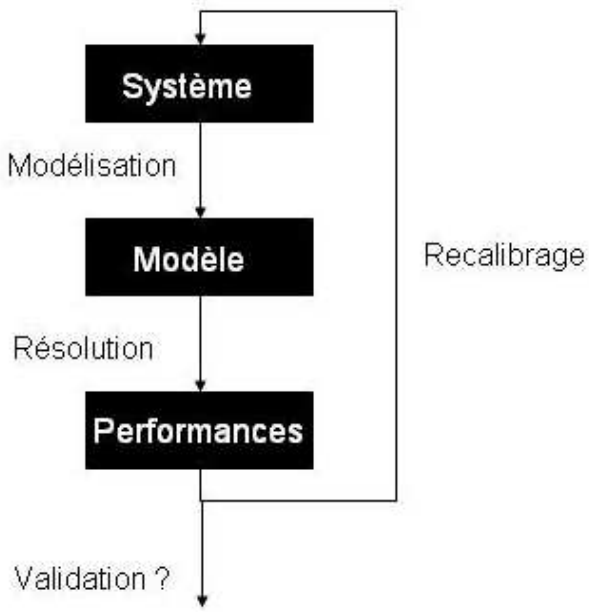


Fig. 3 – Processus de modélisation

de ces trois étapes, l'analyste dispose d'un modèle adapté et correctement calibré qu'il pourra exploiter pour évaluer les performances du système.

Nous avons noté précédemment qu'idéalement un modèle doit recouvrir tous les aspects liés aux objectifs d'évaluation et éclipser tout ceux n'intervenant pas dans l'usage escompté. De cette remarque on déduit deux conséquences pratiques. D'abord que le bon degré de complexité d'un modèle est le plus petit parmi ceux qui suffisent à reproduire fidèlement le comportement du système étudié. Et puis que la valeur d'un modèle n'est pas une notion absolue mais une notion relative à l'usage qui en est fait.

1.4 La balance entre un modèle à la fois fidèle et soluble

L'évaluation des performances d'un système par une technique de modélisation s'accompagne immanquablement d'une erreur comme l'explique [3]. [1] En fait à y voir de plus près, cette erreur est double car une erreur est introduite à chaque étape du processus de modélisation. Lors de l'étape de modélisation, une erreur (E1) apparaît inévitablement due à l'abstraction réalisée sur le système

[1] Nous reprenons ici leur explication

qui éclipse certains détails : le comportement du modèle ne reproduit pas exactement le comportement réel du système. Cette erreur est difficilement quantifiable (sauf lorsqu'il est possible de comparer les résultats de la modélisation avec ceux des mesures). Quant à la deuxième erreur, (E2), elle intervient lors de l'étape de résolution du modèle. Sa valeur dépend de la technique de résolution mise en oeuvre : elle est nulle pour les méthodes exactes, variable et difficile à évaluer pour la simulation, faible à parfois significative pour des méthodes approchées (typiquement 5 à 10 % d'erreur). Plus précisément pour les simulations, les taux d'erreurs s'exprimeront par des intervalles de confiance.

En pratique, (E1) et (E2) sont souvent corrélées. La réduction de l'une entraîne généralement l'augmentation de l'autre et vice-versa. Par exemple la description fine d'un système permet de diminuer l'erreur (E1) mais conduit à un modèle complexe souvent insoluble par des méthodes exactes. Si l'on exclut la simulation, seule une méthode approchée peut être envisagée avec pour conséquence l'accroissement de l'erreur (E2).
Et inversement, la construction d'un modèle simplifié ne traduit que très approximativement le comportement réel du système étudié et aura pour conséquence d'accroître (E1) mais de maintenir (E2) à un valeur nulle ou bien très basse. Une difficulté majeure réside donc dans le choix d'un compromis permettant d'atteindre l'objectif recherché, c'est-à-dire un modèle fidèle et soluble avec un coût (de calcul) minimal. L'habilité des experts en modélisation est donc de savoir identifier les pôles critiques et décisifs d'un système pour le comportement à reproduire, et d'y introduire ponctuellement un degré suffisant de complexité, tout en maintenant un haut niveau d'abstraction sur les autres pôles afin de conserver un modèle soluble et réaliste.

1.5 Les limites de cette approche

La complexité et/ou les dimensions de certains systèmes informatiques ont atteint un tel degré que la balance présentée ci-dessus est rendue impossible. Les facteurs précis de cette impossibilité peuvent être de différentes natures :

- Le système n'est pas suffisamment caractérisé (par exemple la charge de travail est inconnue).
- Le système n'est pas traduisible dans un formalisme connu.
- Le modèle engendré n'est pas soluble (quelle que soit la technique de résolution empruntée).
- L'étape de validation peut parfois être difficile à mettre en oeuvre (dû à l'absence de mesures sur certains systèmes par exemple).

2 L'approche novatrice adoptée

2.1 Une approche *descriptive*

Le projet vise à échafauder une méthodologie systématique et automatisable qui suggère rapidement et simplement une modélisation complète (modèle calibré et charge de travail compris) à partir d'un ensemble de mesures relevées sur un système quelconque. Contrairement à l'approche classique qui s'appuie sur une approche constructive du modèle, cette méthodologie ne présuppose aucune connaissance sur le système. Elle s'apparente à une approche « boîte

noire ». [4] qui a déjà envisagé conceptuellement ce type d'approche, qualifie ce type de méthodologie de purement descriptive par opposition aux approches constructives précédemment présentées. En effet dans sa forme la plus générale, notre démarche repose exclusivement sur les mesures (et pas seulement pour l'étape de calibrage comme c'est le cas communément). Toutefois des informations supplémentaires peuvent être enregistrées et permettront d'orienter plus efficacement la recherche (comme par l'exemple la constatation de pertes en entrée du système).

2.2 Le processus de modélisation revisité

Notre approche s'oppose foncièrement au processus de modélisation classique exposé ci-dessus. Elle s'appuie sur un réservoir prédéfini de modèles génériques et sur un ensemble de mesures relevées sur le système à modéliser. Conceptuellement la méthodologie élaborée vise à trouver parmi tous les modèles génériques disponibles, celui dont le calibrage ad hoc permettra d'obtenir des performances le plus proche possible de celles mesurées sur le système. Ainsi les mesures déterminent à la fois la structure du modèle et son calibrage (et pas uniquement le calibrage comme c'est fréquemment le cas).

2.3 Les usages pressentis/envisagés

Les usages pressentis pour notre projet sont les systèmes peu documentés et/ou incompris ou bien les systèmes dont la complexité/dimension rend impossible toute solution de modélisation classique.

Par ailleurs ce type d'approche pourrait également satisfaire aux demandes pressantes d'analystes qui souhaitent obtenir un modèle rapidement ou bien aisément.

Toutefois soulignons que l'approche descriptive exclut par essence un usage traditionnel des modélisations, à savoir l'évaluation des performances de modèles inexistants (pour lesquels aucune mesure n'est disponible).

2.4 Un sujet à risques

La modélisation d'un système par cette approche ne se fait pas sans risques et sans difficultés. On recense notamment parmi ces écueils :

- L'absence quasi-totale d'information sur le système considéré
- La définition d'un ensemble de modèles génériques respectant autant que possible les 3 règles ci-dessous :
    1. Aussi compact que possible (i.e. un cardinal aussi petit que possible),
    2. Avec des membres aussi « simples » que possible,
    3. Et dont les déclinaisons recouvrent un maximum de comportements observés sur des systèmes réels.
       Bien entendu la subtilité dans la définition de cet ensemble réside dans le compromis à trouver entre la règle 3 et les règles 1 et 2.
- L'existence d'une solution. L'obligation de rechercher un modèle « simple », pour modéliser un système parfois très complexe peut rendre le problème insoluble.

- Le calibrage arbitraire constitue la plus grande menace dans nos travaux. Ce risque apparaît très souvent et se présente comme une solution séduisante et fructueuse alors qu'en réalité il s'avère être une solution tout à fait inutile et inexploitable. En effet la mise en oeuvre d'un calibrage dont les paramètres sont régis par des règles faites sur mesure (on entend par là arbitraire) est tentante car elle permet de reproduire avec une précision aussi fine que voulue le comportement de n'importe quel système. Or un calibrage arbitraire supprime toute valeur prédictive au modèle ainsi établi puisqu'il n'est régi par aucune « logique ». C'est pour cette même raison, qu'on n'utilise pas, par exemple, une simple extrapolation (linéaire ou polynomiale) des mesures relevées pour inférer les performances du système pour d'autres niveaux de charges que ceux mesurés. Le corollaire de cette remarque est que, lors de l'étape de calibration de nos modèle, le maître mot devra être la simplicité. En effet la contrainte d'un modèle simple régi par des lois intuitives constitue la meilleure garantie contre le risque du calibrage arbitraire.
- L'estimation du pouvoir prédictif des modèles obtenus. Ce risque fait écho à la remarque précédente. Il sera capital mais probablement difficile de juger de la valeur prédictive des modèles retenus par notre analyse (par exemple en confrontant leurs prédictions à celles issues d'un modèle constructif).
- Le sens à donner aux paramètres des modèles obtenus. Il s'agit d'une difficulté propre à l'approche desciptive déployée dans nos travaux que nous pouvons énoncer de la façon suivante : Dans quelles mesures peut-on inférer des connaissances sur le fonctionnement interne du système à partir des modèles obtenus ? Les connaissances peuvent s'appliquer au degré de parallélisme, à la taille d'une file ...

2.5 Le contexte restreint du stage

Dans un premier temps, le temps de mon stage de Master II parcours Recherche, les travaux se sont concentrés sur la situation, qui pourra ultérieurement faire office de paradigme, dans laquelle l'analyste dispose de n couples de mesures $(\bar{X}_i ; \bar{R}_i)$ où $\bar{X}_i$ et $\bar{R}_i$ représentent respectivement le débit moyen mesuré en sortie du système et le temps de séjour moyen d'un client dans le système mesuré au point de mesure i (i varie entre 1 et n). Notons qu'une simple transformation par la loi de Little permet de généraliser ce cas à toutes les situations où $\bar{Q}_i$, le nombre de clients moyen au point de mesure i, supplée $\bar{X}_i$ ou bien $\bar{R}_i$. On prendra l'habitude dans ce rapport de tracer sur des graphes les courbes de performances issues de modèles ou de systèmes mesurés en réservant l'axe des abscisses au débit moyen $\bar{X}$ et l'axe des ordonnées au temps de séjour moyen $\bar{R}$.
A présent nous associerons systématiquement le terme *courbe de performance* à l'allure des courbes de performances du modèle ou du système considéré dans un graphe ainsi bâti. Ce contexte restreint du stage permet de traduire de manière plus concrète notre premier objectif. Il consiste à rechercher le calibrage *ad hoc* d'un modèle (faisant partie de l'ensemble prédéfini des modèles génériques) dont la courbe de performance se rapproche le plus de celle engendrée par les points de mesure. Le be but est donc de rapprocher autant que possible les paramètres de performances moyens $\bar{X}_{anal}$ et $\bar{R}_{anal}$ d'un modèle de ceux $\bar{X}$ et $\bar{R}$ mesurés

sur un système. La figure 4 illustre cette idée. La partie suivante présentera en détail les subtilités, les outils, les difficultés inhérentes et les étapes-clés à suivre pour atteindre cet objectif.

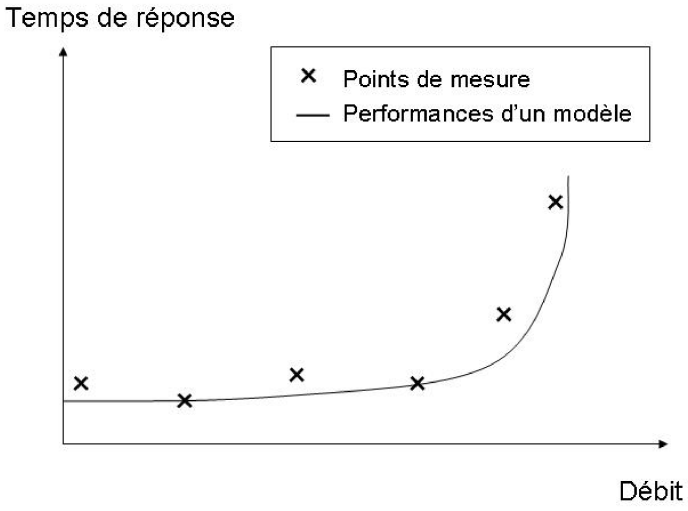


FIG. 4 – Formulation graphique de l'objectif

Troisième partie

Logique sous-jacente au projet

Cette partie a pour but de faire comprendre l'intuition qui encadre nos travaux. Au terme de sa lecture, le lecteur devrait disposer d'une vue globale sur notre méthodologie de calibration automatique de modèles. Les détails techniques et des précisions supplémentaires seront apportés dans la partie 6.

1 La formalisation et la décomposition du sujet du stage

Conceptuellement notre travail vise à rapprocher autant que possible la « courbe de performance »[2] d'un modèle généré automatiquement à celle engendrée par les mesures d'un système. Pour atteindre cet objectif, nous avons, dans un premier temps, opté pour la recherche du « meilleur » modèle parmi un réservoir prédéfini, constitué de modèles génériques que nous nommerons *les briques*. Par conséquent notre approche doit démarrer par la construction d'un ensemble de modèles génériques adapté aux systèmes considérés. Après cette étape préparatoire vient le processus de recherche de la meilleure instanciation de l'un de ces modèles, c'est-à-dire la recherche automatique du meilleur calibrage possible.

2 La définition d'un ensemble de modèles génériques

Au moment où ce rapport est écrit, l'ensemble des briques génériques est subdivisé en 3 sous-ensembles : *les briques de base, les modèles imbriqués* et *les modèles multiclasses*. Sur chacune des briques de ces 3 sous-ensembles peut venir se greffer une M/M/∞ en série. Nous indiquons très rapidement quelques informations sur la nature de ces briques, avant d'explorer en détail la mise en série d'une M/M/∞.

2.1 Les briques de base

2.1.1 Présentation

Le sous-ensemble des brique de bases regroupe les modèles les plus « simples » que nous avons jugé pertinents d'intégrer à nos travaux. L'instantiation optimale de ces briques de base doit permettre de recouvrir une portion importante des comportements observés sur les systèmes informatiques. C'est précisément la recherche de ce calibrage optimal pour chacun des modèles génériques qui fera l'objet de la quatrième partie de ce rapport. Dans cette partie , nous présentons un descriptif très sommaire des briques composant cet ensemble.

M/M/∞ : Modèle ouvert, sans perte, toujours stable, sans attente.

M/M/C : Modèle ouvert, sans perte, avec une condition de stabilité.

[2] Une notion définie précédemment

M/M/C/K : Modèle ouvert, avec perte en entrée (lorsque la file est pleine), toujours stable.

FERME : Modèle fermé avec N sources et un centre de traitement doté d'une file pouvant accueillir N clients, toujours stable.

FERME avec rejet : Modèle fermé avec N sources et une probabilité de rejet au niveau du centre de traitement (lorsque sa file est pleine), toujours stable.

2.1.2 Quelques exemples opportuns

Nous présentons dans la figure 5 un exemple probant des résultats obtenus à partir des briques de base. Les points de mesure sont issus du jeu de mesures 1 [3]. L'essentiel ici est de constater qu'une simple M/M/C/K favorablement calibré permet d'approcher « correctement » un jeu de mesures réelles. La procédure de calibration sera détaillée ultérieurement (dans la partie 4).

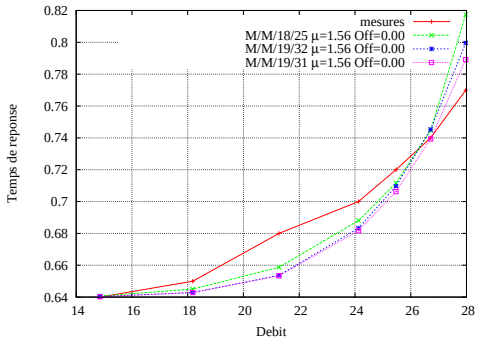


FIG. 5 – Courbes de performances du jeu de mesures 1 et des briques de base

A l'inverse la figure 6 présente un cas, le jeu de mesures 3, où les briques de base s'avèrent totalement incapables de reproduire convenablement les performances mesurées.

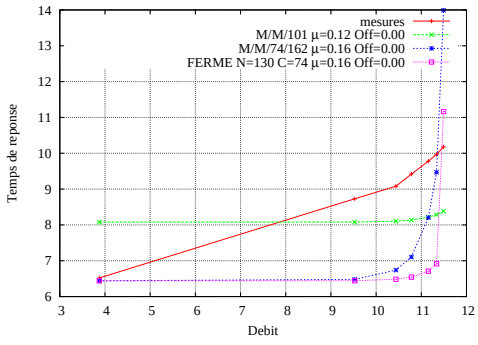


FIG. 6 – Courbes de performances du jeu de mesures 2 et des briques de base

[3] Voir Annexe B pour plus de détails sur les jeux de mesures

2.2 La mise en série d'une M/M/∞ - l'Offset

2.2.1 Présentation.

L'idée de placer une M/M/∞ est apparue consécutivement aux résultats jugés seulement « corrects » obtenus par les briques de base (voir figure 5). Afin d'améliorer ces résultats, nous avons pensé à une combinaison, la plus simple que l'on puisse faire. Elle consiste à placer en série une brique quelconque et une M/M/∞. Dans les modèles à perte, la M/M/∞ est à situer derrière le modèle principal de manière à ce que le temps de séjour des clients refoulés par le modèle principal reste nul.

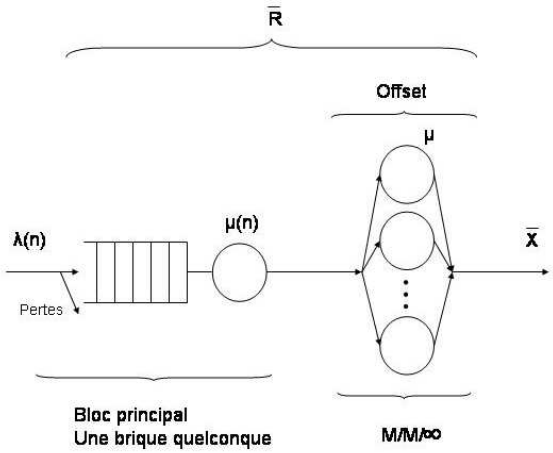


FIG. 7 – Interprétation d'un Offset : une M/M/∞ placée en série

Cette combinaison se démarque des autres briques du second degré (que nous verrons par la suite) car en réalité il s'agit plutôt d'une extension qui vient étoffer toutes les briques existantes. L'ajout de la M/M/∞, dont le temps de séjour quelle que soit la charge est une constante, induit uniquement la sommation d'une constante sur tous les temps de séjour calculés à partir du bloc principal. On peut donc considérer les temps de séjour associés au modèle complet comme composés de 2 éléments : une composante fixe que l'on associe à la M/M/∞ et une partie variable (selon la charge λ) résultant du bloc principal. Ainsi l'introduction d'une M/M/∞ placée en série derrière une brique quelconque impacte uniquement les temps de séjour $\bar{R}_{anal}$ du modèle et laisse indemnes le débit $\bar{X}_{anal}$ du modèle (ce qui explique le nom choisi d'*Offset*). Cette souplesse dans les $\bar{R}$ permet d'améliorer substantiellement les performances de nos briques de base [4] (sans pour autant remettre en cause le besoin pour de nouvelles briques du second ordre).

2.2.2 Interprétation d'une M/M/∞ placée en série.

L'interprétation physique d'une M/M/∞ placée en série derrière un modèle principal est simple et intuitive : la M/M/∞ représente un temps de séjour

[4] La fonction distance, estimatrice de la qualité des modèles, est réduite jusqu'à 6 fois sa valeur

supplémentaire incompressible et immuable, qui opère sur tous les temps de séjour (quel que soit le point de fonctionnement considéré) sans impacter les débits du modèle. Pour de nombreux systèmes, la présence d'une M/M/∞ placée en série dans le modèle associé peut être perçue comme un temps de transfert d'aller/retour fixe des clients pour rallier le centre de traitement.

2.2.3 Un aperçu des gains occasionnés par l'Offset.

Nous présentons graphiquement à travers un exemple, le jeu de mesures 1, les résultats occasionnés par l'Offset. Ces résultats 8 sont à comparer à ceux obtenus sur le même ensemble de mesures mais sans Offset 5. On observe que la présence de l'Offset permet d'améliorer très sensiblement la qualité des modèles retenus et d'obtenir des résultats qui nous conviennent.

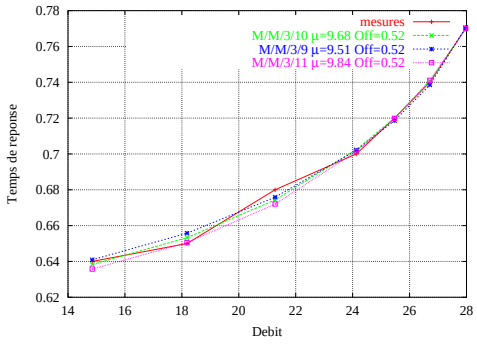


FIG. 8 – Courbes de performances du jeu de mesures 1 et des modèles trouvés avec Offset

En revanche comme l'illustre la figure 9, la présence de l'Offset ne permet toujours pas aux briques de base de reproduire correctement le comportement affiché par le jeu de mesures 2. Il faudra trouver de nouvelles briques.

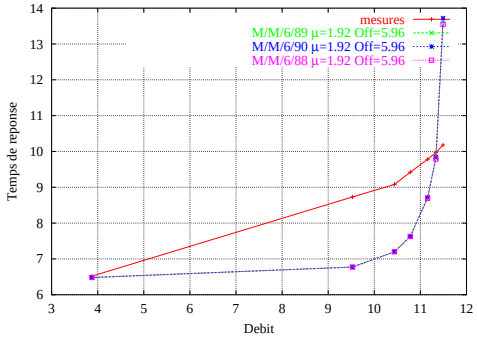


FIG. 9 – Courbes de performances du jeu de mesures 2 et des modèles trouvés avec Offset

Précisons qu'une partie entière est consacrée exclusivement à la présentation détaillée des résultats obtenus au cours de ce projet. Dans cette partie figureront naturellement les figures 5, et 8, accompagnées d'informations supplémentaires.

2.3 Les briques du « second degré »

2.3.1 Une série de modèles indispensables

En plus des briques de bases, nous avons défini deux sous-ensembles de modèles plus sophistiqués qui répondent à des besoins peut être plus ciblés. Historiquement, après avoir construit l'ensemble des briques de base et après avoir récolté quelques résultats prometteurs de génération automatique de modèles calibrés, le besoin pour de nouvelles briques s'est rapidement fait ressentir. Par exemple nous nous sommes rapidement trouvés confronter à des jeux de mesures qui démontrèrent l'insuffisance des briques de base : les jeux de mesures présentaient des comportements qu'aucune brique de base ne peut reproduire. La grammaire constituée uniquement des briques de base ne suffit pas (plus). De ce type de constat est née l'idée des modèles du second degré qui au gré des recherches a aboutit sur l'intégration de nouvelles briques dans notre ensemble de modèles génériques.

Cette remarque intéressante en amène une autre plus générale. Ce sujet, par essence, ouvre beaucoup de libertés dans sa réalisation et notamment dans la construction des ensembles et des sous-ensembles des modèles génériques (parmi lesquels sera rechercher le « meilleur » modèle) puisque les candidats y sont très nombreux (leur nombre est infini). La sélection des modèles intégrant cet ensemble doit donc être décidée judicieusement [5] pour conserver un ensemble de briques cohérent et compact. Les critères principaux de sélection d'un modèle candidat seront sa simplicité (tenant compte de son analyse et du jeu de paramètres associé), sa capacité à modéliser un type de comportements non reproductibles jusqu'alors et à un degré plus faible, de la qualité de l'interprétation physique pour le modèle considéré. En clair, construire une nouvelle brique pour chaque nouvelle courbe de performances serait dans notre approche un très mauvais calcul.

2.3.2 La génèse des modèles imbriqués

Cet assemblage de modèles est apparu en réaction au constat d'échec de notre approche à trouver un bon modèle parmi les briques de base pour les jeux de mesures 2 et 4 [6]. L'observation minutieuse des courbes de performances issues de ces mesures suggérait le besoin de modèles pour lesquels la saturation s'opère progressivement. Or comme l'illustre la figure 10, aucune brique de base ne peut reproduire ce comportement. En effet les briques de base ont toutes un temps de séjour relativement constant avant de saturer pleinement. Il nous fallait donc construire des modèles pour lesquels la saturation intervient de façon plus progressive et plus linéaire. Cela peut se traduire par des modèles dont le temps moyen de service varie avec la charge. C'est l'option que nous avons choisi.

C'est naturellement que nos recherches se sont orientées vers des modèles composés, plus sophistiqués, une famille de modèles, pour lesquels lorsque la charge en entrée du modèle augmente, elle s'accompagne d'un ralentissement du temps de service. Pour synthétiser ces modèles à temps de service variable, on va faire appel à des modèles imbriqués. La solution consiste à entremêler deux briques de base de la façon suivante : le temps de service du modèle principal

[5] Cette idée a déjà été évoquée au 2.4 page 19

[6] voir Annexes B

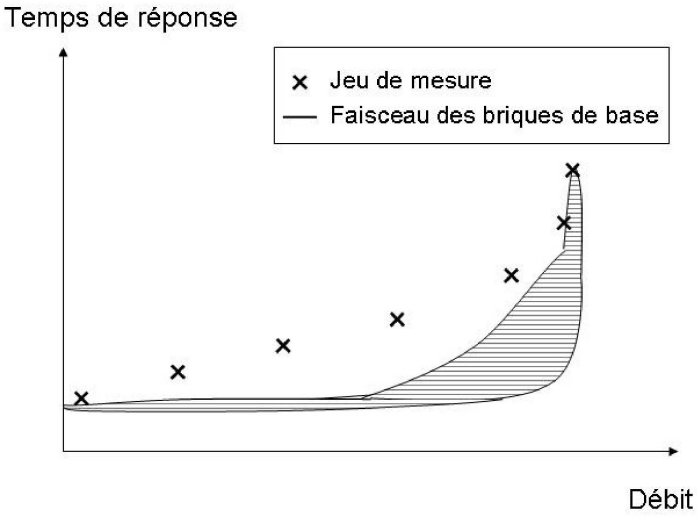


FIG. 10 – Illustration des limites des briques de base à reproduire certains comportement

(celui dont les performances doivent approcher tant que possible celles mesurées) est conditionné par le temps d'attente des clients dans le deuxième modèle, dit interne. Par conséquent plus il y aura de clients dans le modèle (c'est-à-dire de demandes d'accès aux ressources), plus les temps de service du modèle seront longs. Ce phénomène s'apparente à une contention ou bien à une promiscuité des clients. La figure 11 illustre les apports des modèles imbriqués en se basant sur le jeu de mesures 2, dont l'allure était jusqu'à présent non reproductible par les briques de base (comme en témoigne la figure 9 de la page 25).

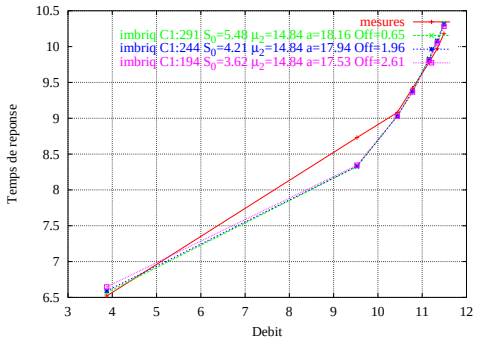


FIG. 11 – Courbes de performances du jeu de mesures 2 et des modèles imbriqués

L'analyse détaillée de cette nouvelle famille de modèles est présentée dans la quatrième partie de ce rapport (page 52).

2.3.3 Les modèles multiclasses

Tous les modèles présentés jusqu'à alors présentent une propriété commune : plus leur débit en sortie augmente, plus le temps de séjour de leurs clients s'allonge. Or il existe des systèmes informatiques au comportement inverse. Pour modéliser ce type de courbe de performance, nous avons opté pour un modèle sans perte à priorités comportant deux classes deux clients (1 et 2). On suppose que la somme des débits moyens du trafic de 1 et de 2 reste constante à tous les points de fonctionnement. On suppose également que le trafic 1 est celui mesuré et que le trafic 2 correspond à un trafic exogène, prioritaire sur le trafic 1. Une fois ces hypothèses faites, nous expliquons conceptuellement pourquoi ce modèle va satisfaire à notre demande en courbes de performances décroissantes (du moins dans un premier temps). Plus le débit des clients 1 est important, moins il y aura de clients 2 dans le système (puisque la somme de leur débit doit rester fixe et que le temps de séjour des clients demeure identique) et donc moins les clients de 1 seront pénalisés par les clients de 2. Inversement lorsque le débit de 1 est faible, les clients de 2 sont majoritaires et retardent d'autant les clients de la classe 1.

La figure 12 présente un exemple des résultats obtenus par la brique multiclasse pour le jeu de mesures décroissants 5.

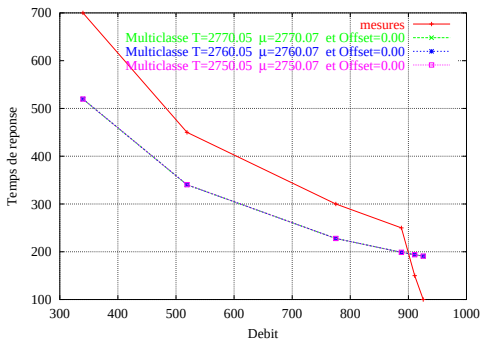


FIG. 12 – Courbes de performances du jeu de mesures 5 et des briques multiclasses

Des précisions sur l'implémentation choisie de ces modèles ainsi que leur analyse seront également détaillées dans la partie suivante de ce rapport.

3 La recherche du meilleur modèle calibré

La recherche du meilleur modèle calibré requiert la mise en place d'une stratégie. Nous verrons que plusieurs approches sont possibles. Mais auparavant, nous décrivons quelques outils indispensables et les étapes-clés pour mener à bien cette recherche.

3.1 Borner l'espace de recherche

Cette étape vise à définir pour chaque paramètre du modèle (exceptée la charge λ) une borne inférieure et une borne supérieure à ne pas dépasser. Pour

obtenir ces relations, on exploite certaines lois sur les modèles couplées aux mesure dont on dispose : conditions de stabilité, loi de Little...

Les bornes inférieures et supérieures d'un paramètre quelconque A seront référencées respectivement comme A_{min} et A_{max}.

Exemple. Cet exemple a pour but d'aider le lecteur a mieux saisir la notion de bornes sur les paramètres. Il existe une borne inférieure simple sur K pour les modèles de type M/M/C/K (i.e. Arrivées Poissoniennes, loi de service exponentielle, C serveurs et une capacité de K). La loi de Little énonce que le nombre moyen de clients $\bar{Q}$ dans le modèle est égale à $\bar{R}\bar{X}$. Dans ce cas le fonctionnement des bornes consiste à penser que si une M/M/C/K est recherchée pour reproduire le système mesurée, alors celle-ci doit satisfaire au moins à la loi de Little. Du coup on imposera au paramètre K d'être supérieur à chacun des i produits $\bar{X}_i\bar{R}_i$.
Malheureusement il n'est pas toujours possible de trouver des bornes supérieures pour certains paramètres. Dans ce cas la solution de dernier recours consiste à incrémenter progressivement la valeur du paramètre en question jusqu'à temps que l'on constate une monotonie négative des résultats (c'est-à-dire que les résultats de param+1 soient régulièrement moins bons que ceux de param). Il va de soi que les plus l'espace entre une borne inférieure et une borne supérieure est mince, plus la recherche du meilleur calibrage pourra être rapide. Du coup fournir un effort de calcul conséquent sur l'obtention de bornes resserées s'avère souvent être un bon calcul.

3.2 Inférer la charge de travail : λ

Le calcul de la charge de travail en entrée est une étape cruciale qui peut s'avérer être une opération parfois triviale, et parfois une opération bien plus subtile.
Dans les modèles ouverts sans perte, le débit moyen mesuré en sortie en régime stationnaire $\bar{X}$ est égal à la charge appliquée en entrée λ. L'obtention de la charge du modèle est donc automatique : $\lambda = \bar{X}$.
En revanche dans les modèles à perte ou à rejet, le débit moyen en entrée et en sortie du modèle diffèrent (à cause des rejets). On a donc pour ces modèles l'inégalité suivante : $\bar{X} \leq \lambda$. Cependant l'obtention rigoureuse du niveau de charge λ (qui a engendré un débit moyen de sortie $\bar{X}$) sera toujours possible pour les briques choisies dans ce rapport. Cette possibilité découlera d'une propriété remarquable et commune à tous les modèles traités dans ce rapport : l'évolution du débit moyen en sortie $\bar{X}$ est strictement monotone avec la charge λ en entrée [7].

3.3 Apprécier la qualité d'un modèle candidat - la fonction distance

Puisqu'on a aligné le débit moyen des modèles à celui mesuré effectivement sur le système ($\bar{X} = \bar{X}_{anal}$), il est naturel de mesurer l'écart entre chaque $\bar{R}_i$ et chaque $\bar{R}_{i,anal}$ associé pour apprécier l'aptitude d'un modèle à reproduire le comportement d'un système. Pour estimer cet écat sur l'ensemble des points de

[7] Se reporter aux parties 4 et 6 pour plus de précisions

mesures, nous avons défini une *fonction de distance d*. Cette fonction prend en entrée un ensemble de mesures ($\bar{X}_i$;$\bar{R}_i$) et un ensemble de performances issues d'un modèle ($\bar{X}_i$;$\bar{R}_{i,anal}$) et retourne une valeur estimatrice de l'écart général séparant ces deux ensembles. Plus cette note sera basse, « meilleur » sera le modèle. Plus formellement nous avons opté pour la fonction distance suivante :

$$d : (\bar{R}_i, \bar{R}_{i,anal}) \longrightarrow \Sigma_{i=1}^{n} w_i \times abs(\frac{\bar{R}_i - \bar{R}_{i,anal}}{\bar{R}_i}) \quad (1)$$

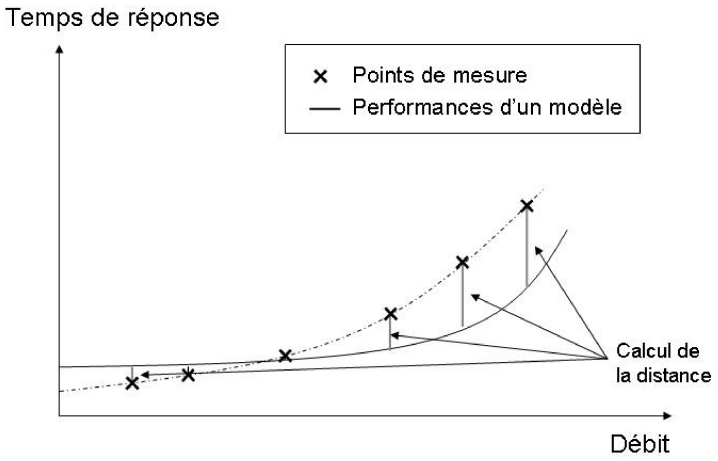


FIG. 13 – Calcul de distance entre les performances mesurées et synthétisées

On note dans cette expression de la fonction distance d la présence de coefficients pondérateurs w_i. Leur rôle exact est explicité plus tard (dans la partie 6, page 82) car ces coefficients ne sont pas indispensables à la compréhension ni à la réalisation du processus de recherche : ils permettent d'étendre ses fonctionnalités. Pour l'instant, on peut supposer tous les w_i égaux et leur somme égale à 1.

3.4 Orchestrer la recherche - Une approche itérative ou une approche intelligente ?

A présent nous allons présenter le déroulement de la recherche de la meilleure brique calibrée. Plusieurs approches sont possibles :

- Une recherche purement systématique
- Une recherche orientée

3.4.1 Une recherche systématique

Pour les briques de base, le nombre « limité » de combinaisons possibles à partir des espaces de recherche (donc des bornes), nous permet d'opter pour une approche itérative. Cette approche simple et exhaustive bien que gourmande se concilie bien avec la phase actuellement exploratoire de ce projet. Son implémentation telle quelle garantit à coup sûr l'obtention de tous les meilleurs candidats sans risquer les effets de bords consécutifs à une recherche intelligente

(risque de passer à côté du modèle optimal). Une garantie fort appréciable lors du stade exploratoire de ce projet. Le temps de calcul est maintenu à des délais raisonnables grâce à la simplicité des modèles mis en jeu conjuguée à l'utilisation d'heuristiques qui permettent de borner l'espace de recherche et d'évincer rapidement certains candidats.

3.4.2 Une recherche orientée

Pour les modèles du second degré, une recherche plus intelligente est rendue obligatoire tant le nombre de combinaisons entre les paramètres est rendu nombreux. Pour le moment, comme nous le verrons ultérieurement dans la partie suivante, l'approche n'est plus exhaustive et demeure en revanche itérative dans des portions restreintes de l'espace de recherche total (à $nb_{parametres}$ dimensions). La différence tient donc à la définiton de ces portions de l'espace dans lesquelles les probabilités de présence du calibrage optimal semble les plus importantes. Ainsi la méthode pourrait être qualifiée de démarche itérative restreinte.

3.4.3 La nature des paramètres

Notons que l'on peut classer les paramètres structurels de nos modèles dans deux catégories selon qu'ils sont de nature continues (comme le taux de service d'un serveur μ) ou de nature discrètes (comme le nombre de serveurs C). L'appartenance d'un paramètre à une de ses catégories détermine la façon dont seront parcourus les valeurs dans l'espace de recherche à tester.
Pour un paramètre discret, c'est simple, il suffit de parcourir l'ensemble des entiers contenus entre la borne inférieure et la borne supérieure de son espace de recherche.
Pour un paramètre continu, la situation est plus compliquée. Il faut définir le *pas* avec lequel nous souhaitons parcourir l'ensemble des valeurs de son espace de recherche. Cette précision sur les paramètres continus devra tenir compte des ordres de grandeurs des mesures en entrées. Pour un paramètre continu quelconque A, nous ferons référence à sa précision par la notation $precision_A$.

Quatrième partie

Analyse détaillée de chaque brique

Pour toute la suite du rapport, nous notons que les paramètres de performance feront systématiquement référence aux paramètres de performance moyens (d'un modèle comme d'un système).

1 M/M/∞

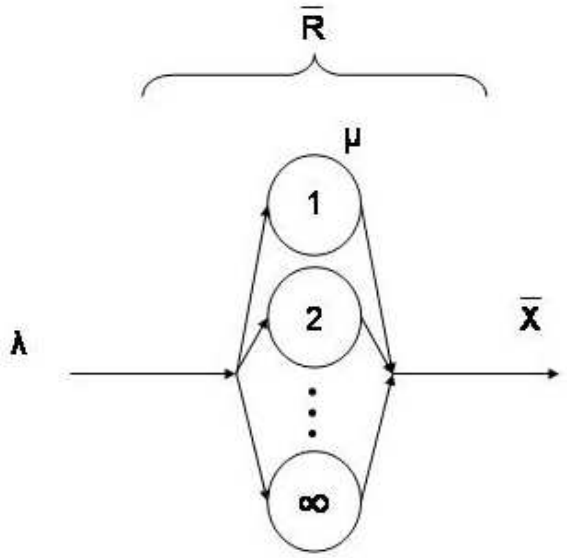


FIG. 14 – Représentation graphique d'une M/M/∞

1.1 Généralités

- Modèle ouvert toujours stable avec un nombre infini de serveurs.
- Paramètres intrinsèque au modèle : Un nombre infini de serveurs, une loi de service exponentielle de taux μ
- Paramètre de charge : un processus d'arrivée des clients selon une loi de Poisson de taux λ.

1.2 Analyse du régime permament

Pour la M/M/∞ les paramètres de performance moyens en régime stationnaire sont très simples (pas besoin de calculer les probabilités d'états stationnaires).

1.2.1 Les paramètres de performance

Nous présentons les expressions littérales des deux paramètres de performances qui nous intéresse pour une M/M/∞.

DÉBIT :	$\bar{X}_{anal} = \lambda$
TEMPS DE SÉJOUR :	$\bar{R}_{anal} = \frac{1}{\mu}$

Le débit moyen λ en entrée d'une M/M/∞ est toujours égal au débit moyen en sortie $\bar{X}_{anal}$. Et le temps moyen de séjour de ses clients est toujours le même : $\frac{1}{\mu}$: il n'y a jamais d'attente dans une M/M/∞. Par conséquent si on ne s'intéresse qu'aux paramètres de performances moyens, l'impact d'une M/M/∞ sur un trafic est nul pour son son débit $\bar{X}_{anal}$ et se traduit par l'ajout d'un délai constant sur tous les temps de séjour $\bar{R}_{anal}$. (rôle de temporisateur).
Cette propriété rend la brique M/M/∞ un peu particulière. Une telle brique n'est pas très utile pour reproduire des systèmes. Elle correspondrait à des systèmes pour lesquels quelque soit le débit moyen des clients, le temps de séjour moyen mesuré reste identique. En revanche nous verrons plus tard que cette brique peut jouer un rôle important en venant se combiner à une autre brique.

2 M/M/C

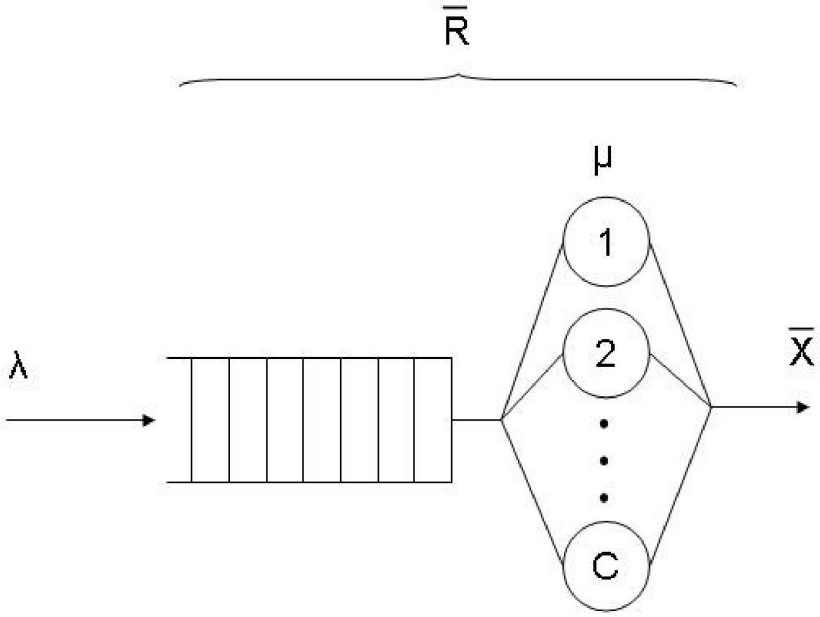


FIG. 15 – Représentation graphique d'une M/M/C

2.1 Généralités

- Modèle ouvert à capacité infinie.
- Paramètres intrinsèque au modèle : C serveurs, une loi de service exponentielle de taux μ
- Paramètre de charge : un processus d'arrivée des clients selon une loi de Poisson de taux λ.
- **Condition de stabilité** : $\lambda < C\mu$

2.2 Analyse du régime permament

2.2.1 Les probabilités d'états stationnaires

Soit $p(n)$ la probabilité d'avoir n clients dans la file.

$$\begin{cases} p(n) = \begin{cases} \frac{\rho}{n} \times p(n-1) = \frac{\rho^n}{n!} \times p(0) & \text{si n} \in [1\,;\text{C}] \\ \frac{\rho}{C} \times p(n-1) = \frac{\rho^n}{C!C^{n-C}} \times p(0) & \text{si } n > C \end{cases} \\ \text{avec } \sum_{n=0}^{\infty} p(n) = 1 \end{cases}$$

2.2.2 Les paramètres de performance

Nous présentons les expressions littérales des deux paramètres de performances qui nous intéresse pour une M/M/C.

DÉBIT :	$\bar{X}_{anal} = \sum_{n=1}^{C-1} p(n) n\mu + \sum_{n=C}^{\infty} p(n) C\mu = \lambda$
TEMPS DE SÉJOUR :	$\bar{R}_{anal} = \bar{S}_{anal} + \bar{W}_{anal} = \frac{1}{\mu} + \sum_{n=C}^{\infty} p(n) \frac{n+1-C}{C\mu}$

2.3 Recherche du calibrage *ad hoc*

Nous détaillons ci-dessous toutes les éléments entrant en jeu dans la recherche d'un calibrage d'une M/M/C répondant à certains critères. Dans notre cas, il s'agit de rechercher le calibrage « optimal » permettant d'approcher au plus près des points de mesure. Mais avant, nous présentons l'allure des courbes de performances d'une M/M/C

2.3.1 L'allure des courbes de performance

Dans cette partie, nous présentons la physionomie des courbes de performance [8] des M/M/C et puis, plus précisément, l'influence des paramètres C et μ sur cette allure (voire figure 16). On commence par constater que les courbes de performances d'une M/M/C sont *monotones* ($\bar{R}$ croît inmanquablement avec λ) et *bimodales*, c'est-à-dire globalement composées de deux parties qui définissent le comportement à faible charge et à charge élevée.

On observe que pour de faibles charges, le temps moyen de réponse $\bar{R}_{anal}$ varie peu et reste très proche de la valeur du temps de service moyen d'un client $\frac{1}{\mu}$. Cette observation résulte du service quasi-immédiat des clients entrant dans la file à faible charge.

On observe également une asymptote verticale en la valeur $C\mu$. En effet, le débit moyen maximum d'une M/M/C est $C\mu$ (il correspond à des charges proches de $C\mu$) et les niveaux de charges associés à des débits très élevés engendrent des temps de séjour $\bar{R}_{anal}$ très longs qui tendent vers l'infini.

Par ailleurs, nous nous sommes rendus compte que plus une M/M/C a de serveurs, plus la courbure de sa courbe de performances est étroite. Et inversement que plus C est petit, plus le rayon de courbure de la courbe est ouvert. Intuitivement cela s'explique de la façon suivante. Il est clair que plus on se rapproche du seuil de saturation d'une M/M/C, plus le temps de séjour moyen des clients augmente. Et cette augmentaion va en grandissant : la pente est exponentielle. Or cette accélération sera d'autant plus marquée qu'il y a de serveurs dans la M/M/C. En effet pour un débit moyen en entrée égale à λ pour deux M/M/C, les chances pour un client d'être traité sans attente (c'est à dire de trouver une place directement dans un serveur) sont d'autant plus grande, que le nombre de serveurs est grand. Autrement dit, pour un même $\bar{X}_{anal}$, c'est la M/M/C avec le plus petit nombre de serveurs C qui aura un temps de séjour moyen $\bar{R}_{anal}$ le plus long. Cette distinction s'estompe pour les faibles charges (quel que soit le nombre de serveurs C du modèle, les clients sont quasi-toujours servis sans attente) et pour les charges très élevées (quel que soit le nombre de serveurs, la probabilité pour un client d'être servis directement est quasi nulle et ce quel que soit le nombre de serveurs du modèle). Ainsi plus C sera grand,

[8] A nouveau, on considère uniquement $\bar{R}$ en fonction de $\bar{X}$

plus la courbure sera prononcée et inversement plus C est petit plus le rayon de courbure est droit.

La liste-dessous et la figure 16 récapitulent l'impact des paramètres dans ce contexte.

* Le quotient $\frac{1}{\mu}$ (sommé éventuellement à un Offset) détermine la valeur du temps de réponse à faible charge.
* Le produit $C\mu$ spécifie la charge et le débit maximum supportés par le modèle, c'est-à-dire le seuil de saturation.
* La valeur de C règle le rayon de courbure de la courbe de performances.

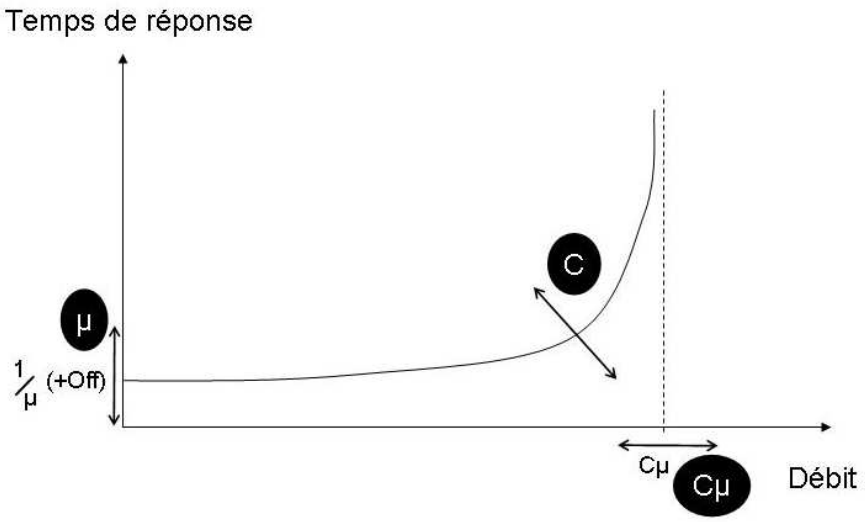


FIG. 16 – Influence des paramètres d'une M/M/C sur ses performances

L'observation de la figure 16 souligne l'existence de trois degrés de liberté pour calibrer une courbe de performances d'une M/M/C (un pour chaque asymptote et le dernier pour le rayon de courbure). Or dans sa forme initiale, le modèle M/M/C ne dispose que de deux paramètres : C et μ. Dans ces conditions, l'ajustement de la courbe pour deux de ses degrés de liberté définit une valeur pour C et pour μ. Par conséquent le degré de liberté restant n'en est plus un (!) : il est imposé (puisqu'il dépend lui aussi d'une combinaison de C et de μ) et n'est donc plus du ressort de l'analyste [9].

2.3.2 Quelques propriétés remarquables

De manière générale, les propriétés énoncées dans cette partie seront mises à contribution pour optimiser certaines recherches récursives de paramètres. Typiquemment lorsqu'on recherche la valeur à donner au dernier paramètre à fixer d'une M/M/C (en pratique il peut s'agir de μ) en vue de reproduire,

[9] On verra dans la sixième partie de ce rapport comment l'Offset pallie ce problème

avec une précision arbitraire pour les paramètres continues, un paramètre de performance mesuré tel que $\bar{R}$.

- $\bar{X} = \lambda$ (puisqu'il n'y a pas de perte)
- Plus λ croît [10], plus le débit en sortie du modèle augmente et plus le temps de réponse s'allonge.
- Plus μ diminue, plus le temps de séjour s'allonge et inversement en augmentant μ, on réduit le temps de séjour.
- Nous constatons également que pour des valeurs de λ proches de $C\mu$, le débit en sortie s'approche de $C\mu$ (sans jamais le dépasser) et qu'en revanche, le temps de réponse explose, tendant vers des valeurs infinies.

2.3.3 Les bornes sur les paramètres

Type de borne	*Expression littérale*	*Signification*
Inférieure sur C	1	
Inférieure sur μ	$\max_{i\in[1;n]}(\frac{X_i}{C})$	Traduction de la condition de stabilité du modèle $\forall i \in [1;n], \lambda_i < C\mu$
Inférieure sur μ	$\max_{i\in[1;n]}(\frac{1}{R_i})$	N'importe quel temps de séjour mesuré est plus grand que le temps moyen de service seul $\forall i \in [1;n], \bar{R}_i \geq \frac{1}{\mu}$

TAB. 1 – Bornes pour une M/M/C

Evidemment lorsque plusieurs bornes sont disponibles pour un même paramètre, on conserve, s'il s'agit d'une borne supérieure, la plus petite de valeurs et la plus grande des valeurs dans le cas contraire.

2.3.4 Le processus de recherche du calibrage adéquat/optimal

Une recherche systématique. Pour le moment, pour les raisons énoncées précédemment (page 30), la démarche consiste à parcourir itérativement toutes les combinaisons autorisées des paramètres. Pour cela on ordonne la recherche par une séquence de boucles imbriquées. L'ordre des boucles ne doit pas être choisi au hasard car certaines séquences s'appréhendent plus facilement que d'autres (essentiellement pour définir des bornes). La recherche itérative peut déboucher sur un nombre considérable de modèles à tester.

Une recherche orientée. Toutefois cette approche exhaustive, adaptée aux besoins d'une étude prospective, ne doit pas faire oublier que d'autres approches, plus « intelligentes » sont possibles. On pourrait par exemple, dans le cas où l'on s'autorise l'Offset, choisir C en fonction du rayon de courbure observé sur la courbe des mesures, décider d'une valeur de μ de manière à situer le produit $C\mu$ à la valeur maximale de débit que semble présenter la courbe des mesures et enfin retenir une valeur de Off qui égalise le temps de séjour à faible charge du modèle à celui mesuré.

[10] Toute chose étant égale par ailleurs

2.3.5 L'algorithme de recherche mis en oeuvre

Nous présentons ici la structure de la recherche exhaustive mise en oeuvre pour la brique M/M/C.

1. Boucle sur C depuis 1 jusqu'à ce que les augmentations successives sur C engendrent des résultats systématiquement plus mauvais.
2. Calcul de la borne inférieure sur μ pour la valeur de C à l'essai.
3. Boucle sur μ depuis la borne inférieure sur μ jusqu'à ce que les augmentations successives sur μ engendrent des résultats systématiquement plus mauvais. Le pas de recherche sur μ est fixé en tenant compte de l'ordre de grandeur des mesures et de la précision souhaitée sur les résultats (plus de détails à la page 31).
4. Calcul (trivial) des λ associés aux $\bar{X}$ (cette étape peut être réalisée une bonne fois pour toute à l'amorce de la recherche puisque pour toutes les M/M/C, $\lambda = \bar{X}$).
5. Calcul du temps de séjour moyen des clients $\bar{R}_{i,anal}$ pour chaque niveau de charge λ_i.
6. Calcul de la fonction distance (se fiant aux écart entre $\bar{R}_{i,anal}$ et $\bar{R}_i$) pour le modèle à l'essai. Si le résultat du calcul situe le modèle parmi les 3 meilleurs modèles jusqu'à présent testés alors le modèle est retenu et remplace un modèle dans la liste provisoire des meilleurs modèles.

A présent nous sommes en mesure de dénombrer très exactement le nombre de modèles testés pour la brique M/M/C. Ce nombre est égal à : $\sum_{C=1}^{C_{max}} \frac{\mu_{max}-\mu_{min}}{precision_\mu}$. Notons que les valeurs de C_{max} et de μ_{max} ne sont pas connues « à priori » : elles se déterminent empiriquement [11]. Par conséquent il n'est pas possible, avec notre approche de recherche, de prédire le temps de calcul de notre recherche.

[11] Pour μ, l'optimisation par les μ locaux permettra d'obtenir une borne supérireure, page 86

3 M/M/C/K

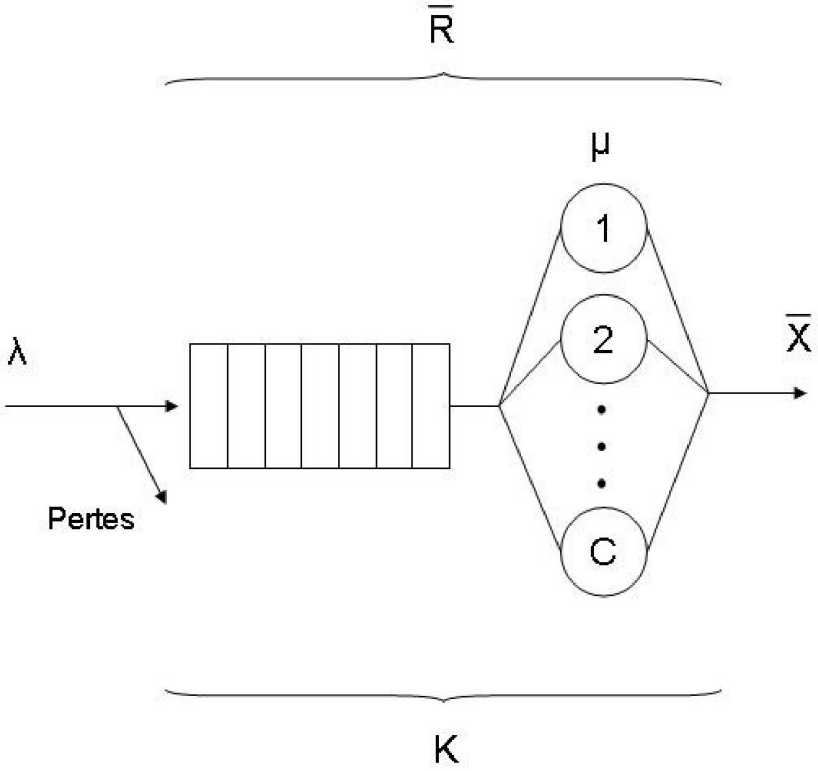


FIG. 17 – Représentation graphique d'une M/M/C/K

3.1 Généralités

- Modèle ouvert à capacité finie.
- Système limité (avec rejets/pertes) donc toujours stable.
- Paramètres intrinsèques au modèle : C serveurs, capacité de K clients et une loi de service exponentielle de taux μ .
- Paramètre de charge : un processus d'arrivée des clients selon une loi de Poisson de taux λ.

3.2 Analyse du régime permament

On définit ρ comme étant le rapport de λ sur μ (i.e. $\frac{\lambda}{\mu}$).

3.2.1 Les probabilités d'états stationnaires

Soit $p(n)$ la probabilité d'avoir n clients dans la file.

$$\begin{cases} p(n) = \begin{cases} \frac{\rho}{n} \times p(n-1) = \frac{\rho^n}{n!} \times p(0) & \text{si } n \in [1;C] \\ \frac{\rho}{C} \times p(n-1) = \frac{\rho^n}{C!C^{n-C}} \times p(0) & \text{si } n \in [C+1;K] \end{cases} \\ \\ \text{avec } \sum_{n=0}^{K} p(n) = 1 \end{cases}$$

3.2.2 Les paramètres de performance

DÉBIT :	$\bar{X}_{anal} = \lambda \times \sum_{n=0}^{K-1} p(n) = \lambda \times (1 - p(K))$
TEMPS DE SÉJOUR :	$\bar{R}_{anal} = \frac{1}{\mu} + \sum_{n=C}^{K-1} \frac{p(n)}{1-p(K)} \frac{n+1-C}{C\mu}$
NOMBRE DE CLIENTS :	$\bar{Q}_{anal} = \sum_{n=1}^{K} n \times p(n)$

3.3 Recherche du calibrage *ad hoc*

3.3.1 L'allure des courbes de performance

Comme l'illustre le graphique 18, les M/M/C/K exhibent des courbes de performance monotones et bimodales.

On observe que pour de faibles charges, le temps moyen de réponse $\bar{R}_{anal}$ varie peu et reste très près de la valeur du temps de service moyen d'un client $\frac{1}{\mu}$. Cette observation résulte du service quasi-immédiat des clients entrant dans la file à faible charge.

Cependant ces courbes se différencient substantiellement de celles des M/M/C de par l'apparition d'un point d'accumulation. Pour des niveaux de charge λ élevés (supérieurs à $C\mu$), le débit et le temps de séjour d'une M/M/C/K se rapprochent très près de leur valeur maximale respective. C'est pourquoi les courbes de performance des M/M/C/K exhibent un point d'accumulation aux coordonnées $(C\mu; \frac{K}{C\mu})$.La justification des coordonnées de ce point d'accumulation est relativement simple. Son abscisse correspond à la valeur maximale de débit moyen en sortie. Cette situation intervient lorsque la charge λ est très élevée et que donc la file sature en permamence. Elle débite alors en sortie à son taux maximum, soit $C\mu$. Quant à son ordonnée, elle traduit le temps de séjour moyen maximum d'un client dans le modèle. Ce temps de séjour sera à son maximum lorsque la file sature complètement. Le client entrant est alors systématiquement accueilli en dernière position de la file ce qui signifie qu'il y a en moyenne K clients dans la file. La loi de Little nous donne alors très simplement l'expression du temps de séjour moyen : $\bar{R} = \frac{K}{C\mu}$.

Comme pour la M/M/C et pour des raisons similaires (page 35), le nombre de serveurs modifie la courbure de la courbe. Plus ce nombre sera grand, plus le rayon de courbure sera serré. Et inversement plus C est petit, plus la courbure sera droite.

La liste ci-dessous et la figure 18 récapitulent l'impact des paramètres dans ce contexte.

* Le quotient $\frac{1}{\mu}$ (sommée éventuellement à Off) détermine la valeur du temps de réponse à faible charge

* Le produit $C\mu$ spécifie le débit maximal supporté par le modèle et participe à situer le point d'accumulation de la courbe
* La valeur de C règle le rayon de courbure de la courbe de performances
* La valeur de K situe le point d'accumulation sur la courbe en imposant comme borne supérieure sur le temps de réponse : $\frac{K}{C\mu}$

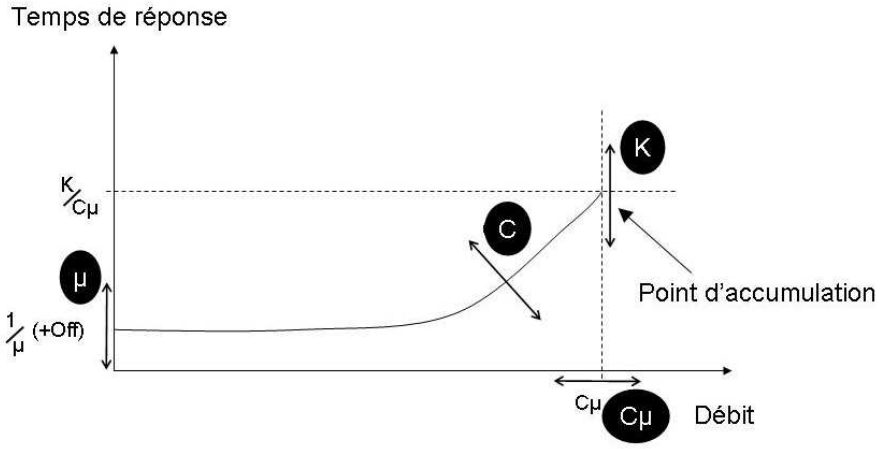


FIG. 18 – Influence des paramètres d'une M/M/C/K sur ses performances

3.3.2 Quelques propriétés remarquables

1. $\bar{X} \leq \lambda$ (à cause des pertes éventuelles)
2. Plus λ croît, plus le débit en sortie du modèle augmente et plus le temps de réponse s'allonge jusqu'à atteindre leur valeur maxaimale respective : $C\mu$ et $\frac{K}{C\mu}$.
3. Plus μ diminue, plus le temps de séjour s'allonge et inversement en augmentant μ, on réduit le temps de séjour.
4. Plus ρ croît, plus le nombre de clients dans le modèle augmente jusqu'à atteindre sa valeur maximale K. Similairement à la deuxième remarque, cette propriété permet pour un nombre de clients moyen mesuré $\bar{Q}$ de retrouver la valeur de ρ qui engendre un $\bar{Q}_{anal}$ identique.

3.3.3 Les bornes sur les paramètres

3.3.4 Le processus de recherche du calibrage adéquat/optimal

Une recherche systématique. Pour l'instant, pour les raisons que nous avons énoncées auparavant (page 30), nous avons opté pour une démarche exhaustive. La démarche consiste à parcourir itérativement toutes les combinaisons autorisées des paramètres. Pour cela on ordonnance la recherche par une séquence de boucles imbriquées. A nouveau l'ordre des boucles ne doit pas être choisi au hasard car certaines séquences s'appréhendent plus facilement que d'autres (pour définir des bornes et pour réduire le nombre d'apparition de certains calculs fastidieux).

Type de borne	*Expression littérale*	*Signification*
Inférieure sur C	1	
Inférieure sur K	$max_{i\in[1;n]}(R_i \times X_i)$	Loi de Little sur la relation $\forall i \in [1;n], K > \bar{Q}_i$
Inférieure sur μ	$\max_{i\in[1;n]}(\frac{X_i}{C})$	Le débit en sortie du modèle ne peut pas dépasser $C\mu$ $\forall i \in [1;n], \bar{X}_i \leq C\mu$
Inférieure sur μ	$\max_{i\in[1;n]}(\frac{1}{R_i})$	Les R_i sont plus grands que le temps moyen de service seul $\forall i \in [1;n], \bar{R}_i \geq \frac{1}{\mu}$
Supérieure sur μ	$\min_{i\in[1;n]}(\frac{K}{C\times R_i})$	A partir de la relation intuitive $\forall i \in [1;n], \bar{R}_i \leq \frac{K}{C\mu}$

TAB. 2 – Bornes pour une M/M/C/K

Dans cette situation il nous est apparu pratique de démarrer par une boucle sur C suivie d'une boucle sur K. Puis nous poursuivons avec une boucle sur μ. Une fois les paramètres C et K et μ fixés, on est en mesure de calculer pour chaque point de mesure i, la valeur exacte à donner à λ_i pour que le débit moyen en sortie du modèle $\bar{X}_{i,anal}$ soit égal à celui mesuré sur le système $\bar{X}_i$.

Une recherche orientée. Il serait possible d'échafauder une méthode plus rapide et plus « intelligente » de recherche du calibrage optimal en tirant profit de notre connaissance sur l'impact des paramètres vis-à-vis des courbes de performances. La méthode pourrait démarrer en fixant la valeur de C en fonction du rayon de courbure observé sur les mesures, puis de choisir μ tel que $C\mu$ soit à peu près égal au plus grand débit observé sur les mesures, ensuite de décider d'Off de façon à obtenir le temps de séjour mesuré à faible charge et enfin de fixer K en se rapportant au temps de réponse maximum mesuré. Ce type d'approche se révélera obligatoire pour les modèles imbriqués où le nombre de combinaisons possibles sur les paramètres est démentiel.

3.3.5 L'algorithme de recherche mis en oeuvre

Nous présentons ici la structure de la recherche exhaustive mise en oeuvre pour la brique M/M/C.

1. Boucle sur C depuis 1 jusqu'à ce que les augmentations successives sur C engendrent des résultats systématiquement plus mauvais.
2. Boucle sur K depuis la borne inférieure sur K jusqu'à ce que les augmentations successives sur K engendrent des résultats systématiquement plus mauvais.
3. Calcul de la borne inférieure sur μ pour les valeurs de C et de K à l'essai.
4. Boucle sur μ depuis la borne inférieure sur μ jusqu'à ce que les augmentations successives sur μ engendrent des résultats systématiquement plus mauvais. Le pas de recherche sur μ est fixé en tenant compte de l'ordre de grandeur des mesures et de la précision souhaitée sur les résultats (plus de détails à la page 31).

5. Calcul des λ associés aux $\bar{X}$ (cette étape fait appel à la propriété remarquable 2). On recherche par dichotomie la valeur à donner à chaque λ_i pour engendrer un débit moyen en sortie du modèle $\bar{X}_{i,anal}$ identique à celui mesuré $\bar{X}$ sur le système. Les explications détaillées relatives à cette étape sont présentées à la page 85.
6. Calcul du temps de séjour moyen des clients $\bar{R}_{i,anal}$ pour chaque niveau de charge λ_i.
7. Calcul de la fonction distance (se fiant aux écart entre $\bar{R}_{i,anal}$ et $\bar{R}_i$) pour le modèle à l'essai. Si le résultat du calcul situe le modèle parmi les 3 meilleurs modèles jusqu'à présent testés alors le modèle est retenu et remplace un modèle dans la liste provisoire des meilleurs modèles.

La recherche itérative peut déboucher sur un nombre considérable de modèles à être testés. Dans sa forme la plus générale, ce nombre atteint : $\sum_{C=1}^{C_{max}} \sum_{K=C}^{K_{max}} \frac{\mu_{max}-\mu_{min}}{precision_\mu}$. Notons que les valeurs de K_{max}, C_{max} et de μ_{max} ne sont pas connues « à priori » : elles se déterminent empiriquement [12]. Par conséquent il n'est pas possible, avec notre approche de recherche, de prédire le temps de calcul de notre recherche.

[12] Pour μ, l'optimisation par les μ locaux permettra d'obtenir une borne supérireure, page 86

4 FERME

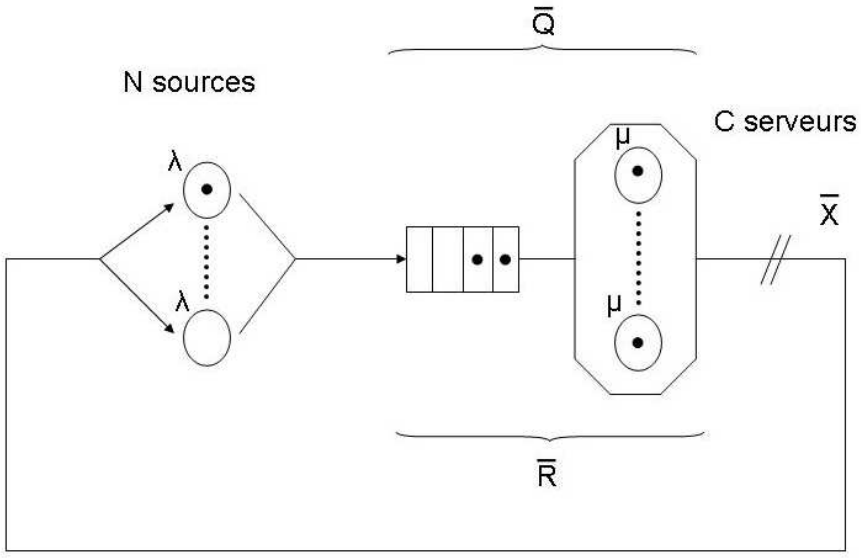


FIG. 19 – Représentation graphique d'un modèle FERME

4.1 Généralités

- Modèle fermé à capacité infinie (ou bien de taille telle que les clients trouvent toujours une place disponible dans la file).
- Toujours stable.
- Paramètres intrinsèques au modèle : N sources, C serveurs, une loi de service exponentielle de taux μ.
- Paramètre de charge : un processus d'arrivée des clients selon une loi de Poisson de taux λ.

4.2 Analyse du régime permament

4.2.1 Les probabilités d'états stationnaires

Soit $p(n)$ la probabilité d'avoir n clients dans la file.

$$\begin{cases} p(n) = \begin{cases} \frac{\rho}{n} \times (N-n+1) \times p(n-1) = \frac{\rho^n}{n!} \times \frac{N!}{(N-n)!} \times p(0) & \text{si n} \in [1\,;\text{C}] \\ \frac{\rho}{C} \times (N-n+1) \times p(n-1) = \frac{\rho^n}{C!C^{n-C}} \times \frac{N!}{(N-n)!} \times p(0) & \text{si n} \in [\text{C}+1\,;\text{N}] \end{cases} \\ \text{avec } \sum_{n=0}^{N} p(n) = 1 \end{cases}$$

4.2.2 Les paramètres de performance

DÉBIT :	$\bar{X}_{anal} = \sum_{n=1}^{C-1} p(n)n\mu + \sum_{n=C}^{N} p(n)C\mu$	
NOMBRE DE CLIENTS :	$\bar{Q}_{anal} = \sum_{n=1}^{N} p(n)n$	
TEMPS DE SÉJOUR :	$\bar{R}_{anal} = \frac{\bar{Q}_{anal}}{\bar{X}_{anal}}$	*Loi de Little*

4.3 Recherche du calibrage *ad hoc*

4.3.1 L'allure des courbes de performance

Les modèles FERME dispensent également des courbes de performance monotones et bimodales.

On observe que pour de faibles charges, le temps moyen de réponse $\bar{R}_{anal}$ varie peu et reste très près de la valeur du temps de service moyen d'un client $\frac{1}{\mu}$. Cette observation résulte du service quasi-immédiat des clients entrant dans la file à faible charge.

On remarque la présence d'un point d'accumulation vers lequel se regroupent les performances en débit et en temps de réponse du modèle pour des niveaux de charge λ élevés. Ce point d'accumulation prend place aux coordonnées $(C\mu; \frac{N}{C\mu})$. La justification des coordonnées de ce point d'accumulation est relativement simple. Son abscisse correspond à la valeur maximale de débit moyen en sortie. Cette situation intervient lorsque la charge λ est très élevée et que donc la file sature en permamence. Elle débite alors en sortie à son taux maximum, soit $C\mu$. Quant à son ordonnée, elle traduit le temps de séjour moyen maximum d'un client dans le modèle. Ce temps de séjour sera à son maximum lorsque la file sature complètement. Le client entrant est alors systématiquement accueilli en dernière position de la file ce qui signifie qu'il y a en moyenne N clients dans la file. La loi de Little nous donne alors très simplement l'expression du temps de séjour moyen : $\bar{R} = \frac{N}{C\mu}$.

Comme pour la M/M/C et pour des raisons similaires (page 35), le nombre de serveurs modifie la courbure de la courbe. Plus ce nombre sera grand, plus le rayon de courbure sera serré. Et inversement plus C est petit, plus la courbure devient droite.

La liste ci-dessous et la figure 20 récapitulent l'impact des paramètres dans ce contexte.

* Le quotient $\frac{1}{\mu}$ (sommée éventuellement à Off) détermine la valeur du temps de réponse à faible charge
* Le produit $C\mu$ spécifie le débit maximal supporté par le modèle et participe à situer le point d'accumulation de la courbe
* La valeur de C règle le rayon de courbure de la courbe de performances
* La valeur de N situe le point d'accumulation sur la courbe en imposant comme borne supérieure sur le temps de réponse : $\frac{N}{C\mu}$

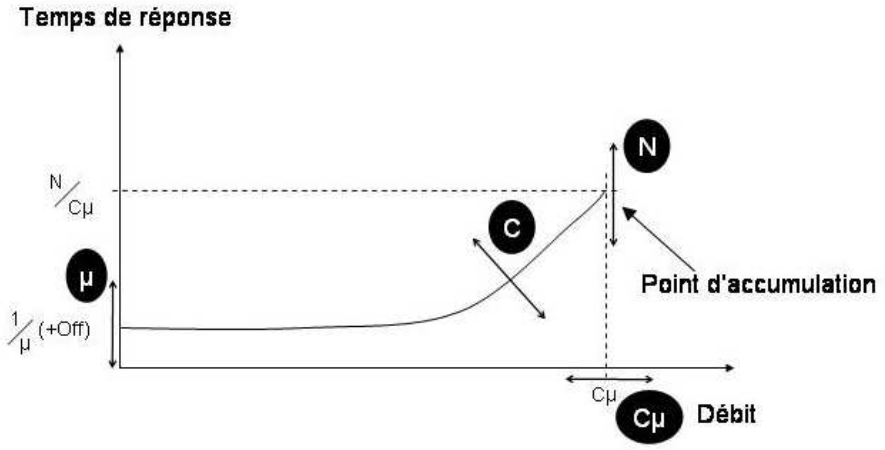


FIG. 20 – Influence des paramètres d'un modèle FERME sur ses performances

4.3.2 Quelques propriétés remarquables

1. $\bar{X} \leq \lambda$
2. La loi de conservation des jetons s'exprime comme :

$$\begin{aligned} & (N - \bar{Q}) \times \lambda = \bar{X} \\ \Leftrightarrow \quad & \lambda = \frac{\bar{X}}{N - \bar{Q}} \\ \Leftrightarrow \quad & \lambda = \frac{\bar{X}}{N - \bar{R} \times \bar{X}} \end{aligned}$$

3. Plus λ croît, plus le débit en sortie du modèle augmente et plus le temps de réponse s'allonge jusqu'à leur valeur maximale respective : $C\mu$ et $\frac{N}{C\mu}$.
4. Plus μ diminue, plus le temps de séjour s'allonge et inversement en augmentant μ, on réduit le temps de séjour.

4.3.3 Les bornes sur les paramètres

Type de borne	*Expression littérale*	*Signification*
Inférieure sur N	$\max_{i\in[1;n]}(R_i \times X_i)$	Loi de Little
Supérieure sur C	N	Si $C \geq N$ alors tous les clients séjourneraient en moyenne un temps identique dans le modèle égal au temps moyen de service $\frac{1}{\mu}$
Inférieure sur μ	$\max_{i\in[1;n]}(\frac{X_i}{C})$	$\bar{X}_{mes}$ ne peut dépasser $C\mu$
Inférieure sur μ	$\max_{i\in[1;n]}(\frac{1}{R_i})$	R_{mes} est toujours plus grand que le temps moyen de service seul $\frac{1}{\mu}$
Supérieure sur μ	$\min_{i\in[1;n]}(\frac{N-1}{C\times R_i})$	A partir de la relation intuitive que pour tout i ∈ [1 ;n], $\bar{R}_i \leq \frac{N}{C\mu}$

TAB. 3 – Bornes pour un modèle FERME

4.4 Le processus de recherche du calibrage adéquat/optimal

Une recherche systématique. Toujours pour les mêmes raisons (énoncées 30), nous avons opté, pour commencer, par une recherche exhaustive par itération. La démarche consiste à parcourir itérativement toutes les combinaisons autorisées des paramètres. Dans cette situation il semble pratique de démarrer par une boucle sur N, puis C et puis de terminer par une boucle sur μ. Au terme de ces trois étapes, nous calculons pour chaque point de charge λ_i, les performances $\bar{X}_i$ et $\bar{R}_i$ du modèle considéré.

Une recherche orientée. Ici aussi il serait possible de mener une recherche plus rapide en orientant la recherche du calibrage. Le procédé serait identique à celui exposé pour une M/M/C/K à la seule différence que le paramètre K serait remplacé par le nombre de clients du modèle FERME N.

4.5 L'algorithme de recherche mis en oeuvre

Nous présentons ic la structure de la recherche exhaustive mise en oeuvre pour la brique FERME.

1. Boucle sur C depuis 1 jusqu'à ce que les augmentations successives sur C engendrent des résultats systématiquement plus mauvais.
2. Boucle sur N depuis la borne inférieure sur N jusqu'à ce que les augmentations successives sur N engendrent des résultats systématiquement plus mauvais.
3. Calcul de la borne inférieure sur μ pour les valeurs de C et de N à l'essai.
4. Boucle sur μ depuis la borne inférieure sur μ jusqu'à ce que les augmentations successives sur μ engendrent des résultats systématiquement plus mauvais. Le pas de recherche sur μ est fixé en tenant compte de l'ordre de grandeur des mesures et de la précision souhaitée sur les résultats (plus de détails à la page 31).
5. Calcul des λ associés aux $\bar{X}$ (cette étape fait appel à la propriété remarquable 2). On recherche par dichotomie la valeur à donner à chaque λ_i pour engendrer un débit moyen en sortie du modèle $\bar{X}_{i,anal}$ identique à celui mesuré $\bar{X}$ sur le système.
6. Calcul du temps de séjour moyen des clients $\bar{R}_{i,anal}$ pour chaque niveau de charge λ_i.
7. Calcul de la fonction distance (se fiant aux écart entre $\bar{R}_{i,anal}$ et $\bar{R}_i$) pour le modèle à l'essai. Si le résultat du calcul situe le modèle parmi les 3 meilleurs modèles jusqu'à présent testés alors le modèle est retenu et remplace un modèle dans la liste provisoire des meilleurs modèles.

La recherche itérative peut déboucher sur un nombre considérable de modèles à tester. Au plus ce nombre peut atteindre : $\sum_{N=C+1}^{N_{max}} \sum_{C=1}^{C_{max}} \frac{\mu_{max}-\mu_{min}}{precision_\mu}$. Notons que les valeurs de N_{max}, C_{max} et de μ_{max} ne sont pas connues « à priori » : elles se déterminent empiriquement [13]. Par conséquent il n'est pas possible, avec notre approche de recherche, de prédire le temps de calcul de notre recherche.

[13] Pour μ, l'optimisation par les μ locaux permettra d'obtenir une borne supérireure, page 86

5 FERME avec rejet

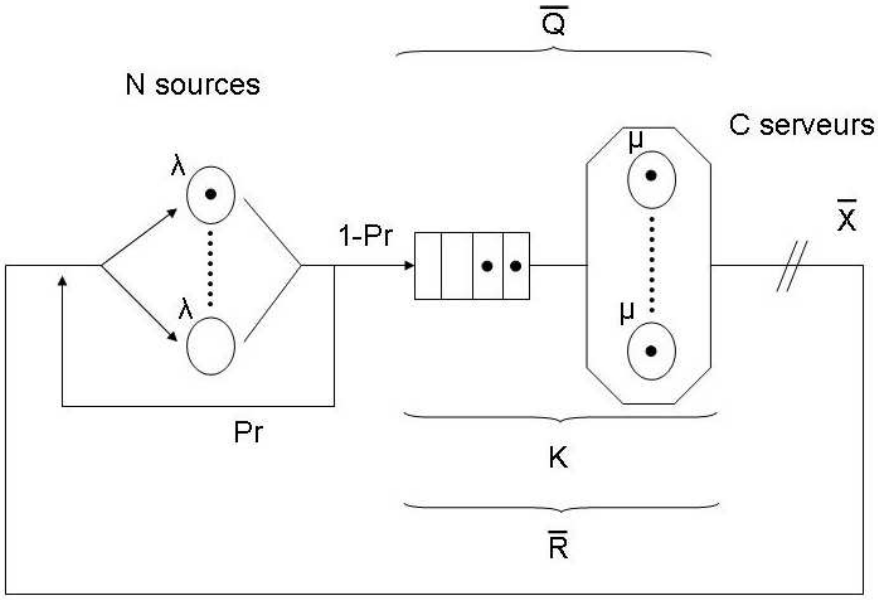


FIG. 21 – Représentation graphique d'un modèle FERME avec rejet

5.1 Généralités

- Modèle fermé à capacité finie. La taille de la file est inférieure au nombre de sources.
- Taux de rejet, les clients peuvent ou bien trouvé une place ou bien être rejetés en entré de la file.
- Toujours stable.
- Paramètres intrinsèques au modèle : N sources, C serveurs, une capacité K et une loi de service exponentielle de taux μ.
- Paramètre de charge : un processus d'arrivée des clients selon une loi de Poisson de taux λ.

5.2 Analyse du régime permament

5.2.1 Les probabilités d'états stationnaires

Soit $p(n)$ la probabilité d'avoir n clients dans la file.

$$\begin{cases} p(n) = \begin{cases} \frac{\rho}{n} \times (N-n+1) \times p(n-1) = \frac{\rho^n}{n!} \times \frac{N!}{(N-n)!} \times p(0) & \text{si n} \in [1\,;\text{C}] \\ \frac{\rho}{C} \times (N-n+1) \times p(n-1) = \frac{\rho^n}{C!C^{n-C}} \times \frac{N!}{(N-n)!} \times p(0) & \text{si n} \in [\text{C+1}\,;\text{K}] \end{cases} \\ \text{avec } \sum_{n=0}^{K} p(n) = 1 \end{cases}$$

5.2.2 Les paramètres de performance

DÉBIT :	$\bar{X}_{anal} = \sum_{n=1}^{C-1} p(n)n\mu + \sum_{n=C}^{K} p(n)C\mu$	
NOMBRE DE CLIENTS :	$\bar{Q}_{anal} = \sum_{n=1}^{K} p(n)n$	
TEMPS DE SÉJOUR :	$\bar{R}_{anal} = \frac{\bar{Q}_{anal}}{\bar{X}_{anal}}$	*Loi de Little*

5.3 Recherche du calibrage *ad hoc*

5.3.1 L'allure des courbes de performance

Les modèles FERME avec rejet dispensent également des courbes de performance monotones et bimodales.

On observe que pour de faibles charges, le temps moyen de réponse $\bar{R}_{anal}$ varie peu et reste très près de la valeur du temps de service moyen d'un client $\frac{1}{\mu}$. Cette observation résulte du service quasi-immédiat des clients entrant dans la file à faible charge.

On remarque la présence d'un point d'accumulation vers lequel se regroupent les performances en débit et en temps de réponse du modèle pour des niveaux de charge λ élevés. Ce point d'accumulation prend place aux coordonnées $(C\mu; \frac{K}{C\mu})$. La justification des coordonnées de ce point d'accumulation est relativement simple. Son abscisse correspond à la valeur maximale de débit moyen en sortie. Cette situation intervient lorsque la charge λ est très élevée et que donc la file sature en permamence. Elle débite alors en sortie à son taux maximum, soit $C\mu$. Quant à son ordonnée, elle traduit le temps de séjour moyen maximum d'un client dans le modèle. Ce temps de séjour sera à son maximum lorsque la file sature complètement. Le client entrant est alors systématiquement accueilli en dernière position de la file ce qui signifie qu'il y a en moyenne K clients dans la file. La loi de Little nous donne alors très simplement l'expression du temps de séjour moyen : $\bar{R} = \frac{K}{C\mu}$.

Comme pour la M/M/C et pour des raisons similaires (page 35), le nombre de serveurs modifie la courbure de la courbe. Plus ce nombre sera grand, plus le rayon de courbure sera serré. Et inversement plus C est petit, plus la courbure devient droite.

Enfon on a remarqué à travers des exemples que le paramètre N permet également de modifier la courbure de la courbe de performances. Plus la valeur de N est petite, plus le rayon de courbure de la courbe est serré. On peut justifier intuitivement cette observation de la façon suivante. Pour des charges « moyennes »(associées à des débits de l'ordre de $\frac{C\mu}{2}$), les clients ont plus de chance de rentrer en queue de file pour le modèle comportant le plus grand nombre de sources, et ce bien que le débit moyen soit le même. Par conséquent c'est le temps de réponse de la file ayant le plus de jetons qui est le plus long. Cette différence s'estompe à faible charge où les clients ne subissent pratiquement plus d'attente (temps de séjour quasi-constant) et à charge élévée parce qu'alors les clients, quelque soit le nombre de sources considéré, rentrent toujours dans la file en dernière position (à la position K).

La liste ci-dessous et la figure 22 récapitulent l'impact des paramètres dans ce contexte.

* Le quotient $\frac{1}{\mu}$ (sommée éventuellement à Off) détermine la valeur du temps de réponse à faible charge
* Le produit $C\mu$ spécifie le débit maximal supporté par le modèle et participe à situer le point d'accumulation de la courbe
* La valeur de C règle le rayon de courbure de la courbe de performances
* La valeur de K situe le point d'accumulation sur la courbe en imposant comme borne supérieure sur le temps de réponse : $\frac{K}{C\mu}$
* La valeur de N permet également de régler le rayon de courbure des de courbes de performances.

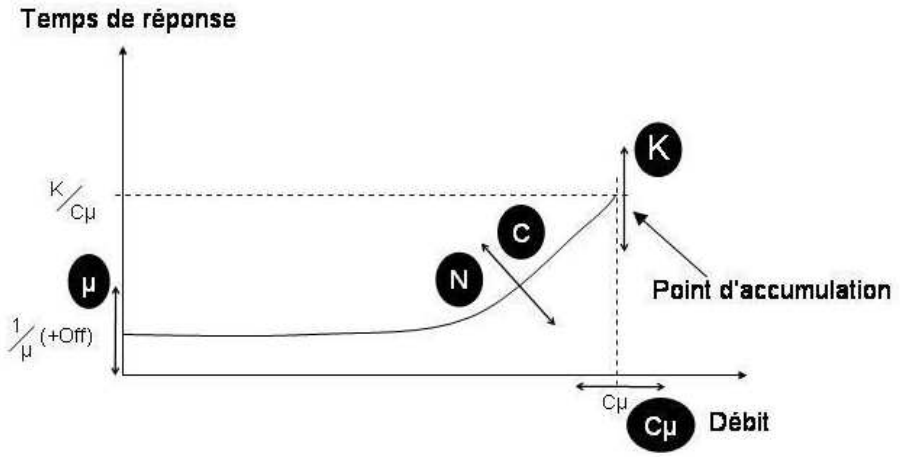


FIG. 22 – Influence des paramètres d'un modèle FERME avec rejet sur ses performances

5.3.2 Quelques propriétés remarquables

1. $\bar{X} \leq \lambda$
2. La loi de conservation des jetons s'exprime comme :

$$\begin{aligned} & (K - \bar{Q}) \times \lambda = \bar{X} \\ \Leftrightarrow \quad & \lambda = \frac{\bar{X}}{K - \bar{Q}} \\ \Leftrightarrow \quad & \lambda = \frac{\bar{X}}{K - R \times \bar{X}} \end{aligned}$$

3. Plus λ croît, plus le débit en sortie du modèle augmente et plus le temps de réponse s'allonge jusqu'à atteindre leur valeur maximale respective $C\mu$ et $\frac{K}{C\mu}$.
4. Plus μ diminue, plus le temps de séjour s'allonge et inversement en augmentant μ, on réduit le temps de séjour.

5.3.3 Les bornes sur les paramètres

Type de borne	*Expression littérale*	*Signification*
Inférieure sur N	$\max_{i\in[1;n]}(R_i \times X_i)$	La loi de Little
Supérieure sur C	N	Si $C \geq N$ alors tous les clients séjourneraient en moyenne un temps identique dans le modèle égal au temps moyen de service $\frac{1}{\mu}$
Inférieure sur K	C	
Supérieure sur K	N	
Inférieure sur μ	$\max_{i\in[1;n]}(\frac{X_i}{C})$	$\bar{X}_{mes}$ ne peut dépasser $C\mu$
Inférieure sur μ	$\max_{i\in[1;n]}(\frac{1}{R_i})$	R_{mes} est toujours plus grand que le temps moyen de service seul $\frac{1}{\mu}$
Supérieure sur μ	$\min_{i\in[1;n]}(\frac{K-1}{C\times R_i})$	A partir de la relation intuitive que pour tout i $\in$ [1 ;n], $\bar{R}_i \leq \frac{K}{C\mu}$

TAB. 4 – Bornes pour un modèle FERME avec rejet

5.3.4 De l'utilité de cette brique

L'étude des courbes de performance réalisée précédemment (page 49) a mis en évidence que et C et N permettent de modifier le rayon de courbure d'une M/M/C. Leur impact n'est probablement pas exactement identique mais cette « redondance » nous est inutile et constitue donc un handicap pour cette brique. A performances égales, on préfèrera toujours le modèle avec le moins de paramètres lorsque deux modèles sont en compétition. Ainsi on préfèrera faire appel à un modèle FERME (3 paramètres intrinsèques) plutôt qu'à un modèle FERME rejet (4 paramètres intrinsèques) si leur capacité à reproduire un système est identique. Or nous nous sommes aperçus qu'il est toujours possible d'approcher avec une précision quasi-arbitraire les performances d'un modèle FERME avec rejet ($\bar{R}$ selon $\bar{X}$) par un modèle FERME tout court. C'est pourquoi, nous avons jugé inutile de prendre en compte la brique FERME avec rejet pour les résultats exposés dans la partie suivante.
En revanche le modèle FERME avec rejet pourrait nous être très utile dans certains cas. Il permet sans aucun doute de calibrer également le taux de rejet des clients $Pr.$ du modèle. Or ce paramètre de performances n'est pas pris en compte dans le contexte restreint de ce stage : on se contente de reproduire le comportement de $\bar{R}$ selon $\bar{X}$. Voilà pourquoi nous l'avons provisoirement évincée.

6 Modèles Imbriqués

6.1 Généralités

Les modèles imbriqués constituent le deuxième type de combinaisons de briques de base étudié dans ce rapport. La première combinaison référait à la mise en série d'une M/M/∞. Comme expliqué précédemment (page 26), les modèles imbriqués résultent du besoin de construire automatiquement des modèles à temps de service variable en fonction de la charge soumise en entrée, une propriété inaccessible aux briques de base. A nouveau, nous avons opté parmi tous les familles de modèles qui répondaient le plus à notre besoin, pour la brique qui paraît la plus simple et dont la logique s'interprète facilement.
Globablement la solution adoptée met en jeu deux modèles markoviens, consituant deux blocs, tous deux soumis à la même charge λ. On verra plus loin pourquoi cette hypothèse revient dans notre configuration à supposer simplement que les charges des deux blocs sont proportionnelles. La relation entre les deux blocs de ce modèle est que le temps d'attente dans l'un des deux (que nous désignerons comme étant le bloc interne) conditionne le temps de service de l'autre bloc (le bloc principal). Enfin, les performances en débit et en temps de séjour à confronter aux séries de mesure seront celles engendréés par le modèle principal. Il est clair que cette implémentation garantit un modèle à temps de service variable dans lequel plus la charge augmente, plus le temps de service du bloc principal est important.

6.2 Solution adoptée - Un temps de service variable selon la charge

Nous avons opté pour un bloc principal modélisé par une M/M/C et composé de C_1 serveurs, tous de taux de service μ_1 variable (selon la charge). Il s'agit donc d'un modèle ouvert de capacité illimitée dont le régime permament est soumis à une condition de stabilité : $\lambda_1 < C_1\mu_1$

Pour le bloc interne, nous avons décidé d'une M/M/1 de taux de service μ_2 (fixe). Ce modèle également ouvert et de capacité illimité est monoserveur et doit satisfaire la condition de stabilité $\lambda < \mu_2$ pour permettre l'analyse de son régime permanent.
De façon très sommaire (on y reviendra en détails un peu plus tard), le temps moyen de service d'un client μ_1 du modèle 1 est fonction de la charge λ en entrée du modèle et plus précisément du temps d'attente pour ce niveau de charge dans le bloc interne.

6.2.1 Terminologie des paramètres et des variables mis en jeu

Le bloc principal, auquel on affiliera les indices en 1, dispose d'un temps de service variable S_1 mais qui contient une partie irréductible nommée S_0. Quant au bloc interne, auquel se rattache les indices en 2, son taux de service μ_2 est fixe.

Le paramètre de sensibilité a, détermine le niveau de corrélation entre les deux modèles c'est-à-dire le degré d'influence du temps d'attente dans le bloc interne W_2 , sur le temps de service du bloc principal S_1. Plus a est élevé, plus W_2 sera pris en considération dans le calcul de S_1. Inversement, attribuer à a

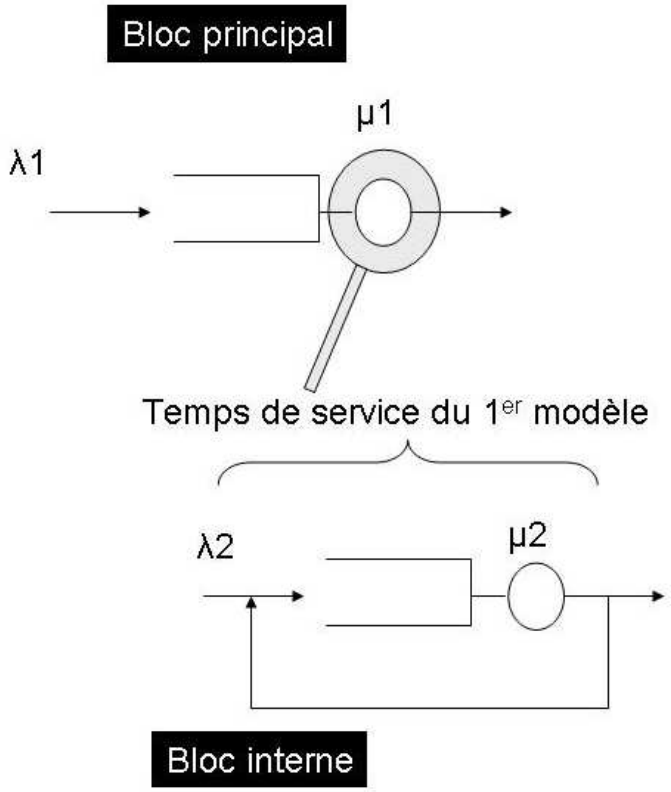


FIG. 23 – Architecture des modèles imbriqués

une valeur très basse conduira à un modèle à temps de service pratiquement constant.

A présent, nous associons à des variables les points-clefs (pour notre analyse) d'une courbe de points de mesure que l'on a choisie représentative du comportement à reproduire avec les modèles imbriqués. Cette courbe pourrait correspondre aux jeux de mesure 2 et 4. Pour clarifier notre analyse, on supposera que λ_a [14] correspond à une faible charge et λ_b à une charge élevée (proche de la saturation), niveaux de charges auxquels correspondent respectivement les temps de réponse R_a et R_b comme l'illustre la figure 24.

6.2.2 Justification et interprétation des hypothèses fondatrices

On suppose que plusieurs blocs principaux identiques, mettons N, sont branchés sur le bloc interne ce qui dénote un accès compétitif aux ressources du bloc 2, et fait de ce bloc un éventuel goulot d'étranglement. On fait l'hypothèse que tous les blocs principaux recoivent un trafic Poissonien de requêtes de même intensité λ_1 et relaient ces requêtes à destination du bloc 2 (cela traduit le fait que les augmentations de charge s'effectuent simultanement sur tous les blocs et dans les mêmes proportions). Ce dernier doit donc faire face à une somme

[14] on peut parler de λ ou de $\bar{X}$ indifférement puisque les modèles mis en jeu sont stables sans perte

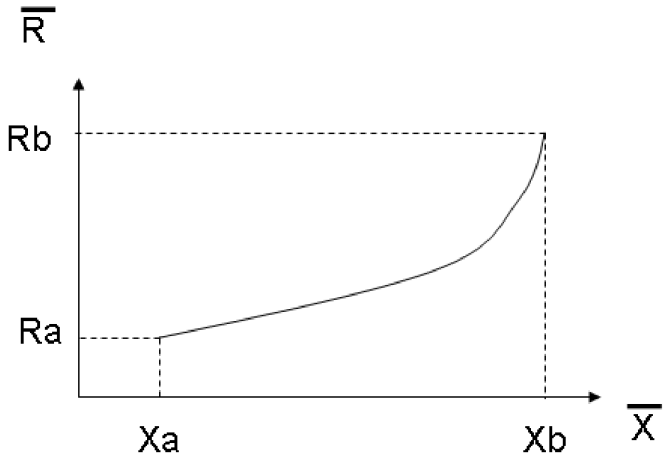


FIG. 24 – Allure de la courbe de performances des systèmes à temps de service variable

de N trafics, que l'on suppose Poissionniens, qui peuvent être vus comme un seul trafic Poissonnien d'intensité $N\lambda_1$. Or que le bloc interne soit soumis à un trafic $N\lambda_1$ avec un taux de service μ_2 ou bienqu'il soit soumis à un trafic λ_1 avec un taux de service $\frac{\mu_2}{N}$, le temps d'attente moyen (et pas de service) d'un de ses clients sera le même (en régime permament ce temps dépend seulement du quotient $\frac{\lambda}{\mu}$). C'est pourquoi on pourra supposer que le bloc interne étudié reçoit en entrée un trafic de taux λ_1 et plus généralement qu'il existe un trafic d'entrée unique de taux λ pour tous les blocs.
Pour a, son interprétation est simple, il correspond au nombre moyen d'accès au bloc interne pour traiter une requête du bloc principal.
Enfin S_0 peut être perçu comme le délai incompressible dû au traitement d'un client par le bloc principal dans les conditions les plus favorables (lorsque le client se retrouve seul dans le bloc interne, sans attente).

6.3 Dialectique des modèles imbriqués

Le lien qui existe entre les deux blocs qui composent un modèle imbriqué s'exprime à travers le calcul du temps de service S_1 du modèle 1. Son expression arithmétique est la suivante :

Temps de service moyen d'un client du bloc principal : $S_1 = S_0 + aW_2$

$$\text{d'où} \quad \mu_1 = \frac{1}{S_1} = \frac{1}{S_0 + aW_2} \tag{2}$$

La relation causale entre le bloc interne et les performances du bloc principal transparaît clairement à travers la formule littérale (2).

6.4 Calcul des paramètres de performance

Le calcul des paramètres de performance démarre par l'analyse du bloc interne. Pour un niveau de charge λ, on détermine le temps d'attente dans la file W_2. Les connaissances sur la M/M/1 permettent d'exprimer simplement ce temps d'attente.

DÉBIT :	$\bar{X}_{1,anal} = \lambda$
du bloc principal	
TEMPS D'ATTENTE :	$\bar{W}_2 = \frac{1}{\mu_2 - \lambda} - \frac{1}{\mu_2} = \frac{\lambda}{\mu_2(\mu_2 - \lambda)}$
dans le bloc interne	
TEMPS DE SÉJOUR :	$\bar{R}_{1,anal} = \frac{1}{\mu_1} + \sum_{n=C_1}^{\infty} p(n) \frac{n+1-C_1}{C_1 \mu_1}$
dans le bloc principal	

Une fois, W_2 obtenu, on infère μ_1 à partir de la relation 2 la valeur de μ_1. A présent pour obtenir les performances du modèle imbriqué étudié pour une charge égale à λ, il suffit de calculer les performances d'une M/M/C classique avec C_1 serveurs de taux μ_1 pour une charge égale à λ. On notera que puisqu'il s'agit d'un modèle stable et sans perte, le calcul du débit moyen est automatique : $\bar{X}_{1,anal} = \lambda$.

6.5 Quelques propriétés remarquables

On constate que :

1. $\bar{X}_1 = \lambda$
2. Plus λ croît, plus μ_1 diminue et donc plus généralement, plus la limite théorique de saturation $C\mu_1$ est réduite
3. Que le débit maximal autorisé est majoré par $\min(\mu_2; C_1 \times \frac{1}{S_0})$. $\bar{X} < \mu_2$ pour satisfaire la condition de stabilité du bloc interne. Et puisque μ_1 sera toujours majoré par $\frac{1}{S_0}$, $\bar{X} < \frac{1}{S_0}$ est nécessaire (mais pas suffisant !) pour assurer la condition de stabilité du bloc principal
4. Plus λ croît, plus le débit augmente et se rapproche de sa limite théorique $C_1\mu_1$ et parallèlement plus le temps de réponse s'allonge en tendant vers des valeurs infinies.

6.6 Les bornes sur les paramètres

Le tableau 6.6 référence les bornes qui permettent de circonscrire les intervalles de valeurs sur les paramètres du modèle.

6.7 Processus de recherche du calibrage *ad hoc*

Pour cette brique, la démarche employée pour rechercher le calibrage optimal marque une rupture nette avec les approches utilisées jusqu'ici. En effet la présence de 3 paramètres continus prévient une approche exhaustive par itération successive. C'est pourquoi on a recours à une méthodologie plus ciblée et plus spécifique au comportement de la brique. Le procédé de calibrage des paramètres fait appel à notre compréhension du modèle et requiert plus d'intuition sur le jeu de paramètres. Voici présentées ci-dessous les valeurs théoriques qu'on associera aux paramètres pour obtenir le calibrage souhaité.
On s'appuie sur la description et sur la terminologie de la série de mesures faites auparavant. Ici ϵ représente un réel positif très proche de 0.
On notera par ai

1. On choisit μ_2 tel que $\mu_2 = \lambda_b + \epsilon$ de manière à assurer la stabilité du bloc interne.

Type de borne	*Expression littérale*	*Signification*
Inférieure sur μ_2	$\max_{i\in[1;n]}(X_i)$	Traduction de la condition de stabilité sur le bloc interne
Inférieure sur μ_2	$\max_{i\in[1;n]}(\frac{1}{\bar{R}_i})$	Le temps moyen de séjour d'un client est toujours plus grand que le temps moyen de service seul $\Leftrightarrow \forall i \in [1;n], \bar{R}_i \geq \frac{1}{\mu}$
Inférieure sur S_0	$\min_{i\in[1;n]}(\frac{C_1}{\bar{X}_i})$	La condition de stabilité sur le bloc principal $\Leftrightarrow \forall i \in [1;n], C_1\mu_1 > \bar{X}_i$ Or $\forall \lambda, \mu_1 \leq \frac{1}{S_0}$. Ainsi il est nécessaire (mais pas suffisant) que : $\forall i \in [1;n], S_0 < \frac{C_1}{\bar{X}_i}$
Supérieure sur a	$(\frac{C_1}{\lambda_b} - S_0) \times \frac{\mu_2(\mu_2-\lambda_b)}{\lambda_b}$	A nouveau, la condition de stabilité sur le bloc principal $\Leftrightarrow \forall i \in [1;n], \lambda_i < C_1 \times \frac{1}{S_0+a\times(\frac{\lambda_i}{(\mu_2-\lambda_i)\mu_2})}$ $\Leftrightarrow \forall i \in [1;n], a < (\frac{C_1}{\bar{X}_i} - S_0) \times \frac{\mu_2(\mu_2-\bar{X}_i)}{\bar{X}_i}$ car $\forall i \in [1,n](\lambda_i = \bar{X}_i)$

TAB. 5 – Bornes pour une M/M/1 imbriquée dans une M/M/C

2. On fixe $S_0 = \bar{R}_a - \epsilon$ ce qui assure que les clients du bloc principal ont un temps de séjour proche et légèrement inférieur à $\bar{R}_a$ à faible charge.
3. A présent, on décide d'une valeur de C_1 de manière à assurer la stabilité du bloc principal dans le cas de la charge mesurée la plus élevée. Autrement dit, C_1 tel que $C_1\mu_1 > \lambda_b \Leftrightarrow C_1 > \frac{\lambda_b}{S_0+a\times\frac{\lambda_b}{\mu_2(\mu_2-\lambda_b)}}$
4. Reste le paramètre de sensibilité a à calibrer. a détermine le degré de variabilité du temps de service du bloc principal. L'objectif des modèles imbriqués est de reproduire certains comportements de systèmes pour lesquels l'engorgement intervient de manière plus progressive. Ainsi la valeur de a règle la pente de la courbe de performance avant que la saturation intervienne pleinement. On initie la recherche de la valeur de a par une valeur arbitraire (qui maintient la condition de stabilité du bloc 1). On calcule μ_1 pour chaque point de mesure λ_i. S'il s'avère que la courbe du modèle appraît en dessous de celle engendrée par les mesurées, alors on augmente la valeur de a ce qui aura pour effet de diminuer le taux de service μ_1 et de réhausser la courbe de performances du modèle considéré. En revanche si la courbe du modèle majore celle des mesures, alors on diminuera la valeur de a afin d'accroître μ_1 et donc de réduire les temps de réponse générés par le modèle imbriqué. Le calibrage de a peut nécessiter plusieurs essais mais se réalise efficacement par dichotomie avec une précision arbitraire.

6.8 L'algorithme de recherche mis en oeuvre

On va relâcher les valeurs strictes décidées précédemment pour le jeu de paramètres (autrement dit, autoriser une variabilité sur ϵ) et parcourir itérativement

ces combinaisons de valeurs. Plus précisémment, l'approche suivante :

1. Pour tout i dans $[1, n]$, $\lambda_i = \bar{X}_i$ (aucune perte, aucun rejet)
2. Calcul de la borne inférieure sur S_0
3. Boucle sur S_0 dans l'intervalle $[R_a - 2\epsilon; R_a - \epsilon]$
4. Calcul de la borne inférieure sur μ_2
5. Boucle sur μ_2 dans l'intervalle $[\lambda_b + \epsilon; \lambda_b + 2\epsilon]$
6. Calcul du temps d'attente moyen dans le bloc interne W_2 pour chaque niveau de charge λ_i
7. Puis calcul du temps moyen de service μ_1 du bloc principal pour chacun des niveaux de charge λ_i
8. Calcul de la borne inférieure sur C_1
9. Boucle sur C_1 dans l'intervalle $[S_0\lambda_b; ???$ [15]$]$ (en effet au plus μ_1 vaut $\frac{1}{S_0}$)
10. Pour a, deux approches possibles :
    * Boucle sur a dans l'intervalle $[0\,; (\frac{C_1}{\lambda_b} - S_0)(\frac{\mu_2 \times (\mu_2 - \lambda_b)}{\lambda_b})]$. Le a_{max} correspond à la valeur maximale que peut atteindre a tout en conservant la stabilité du bloc 1.
    * Recherche directe du a tel que $\bar{R}_b = \bar{R}_{b,anal}$. On vérifie d'abord l'existence de la solution (Pour a=0, $R_{b,anal}$ engendré est il bien inférieur à R_b ? Si ce n'est pas le cas, alors quelque soit la valeur de a il sera impossible d'obtenir un $R_{b,anal}$ égal à R_b car l'accroissement de a induit inmanquablement une augmentation du $R_{b,anal}$). Si c'est bien le cas, on recherche la valeur exacte de a permettant d'approcher aussi finement que l'on souhaite $R_{b,anal}$ et R_b par dichotomie.
11. Calcul du temps de séjour moyen des clients $\bar{R}_{i,anal}$ pour chaque niveau de charge λ_i.
12. Calcul de la fonction distance (se fiant aux écart entre $\bar{R}_{i,anal}$ et $\bar{R}_i$) pour le modèle à l'essai. Si le résultat du calcul situe le modèle parmi les 3 meilleurs modèles jusqu'à présent testés alors le modèle est retenu et remplace un modèle dans la liste provisoire des meilleurs modèles.

L'ajustement de a par la deuxième méthode permet au modèle à l'étude de coincider parfaitement au point de mesures $(\bar{X}_b, \bar{R}_b)$ sans avoir à parcourir toutes les valeurs possibles de a. Cependant le comportement du modèle entre les points de charge λ_a et λ_b demeure globalement assez peu maitrisable et imprévisible.

6.9 Limites du modèle et améliorations à apporter

On distingue principalement deux faiblesses à l'approche présentée des modèles imbriqués.

- On a supposé que λ_a et λ_b correspondent respectivement à des charges basses et élevées. Or il se peut (et se sera probablement souvent le cas) que l'on n'ait aucune idée sur le degré de charge relatif aux points de fonctionnement mesuré. Si c'est le cas, il y a toutes les chances que notre calibrage qui s'appuie sur cette hypothèse invalide soit mauvais.

[15] Absence de bornes statiques pour C_1

- Les résultats des tests, exposés dans la partie suivante de ce rapport, montreront que les valeurs trouvés par le calibrage pour C_1 n'ont pas de significations physiques : la valeur de C_1 est seulement régie par le besoin d'assurer la stabilité du modèle pour les charges élevées. A ce propos, soulignons que beaucoup d'autres combinaisons sont possibles pour les modèles imbriqués et que l'on espère, comme nous l'expliquerons dans la dernière partie de ce rapport, que l'utilisation d'un modèle limité ou fermé pour le bloc interne permettra de remédier au problème sur C_1.

7 Modèles multiclasses

Cette classe de modèles est la plus récente à avoir été intégrée etil est probable que sa forme actuelle évolue vers une forme plus aboutie, plus adaptée à nos besoins. Ce sont les expérimentations et l'analyse de leurs résultats qui guideront cette évolution. Cependant, même en cas de changement, l'essence du modèle demeurera très proche de celle présentée ici : seules quelques paramètres supplémentaires remplaceront certains constantes dans l'implémentation actuelles.

7.1 Généralités - Une M/M/1 multiclasse HOL

Le but pressenti pour les modèles multiclasses est de pouvoir reproduire les courbes de performances décroissantes constatées dans certains domaines. Ce comportement fréquent dans les systèmes fonctionnant sur TCP et que l'on retrouve également dans les systèmes de transport de données sur trame ne peut être reproduit par aucune brique de base [16]. Ainsi cette brique initie peut être un sous-ensemble de briques qui devra être élargi et remodelé afin de constituer une grammaire compacte et complète sur les courbes de performance décroissantes.

Par convention on supposera que plus l'indice d'une classe est élevé, plus sa priorité est importante.

7.1.1 Hypothèses fondatrices

Modèle ouvert, sans perte soumis à une condition de stabilité.
On suppose que seulement deux classes de clients (d'indice 1 et 2) rivalisent pour l'accès aux ressources. Suivant notre convention, les clients de la classe 2 ont priorité non préemptive sur ceux de la classe 1.

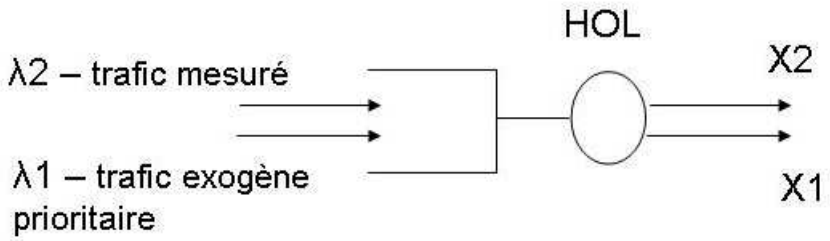


FIG. 25 – Représentation graphique d'un modèle Multiclasse

Plus formellement, les hypothèses faites sont que :
- La file est monoserveur et le buffer est de taille infinie.
- La discipline de service est HOL[17] et on suppose que le taux de service est exponentiel et identique pour tous les clients : μ
- Tous les processus d'arrivée sont supposés Poissonniens. Soient λ_1 et λ_2, les taux d'interarrivées des clients associés à chacune des classes.

[16] Aucune brique jusqu'à présent ne réduit son temps de séjour à mesure que son débit en sortie progresse.
[17] Head Of Line

- On suppose que le modèle fonctionnne à quasi-plein régime mais sans jamais saturer. Ceci se traduit par la relation : $\lambda_1 + \lambda_2 = \mu - \epsilon$ où ϵ est positif et petit comparé à μ.
- On suppose qu'il y a conservation du trafic traversant le système. C'est-à-dire qu'on imposera inmanquablement : $\lambda_1 + \lambda_2 =$ Constante=T. Le corollaire direct de cette relation est que lorsque λ_1 augmente, λ_2 diminue afin de maintenir constant le débit du trafic global.

7.2 Interprétation des hypothèses

Puisque ce modèle a été pensé pour reproduire des comportements issus de TCP, nous allons nous efforcer dans cette partie de donner un sens physique aux hypothèses que nous avons faites précédemment.

- La classe 1 représente le trafic dont on mesure en sortie le débit et le temps de séjour dans le réseau considéré. Les clients de la classe 2 officient le rôle endossé par le trafic exogène (ou de background) dans les réseaux informatiques. C'est-à-dire un trafic concurrent qui rivalise avec le trafic mesuré pour l'obtention des ressources disponibles.
- On suppose que la valeur de $\lambda_1 + \lambda_2$ est proche de celle de μ (bien qu'inférieure pour assurer la stabilité du modèle) et que cette valeur demeure identique à tous les points de mesures. Cette hypothèse reproduit, plus précisément, un comportement remarquable de TCP qui consiste à forcer progressivement l'ensemble des flux existants qui se partagent la capacité d'un lien à baisser leur débit lors de l'arrivée d'un nouveau flux ou bien lorsqu'un flux accroît son débit. Chaque flux tendant vers un débit moyen égal à : $\frac{1.22 \times MTU}{\sqrt{L} \times RTT}$
- C'est la priorité des clients de la classe 2 sur ceux de la classe 1 qui devrait permettre à ce modèle de générer des courbes de performances (pour le trafic 1) où plus le débit croît, plus le temps de réponse diminue. En effet, si le débit de la classe 1 augmente, c'est que celui de la classe 2 a diminué. Les clients de la classe 2 seront en conséquent moins nombreux dans la file d'attente, pénalisant moins les clients de la classe 1 qui la traverseront donc plus rapidement.

7.3 Calcul des paramètres de performance

Les paramètres de performance décisifs sont ceux se rapportant aux clients de la classe 1.

DÉBIT des clients 1 : $\bar{X}_{anal} = \lambda_1$

W_0, variable de calcul[18] $W_0 = \frac{\lambda_1 + \lambda_2}{\mu^2} = \frac{T}{\mu^2}$

TEMPS D'ATTENTE : $W_2 = \frac{W_0}{1-\frac{\lambda_2}{\mu}} = \frac{T}{\mu^2-\lambda_2\times\mu} = \frac{T}{\mu} \times \frac{1}{\mu-\lambda_2}$
des clients 2

TEMPS D'ATTENTE : $W_1 = \frac{\frac{\lambda_2\times W_2}{\mu}}{1-(\frac{\lambda_2}{\mu}+\frac{\lambda_1}{\mu})} = \frac{\frac{T}{\mu^2}+\frac{\lambda_2\times T}{\mu^2\times(\mu-\lambda_2)}}{1-\frac{T}{\mu}} = \frac{T}{\mu^2} \times \frac{\frac{\mu}{\mu-\lambda_2}}{1-\frac{T}{\mu}}$
des clients 1 $W_1 = \frac{T}{(\mu-T)\times(\mu-\lambda_2)}$

TEMPS DE SÉJOUR : $\bar{R}_{anal} = \bar{S} + \bar{W} = \frac{1}{\mu} + W_1$
des clients 1

7.4 Bornes sur les paramètres

Type de borne	*Expression littérale*	*Signification*
Inférieure sur T	$\max_{i\in[1;n]}(X_i)$	Comme le modèle est sans perte, $\bar{X}$ correspond seulement à une partie du débit total T
Inférieure sur μ	T	Condition de stabilité

TAB. 6 – Bornes pour un modèle Multiclasse

7.5 Processus de recherche du calibrage *ad hoc*

La recherche du calibrage consiste à déterminer les valeurs les plus favorables possibles pour deux paramètres continus : T et μ qui n'ont pas d'expressions littérales pour leur borne supérieure. Ici aussi on optera pour une recherche systématique mais restreinte à un espace de recherche inclus dans celui défini par les bornes. Cette recherche orientée par itération requiert une compréhension fine du rôle de chacun de ces deux paramètres. Voici les valeurs théoriques qu'on associera aux paramètres pour obtenir le calibrage souhaité.

1. On fixe la valeur de T à une valeur légèrement supérieure à $\max_{i\in[1,n]}(\bar{X}_i)$. Cette affectation traduit le fait que pour les fortes charges du trafic 1 mesuré, les clients de classe 1 consitueront la grande majorité du trafic circulant dans le système tandis que le trafic du flux 2 sera très faible. Plus la valeur de T sera proche de celle de $\max_{i\in[1,n]}(\bar{X}_i)$, plus les clients de classe 1 bénéficireont d'une amélioration forte sur leur temps moyen de séjour dans le système pour les charges élevées compratively au charge faible. Ainsi d'une certaine manière, la valeur de T détermine la pente de la décroissance des temps moyen de séjour sur les courbes de performances.
2. On fixe μ à une valeur supérieure à T pour assurer la condition de stabilité du modèle. Cependant la valeur de μ doit être relativement proche de celle de T pour éviter une situation où les clients ne subiraient que peu d'attente dans la file avant leur service.

7.6 L'algorithme de recherche mis en oeuvre

On va relâcher les valeurs relativement strictes décidées précédemment pour le jeu de paramètres et parcourir itérativement ces combinaisons de valeurs. ϵ est un réel positif non nul très petit comparé aux paramètres T et μ. Plus précisémment, l'approche suivante :

1. Calcul des $\lambda_{1,i}$. $\forall i \in [1;n], \lambda_{1,i} = \bar{X}_i$
2. Boucle sur T entre $[\max_{i\in[1;n]}(\bar{X}_i) + \epsilon; \max_{i\in[1;n]}(\bar{X}_i) + 2\epsilon]$.
3. A T fixé, l'obtention des $\lambda_{2,i}$ est immédiate. $\forall i \in [1;n], \lambda_{2,i} = T - \lambda_{1,i}$
4. Boucle sur μ entre $[T + \epsilon; T + 2\epsilon]$.
5. Calcul de W_0 puis de W_2 et de W_1 pour chaque niveau de charge $(\lambda_{1,i}, \lambda_{2,i})$. On en déduit le temps de séjour moyen $\bar{R}_{i,anal}$ des clients de classe 1 pour chaque niveau de charge $\lambda_{1,i}$.
6. Calcul de la fonction distance (se fiant aux écart entre $\bar{R}_{i,anal}$ et $\bar{R}_i$) pour le modèle à l'essai. Si le résultat du calcul situe le modèle parmi les 3 meilleurs modèles jusqu'à présent testés alors le modèle est retenu et remplace un modèle dans la liste provisoire des meilleurs modèles.

7.7 Limites du modèles

La principale critique que l'on peut adresser à cette brique est qu'on introduit dans la modélisation un mécanisme à priorité de classes alors que le système initiateur, connexion réseau en TCP, n'en implémente pas. Toutefois cette priorité a vocation à reproduire la répartition « équitable » des trafics telle qu'elle est effectuée dans les réseaux TCP/IP.

8 Les faisceaux des briques de base

Le but de cette partie est de dresser un comparatif entre les allures des courbes de performance des briques de base. Au terme de cette partie, nous serons entre autres capables de positionner le faisceau d'une M/M/C par rapport à celui d'un modèle FERME. On sait à présent que toutes les briques de base présentent des courbes de performance bimodales. Le taux de service μ détermine le temps de réponse à faible charge (en l'absence d'Offset) et le nombre de serveurs du modèle C, le rayon de courbure transitoire entre les deux régimes du modèle. De plus leur produit, $C\mu$ spécifie la plus petite borne supérieure sur le débit du modèle. Ces deux paramètres officiant le même rôle sur toutes les briques de base, nous supposerons leurs valeurs identiques pour tous les modèles cités dans le raisonnement à venir.

Notre compréhension des modèles nous suggère que dans ces conditions [19] la M/M/C définie [20] majore le faisceau de courbes des M/M/C/K. En effet à bas régime, ces deux briques se comportent identiquement. Mais lorsque λ se rapproche de $C\mu$ alors le modèle limité essuie des rejets ce qui réduit le nombre de clients présents dans la file. De plus ce nombre est borné par K. Dans le même temps, la M/M/C voit la taille de sa file grandir autant que nécessaire pour accueillir tous les clients. Par conséquent plus λ croît vers $C\mu$, plus la M/M/C présente des $\bar{X}$ et des $\bar{R}$ supérieurs à ceux de la M/M/C/K. Arrivées au seuil de $C\mu$, la M/M/C présente une asymptote verticale tandis que la M/M/C/K exhibe un point d'accumulation aux coordonnées $(C\mu\,;\frac{K}{C\mu})$.

A présent nous allons nous pencher sur le comparatif entre une M/M/C/K et une modèle FERME. Ces deux familles de modèles arborent un point d'accumulation aux coordonnées $(C\mu\,;\frac{K}{C\mu})$ et $(C\mu\,;\frac{N}{C\mu})$ respectivement. Pour savoir

[19] à mêmes valeurs de C et de μ

[20] le couple (C,μ désigne une et une seule M/M/C)

ce qui différencie ces deux modèles nous allons choisir K (pour le modèle ouvert) égal à N (pour le modèle fermé) ce qui se traduit par un emplacement identique du point d'accumulation pour les deux modèles. Ici notre intuition nous incite à penser que le faisceau généré par le modèle FERME minorera celui des M/M/C/K. L'explication tient au fait que plus λ augmente, plus le nombre de clients en file d'attente croît. Or pour les modèles fermés ce nombre conditionne les arrivées (surtout lorsque $N \gg C$) : il les réduit. A l'inverse dans les modèles ouverts, arrivées et nombre de clients dans le systèmes sont deux variables décorrélées. Du coup le modèle FERME devrait saturer moins brutalement que la M/M/C/K. C'est-à-dire avec une pente plus faible et il devrait se rapprocher moins vite selon le paramètre λ de sa destination finale.

Ces conjectures ont été vérifiées en instanciant des modèles issus des briques et en comparant leurs courbes de performance. La figure 26 illustre les trajectoires de ces courbes.

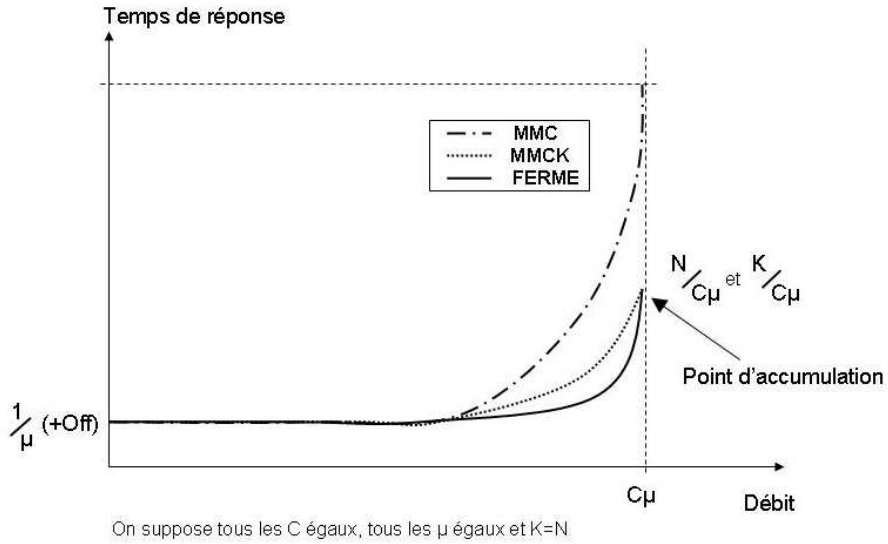


FIG. 26 – Comparaison des allures des courbes de performances des briques de base

Bien que cet exercice comparatif des allures des courbes de performances soit beaucoup plus difficile pour les modèles multiclasses et pour les modèles imbriqués, nous nous efforcerons de le faire dès que leurs formes seront finalisées.

Cinquième partie

Expérimentation et résultats

1 Expérimentation

L'expérimentation de notre programme de génération de modèles calibrés automatiquement utilise des jeux de mesures. Ces derniers peuvent être artificiels ou bien ils peuvent correspondre à des mesures réelles. S'il s'agit de jeux de mesures artificielles, ils ont été généralement générés à partir de modèles simples et éventuellement bruitées. Leur utilisation a permis entre autres de vérifier la cohérence de notre programme en testant sa capacité à retrouver le calibrage du modèle initiateur des mesures. Pour ce qui concerne les jeux de mesures réelles, ils ont été fournis par Alexandre et proviennent d'un contrôleur disque en E/S. Les valeurs de ces jeux de mesures sont données en annexe.

2 Les Résultats obtenus sur les jeux de mesures

2.1 Présentation des jeux de mesures

La figure 27 présente l'allure générale de chacun des quatre jeux de mesures. Des détails sur la nature des mesures et leurs valeurs précises sont disponibles en Annexe B de ce rapport. Ici nous nous contentons de décrire l'allure générale des courbes de mesures. Nous remarquons que les jeux de mesures 1 et 3 présentent un comportement où l'on distingue assez nettement les zones de faibles charge (pente nulle) et la zone de saturation tandis que le jeu 2 et à moindre mesure le jeu 4 semblent exhiber une saturation plus progressive. Une autre façon de dire est que les jeux 1 et 3 ont une pente quasi-nulle pour les niveaux de charges les plus faibles là où les jeux 2 et 4 suivent une pente linéaire.

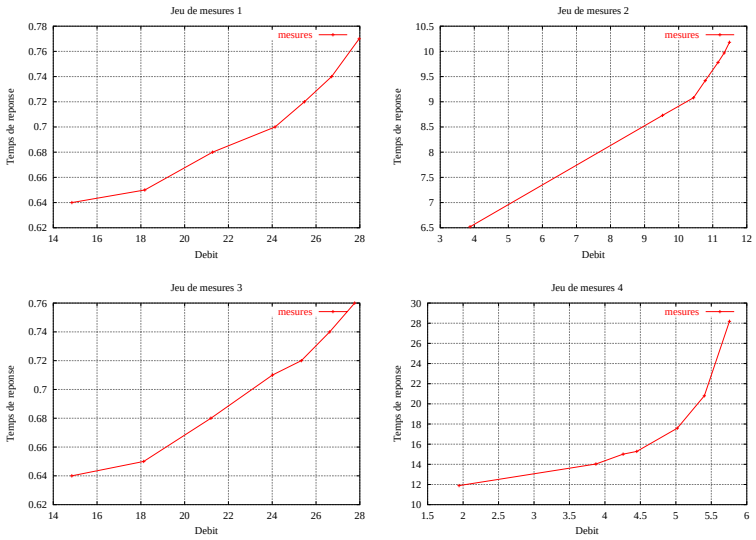


FIG. 27 – Allure générale des jeux de mesures 1 à 4

2.2 Avec les briques de base uniquement (sans l'Offset)

2.2.1 Jeu de mesures 1

Toutes briques confondues On constate que les résultats sont « assez bons » et que ce sont les M/M/C/K qui permettent de reproduire au mieux le comportement du système mesuré. Bien que convenable, ce résultat nous incitera à chercher une meilleure façon de procéder afin d'obtenir un modèle plus fidèle.

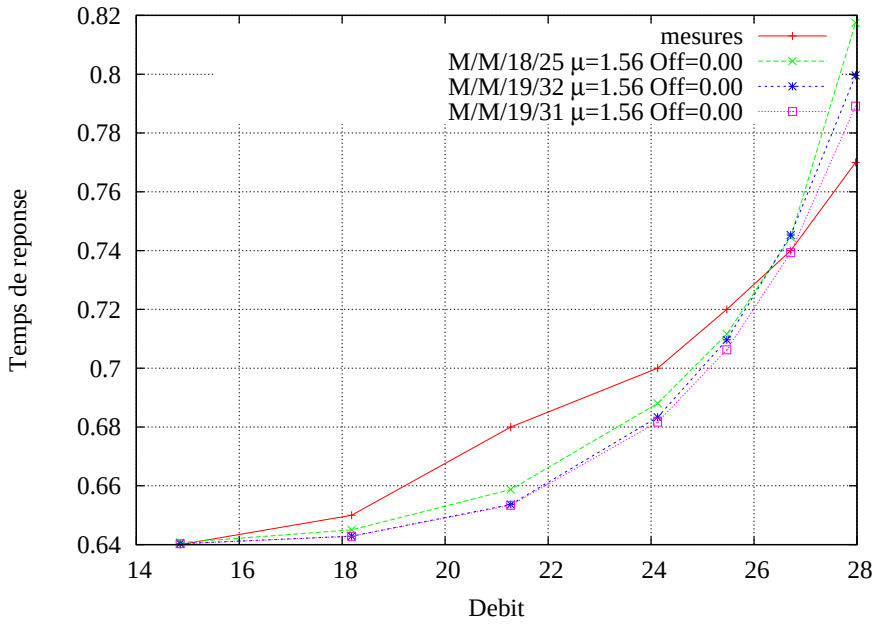


FIG. 28 – Courbes de performances du jeu de mesures 1 et des briques de base

Ordre	*Modèles*	*Distance*
1	M/M/18/25 avec μ :1.5625 et Off :0.000	0.006897
2	M/M/19/32 avec μ :1.5625 et Off :0.000	0.007097
3	M/M/19/31 avec μ :1.5625 et Off :0.000	0.007118

TAB. 7 – Classement des meilleurs modèles calibrés sans Offset

Jeu de mesures 1 avec briques de base sans Offset (suite...)

Avec un représentant par type de brique Il apparaît pour ce jeu de mesures que la M/M/C/K l'emporte assez largement sur les autres briques de base.

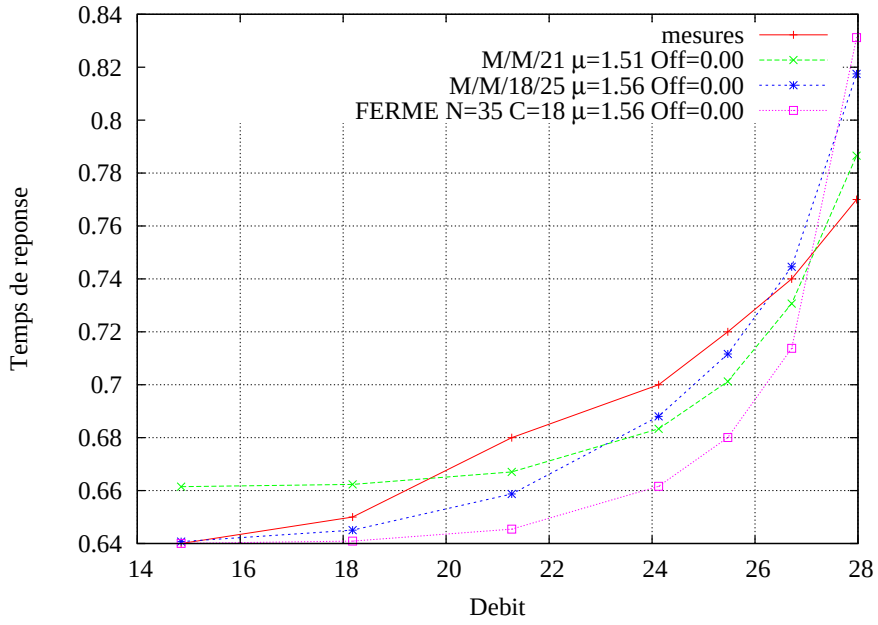


FIG. 29 – Courbes de performances du jeu de mesures 1 et des briques de base

Ordre	*Modèles*	*Distance*
1	M/M/21 avec μ :1.5120 et Off :0.000000	0.007864
2	M/M/18/25 avec μ :1.5625 et Off :0.000000	0.006897
3	FERME avec N :35 et C :18 et μ :1.5625	0.016089

TAB. 8 – Classement des meilleurs modèles calibrés sans Offset

2.2.2 Jeu de mesures 2 avec briques de base sans Offset

Avec un représentant par type de brique On constate qu'aucune brique de base, aussi bien calibrée que possible, ne permet de reproduire le comportement exhibé par le jeu de mesures 2. Les briques de base ne permettent pas de reproduire la saturation linéaire et très progressive observée sur les mesures. Cela signifie qu'il faudra trouver une autre façon de faire.

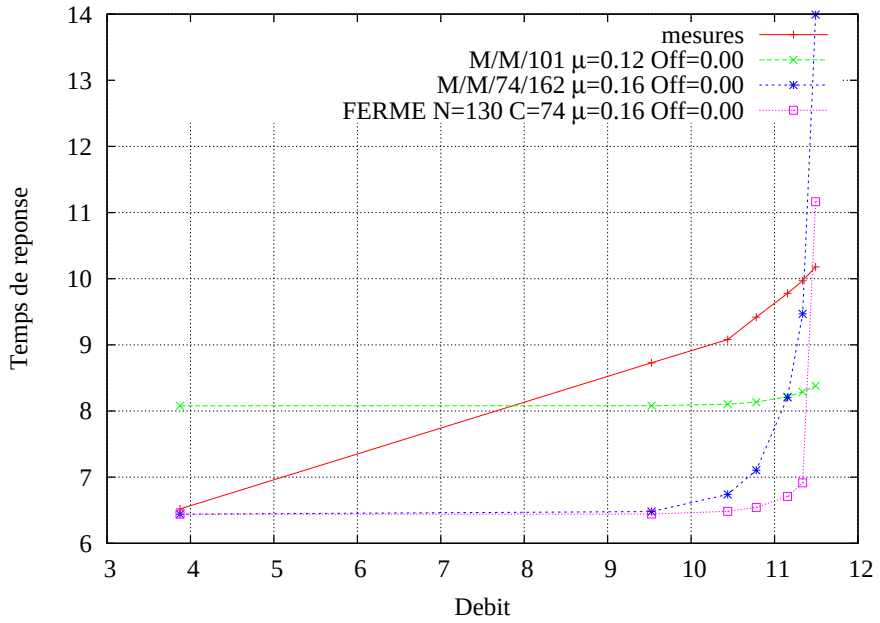


FIG. 30 – Courbes de performances du jeu de mesures 2 et des simples briques de base

1	M/M/101 avec μ :0.123 et Off :0.000	0.061255
2	M/M/74/162 avec μ :0.155 et Off :0.000	0.075427
3	FERME avec N :130, C :74 et μ :0.155 et Off :0.000	0.106397

TAB. 9 – Classement des meilleurs modèles calibrés sans Offset

2.2.3 Jeu de mesures 3 avec briques de base sans Offset

Toutes briques confondues On constate que les résultats sont « assez bons » et que ce sont les M/M/C/K qui permettent de reproduire au mieux le comportement du système mesuré.

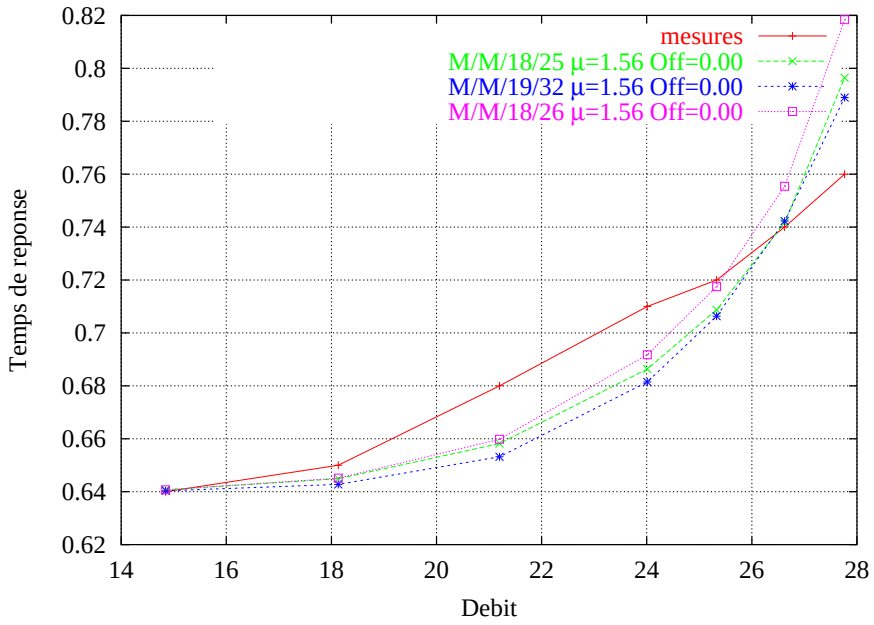


FIG. 31 – Courbes de performances du jeu de mesures 3 et des briques de base

Ordre	*Modèles*	*Distance*
1	M/M/18/25 avec μ :1.562 et Off :0.000	0.007325
2	M/M/19/32 avec μ :1.562 et Off :0.000	0.008382
3	M/M/18/26 avec μ :1.562 et Off :0.000	0.008499

TAB. 10 – Classement des meilleurs modèles calibrés sans Offset

Jeu de mesures 3 avec briques de base sans Offset (suite...)

Avec un représentant par type de brique Il apparaît pour ce jeu de mesures que la M/M/C/K l'emporte assez largement sur les autres briques de base.

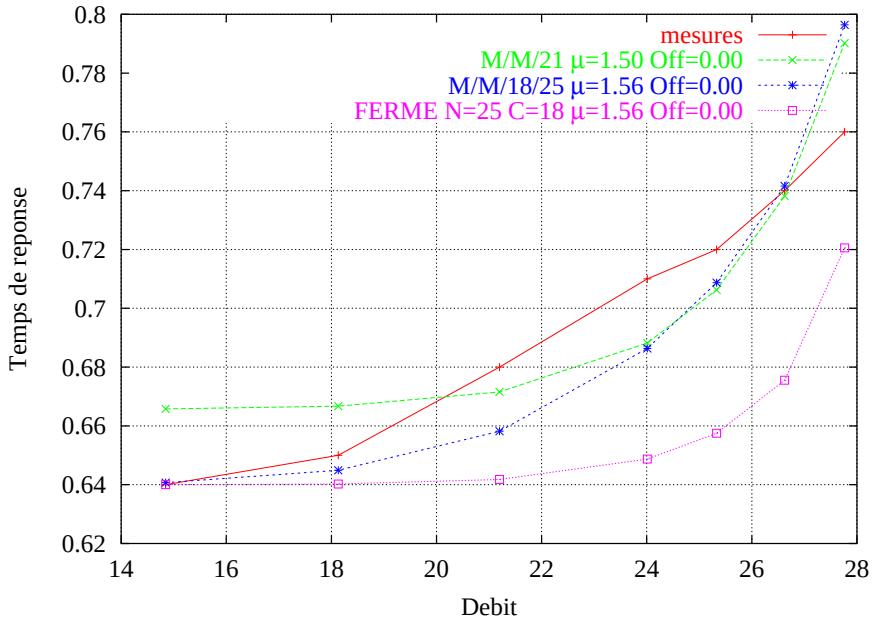


FIG. 32 – Courbes de performances du jeu de mesures 3 et des briques de base

Ordre	*Modèles*	*Distance*
1	M/M/21 avec μ :1.502 et Off :0.000	0.008849
2	M/M/18/25 avec μ :1.562 et Off :0.000	0.007325
3	FERME avec N :25, C :18, μ :1.562 et Off :0.000	0.027389

TAB. 11 – Classement des meilleurs modèles calibrés sans Offset

2.2.4 Jeu de mesures 4 avec briques de base sans Offset

Toutes briques confondues De même que pour le jeu de mesures 2, on constate qu'aucune brique de base, aussi bien calibrée que possible, ne permet de reproduire le comportement exhibé par le jeu de mesures 2. Les briques de base ne permettent pas de reproduire la saturation linéaire et très progressive observée sur les mesures. Cela signifie qu'il faudra trouver une autre façon de faire.

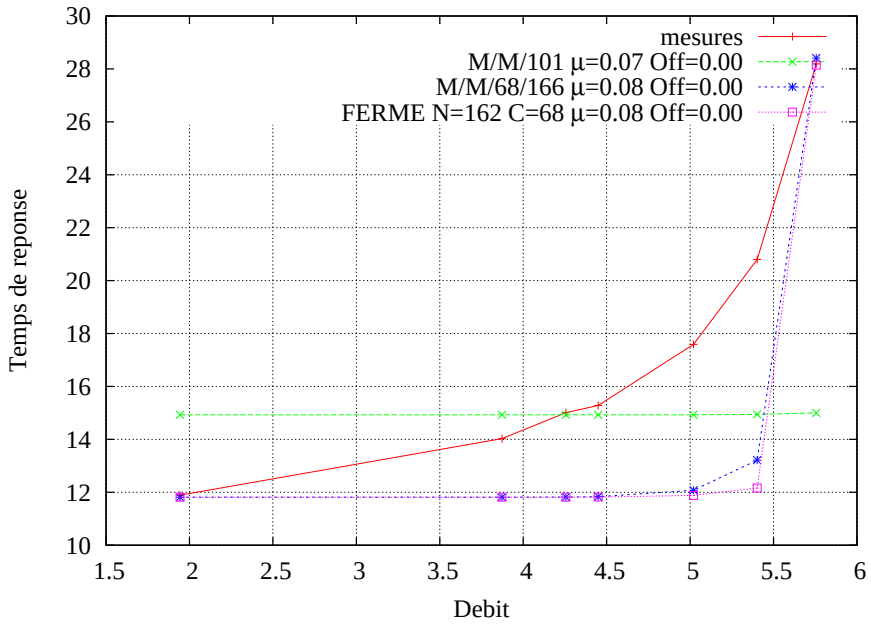


FIG. 33 – Courbes de performances du jeu de mesures 4 et des briques de base

Ordre	*Modèles*	*Distance*
1	M/M/101 avec μ :0.0669 et Off :0.000	0.070194
2	M/M/68/165 avec μ :0.084 et Off :0.000	0.091459
3	FERME avec N :162, C :68, μ :0.084 et Off :0.000	0.096103

TAB. 12 – Classement des meilleurs modèles calibrés sans Offset

2.3 Avec les briques de base et l'Offset

2.3.1 Jeu de mesures 1

Toutes briques confondues On constate que les résultats sont « bons » et que ce sont les M/M/C/K qui permettent de reproduire au mieux le comportement du système mesuré. L'apport de l'Offset s'est traduit par une amélioration très nette des résultats (à comparer avec la figure 28). La valeur de la fonction distance est réduite jusqu'à 6 fois sa valeur.

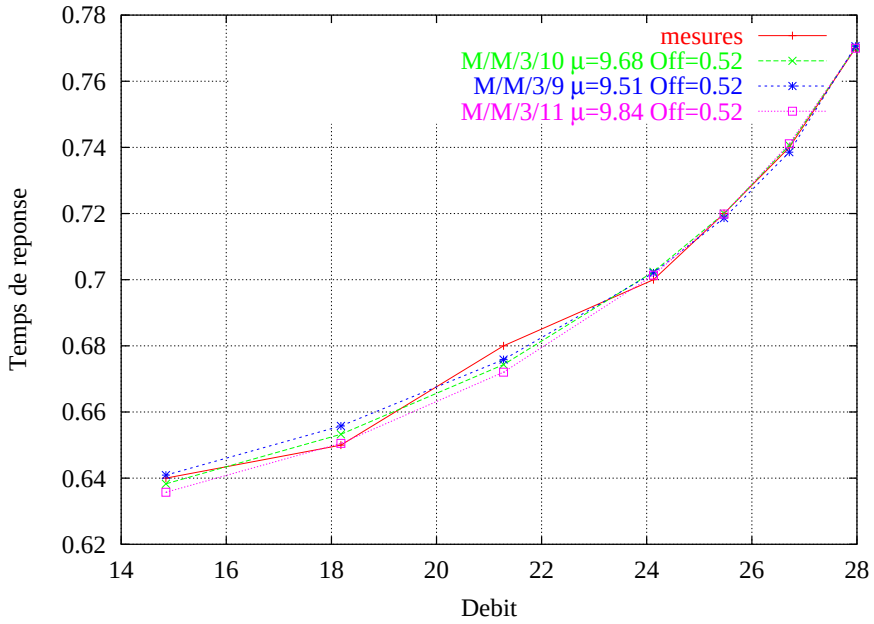


FIG. 34 – Courbes de performances du jeu de mesures 1 et des briques de base

Ordre	*Modèles*	*Distance*
1	M/M/3/10 avec μ :9.679 et Off :0.518	0.001037
2	M/M/3/9 avec μ :9.507 et Off :0.518	0.001228
3	M/M/3/11 avec μ :9.839 et Off :0.518	0.001367

TAB. 13 – Classement des meilleurs modèles calibrés avec Offset

Jeu de mesures 1 avec briques de base et Offset (suite...)

Avec un représentant par type de brique Il apparaît pour ce jeu de mesures que la M/M/C/K l'emporte assez largement sur les autres briques de base. Mais peut être plus important on constate que l'Offset profite substantiellement à toutes les briques. Enfin on note que le nombre de serveurs de toutes les briques a également diminué (à comparer avec le tableau 2.2.1 à la page 66).

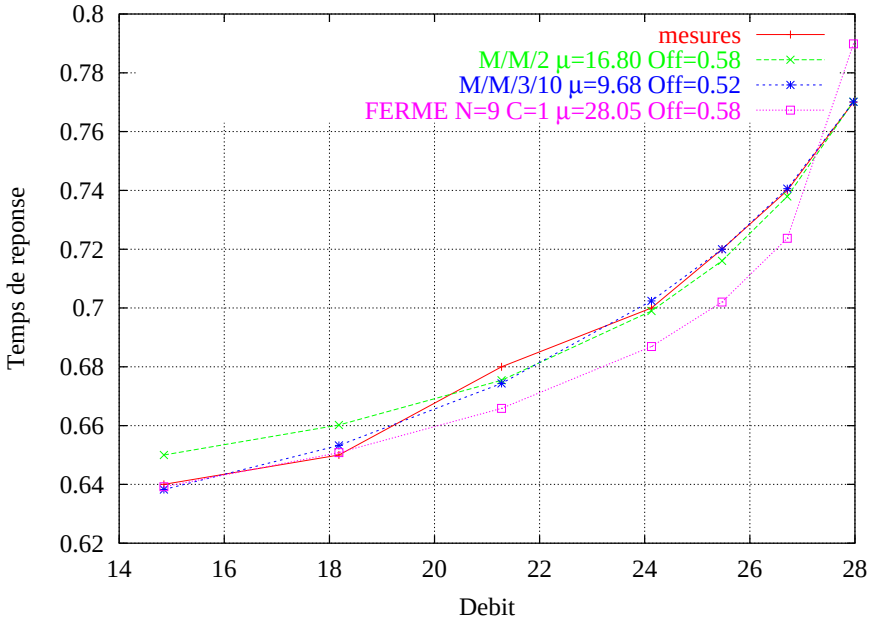


FIG. 35 – Courbes de performances du jeu de mesures 1 et des briques de base

Ordre	*Modèles*	*Distance*
1	M/M/2 avec μ :16.796 et Off :0.576	0.002541
2	M/M/3/10 avec μ :9.679 et Off :0.518	0.001037
3	FERME avec N :9, C :1, μ :28.052 et Off :0.576	0.006557

TAB. 14 – Classement des meilleurs modèles calibrés avec Offset

2.3.2 Jeu de mesures 2 avec briques de base et Offset

Avec un représentant par type de brique Ici le résultat s'oppose à celui constaté pour le scénario précédent. La présence de l'Offset dans les recherches n'a pas permi de trouver un « bon » modèle. La conclusion est qu'aucune brique de base, même dotée d'un Offset ne peut reproduire le comportement exhibé par les mesures : à savoir une saturation linéaire et très progressive. Il nous faudra inventer une nouvelle brique pour pallier ce manque.

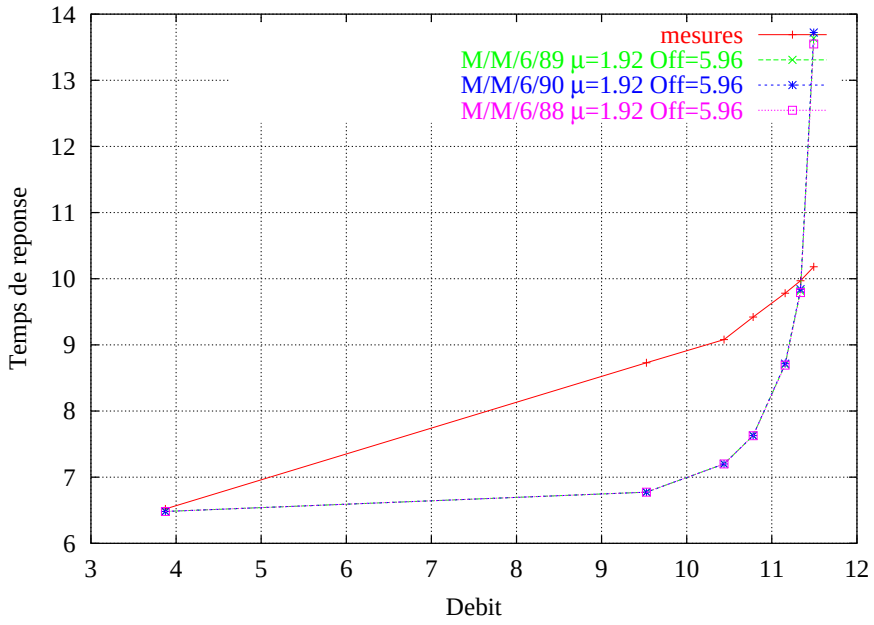


FIG. 36 – Courbes de performances du jeu de mesures 2 et des simples briques de base

1	M/M/6/89 avec μ :1.915 et Off :5.958	0.058961
2	M/M/6/90 avec μ :1.915 et Off :5.958	0.058965
3	M/M/6/88 avec μ :1.915 et Off :5.958	0.058967

TAB. 15 – Classement des meilleurs modèles calibrés avec Offset

2.3.3 Jeu de mesures 3 avec briques de base et Offset

Toutes briques confondues On constate que les résultats sont « bons » et que ce sont les M/M/C/K qui permettent de reproduire au mieux le comportement du système mesuré. L'apport de l'Offset s'est traduit par une amÈlioration très nette des résultats (à comparer avec la figure 31). La valeur de la fonction distance est réduite jusqu'à 6 fois sa valeur.

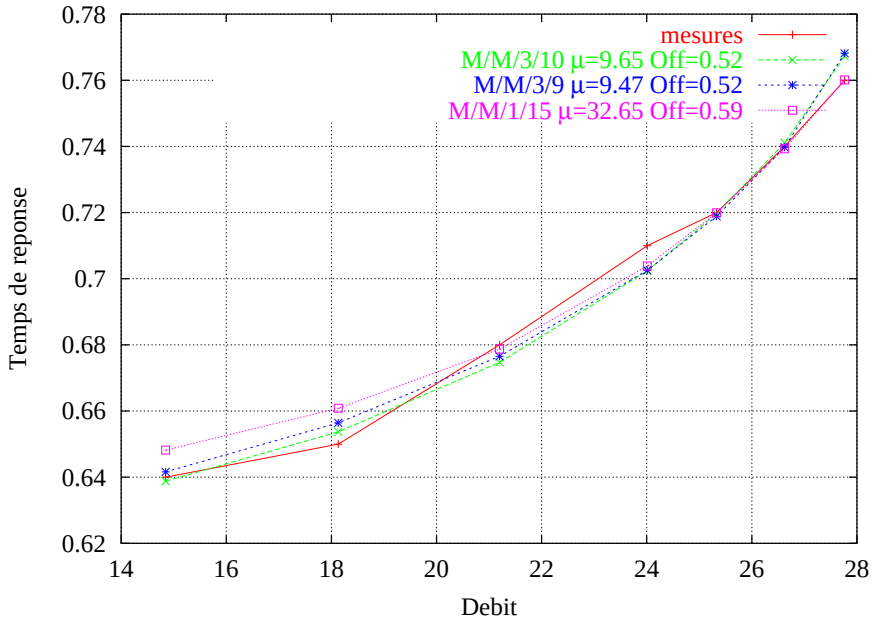


FIG. 37 – Courbes de performances du jeu de mesures 3 et des briques de base

Ordre	*Modèles*	*Distance*
1	M/M/3/10 avec μ :9.650 et Off :0.518	0.001912
2	M/M/3/9 avec μ :9.470 et Off :0.518	0.002070
3	M/M/1/15 avec μ :32.648 et Off :0.592	0.002271

TAB. 16 – Classement des meilleurs modèles calibrés avec Offset

Jeu de mesures 3 avec briques de base et Offset (suite...)

Avec un représentant par type de brique Il apparaît pour ce jeu de mesures que la M/M/C/K l'emporte assez largement sur les autres briques de base. Mais peut être plus important on constate que l'Offset profite substantiellement à toutes les briques. Enfin on note que le nombre de serveurs de toutes les briques a également diminué (à comparer avec le tableau 12 à la page 70).

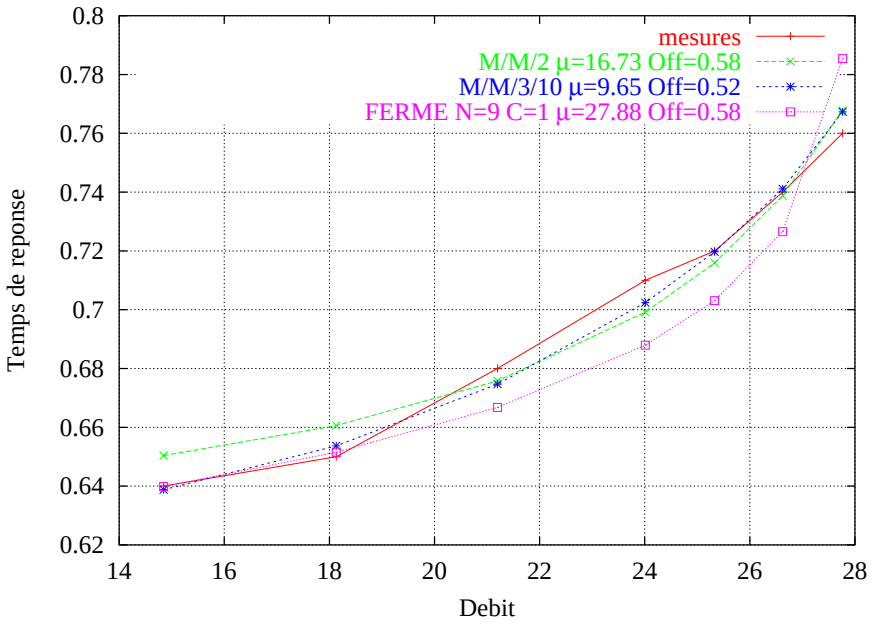


FIG. 38 – Courbes de performances du jeu de mesures 3 et des briques de base

Ordre	*Modèles*	*Distance*
1	M/M/2 avec μ :16.733 et Off :0.576	0.003682
2	M/M/3/10 avec μ :9.650 et Off :0.518	0.001912
3	FERME avec N :9, C :1, μ :27.876 et Off :0.576	0.007082

TAB. 17 – Classement des meilleurs modèles calibrés avec Offset

2.3.4 Jeu de mesures 4 avec briques de base et Offset

Toutes briques confondues Comme pour le jeu de mesures 2, la présence de l'Offset n'a pas suffi pour trouver un « bon » modèle. La conclusion est qu'aucune brique de base, même dotée d'un Offset ne peut reproduire le comportement exhibé par les mesures : à savoir une saturation linéaire et très progressive. Il nous faudra inventer une nouvelle brique pour pallier ce manque.

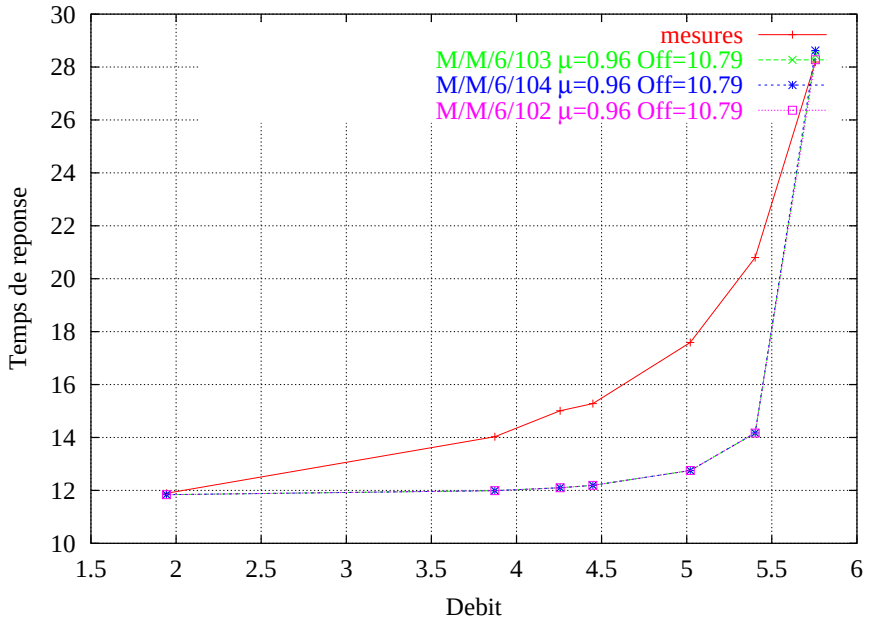


FIG. 39 – Courbes de performances du jeu de mesures 4 et des briques de base

Ordre	*Modèles*	*Distance*
1	M/M/6/103 avec μ :0.959 et Off :10.791	0.081413
2	M/M/6/104 avec μ :0.959 et Off :10.791	0.081413
3	M/M/6/102 avec μ :0.959 et Off :10.791	0.081416

TAB. 18 – Classement des meilleurs modèles calibrés avec Offset

2.4 Avec les modèles imbriqués

Nous l'avons dit précédemment, les modèles imbriqués ont vocation à suppléer les briques de base pour les jeux de mesures présentant une saturation à l'allure progressive. Et comme nous l'avons constaté au cours des pages précédentes, les jeux de mesures 2 et 4 exhibent des courbes de performances dont l'allure n'est pas reproductible par les briques de base et devrait pouvoir l'être par les modèles imbriqués.

2.4.1 Jeu de mesures 2 avec modèles imbriqués et Offset

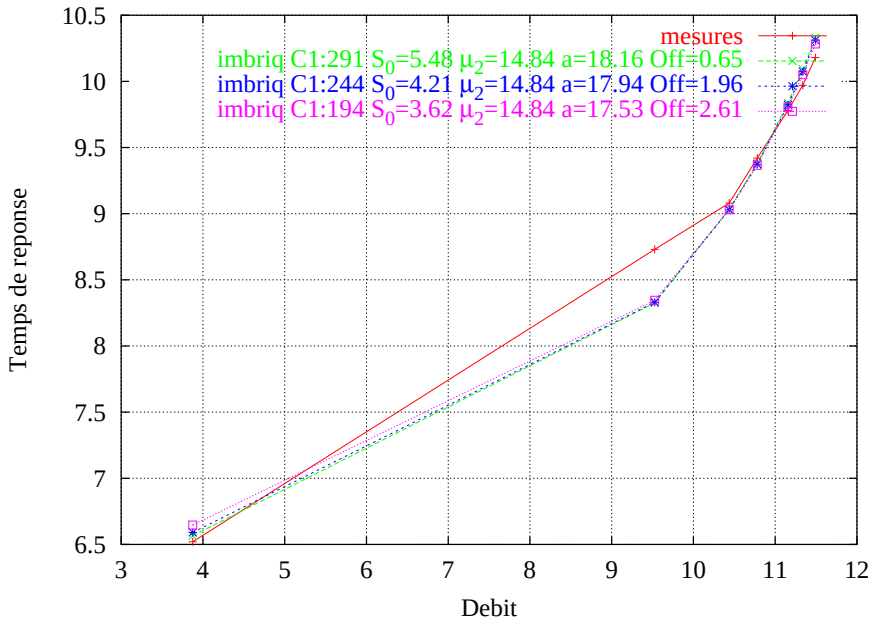


FIG. 40 – Courbes de performances du jeu de mesures 2 et des modèles imbriqués

Ordre	*Modèles*	*Distance*
1	Imbriqué avec C_1 :291 S_0 :5.481 μ_2 :14.841 a :18.155 et Off :0.652	0.004878
2	Imbriqué avec C_1 :244 S_0 :4.207 μ_2 :14.841 a :17.940 et Off :1.956	0.004881
3	Imbriqué avec C_1 :194 S_0 :3.620 μ_2 :14.841 a :17.532 et Off :2.608	0.004881

TAB. 19 – Classement des meilleurs modèles calibrés

2.4.2 Jeu de mesures 4 avec modèles imbriqués et Offset

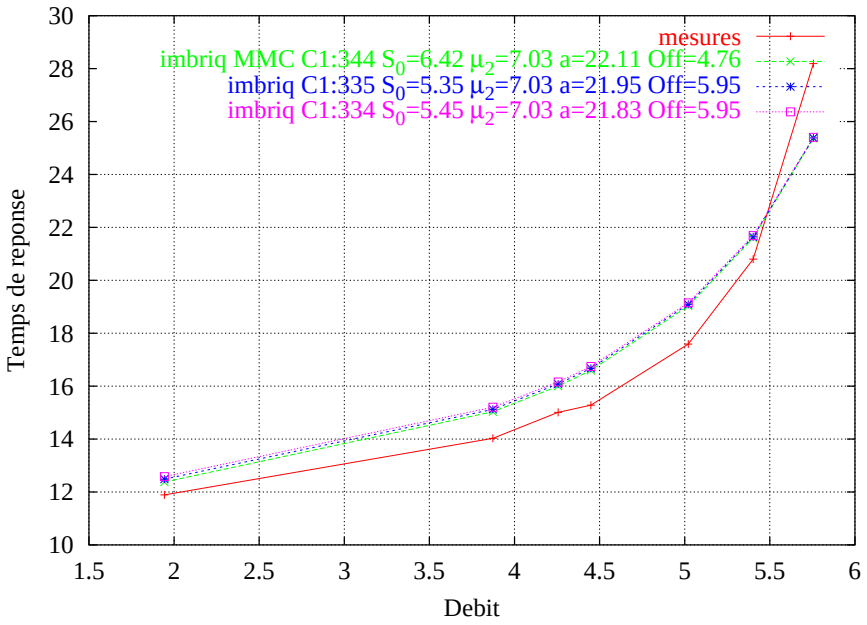


FIG. 41 – Courbes de performances du jeu de mesures 4 et des modèles imbriqués

Ordre	*Modèles*	*Distance*
1	Imbriqué avec C_1 :344 S_0 :6.4206 μ_2 :7.0316 a :22.114089 et Off :4.756	0.028402
2	Imbriqué avec C_1 :335 S_0 :5.3505 μ_2 :7.0316 a :21.951918 et Off :5.945	0.030610
3	Imbriqué avec C_1 :334 S_0 :5.4505 μ_2 :7.0316 a :21.832861 et Off :5.945	0.032579

TAB. 20 – Classement des meilleurs modèles calibrés

On constate à travers ces deux exemples deux mesures que les modèles imbriqués remplissent « correctement » leur mission. Bien calibrés, ils permettent de reproduire assez fidèlement les performances mesurées sur les jeux 2 et 4 (jusqu'alors impossible). Seul bémol, les valeurs de C_1 sont très élevées et ne semblent pas correspondre à une réalité physique. Nous aborderons plus en détail ce point dans la dernière partie de ce rapport.

2.5 Avec les modèles multiclasses sans Offset

Voici les résultats obtenus par la brique multiclasse (sans puis avec Offset) sur les jeux de mesures 5 et 6.

2.5.1 Jeu de mesures 5 avec modèles multiclasses sans Offset

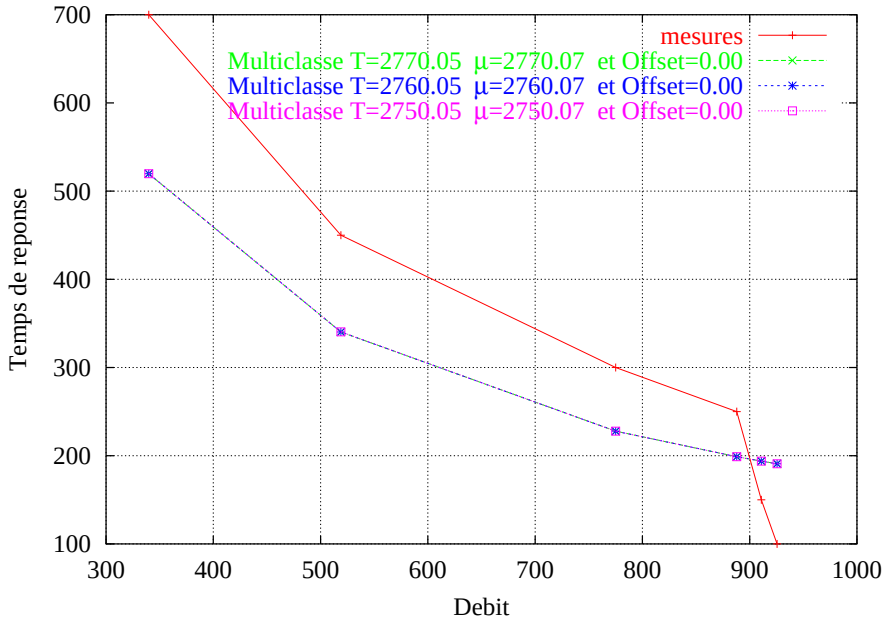


FIG. 42 – Courbes de performances du jeu de mesures 5 et des modèles multiclasses

Ordre	*Modèles*	*Distance*
1	Multiclasse avec T :2770.05, μ :2770.06 et Off :0.000	0.127394
2	Multiclasse avec T :2760.05, μ :2760.06 et Off :0.000	0.127400
3	Multiclasse avec T :2750.05, μ :2750.06 et Off :0.000	0.127406

TAB. 21 – Classement des meilleurs modèles calibrés

2.5.2 Jeu de mesures 6 avec modèles multiclasses sans Offset

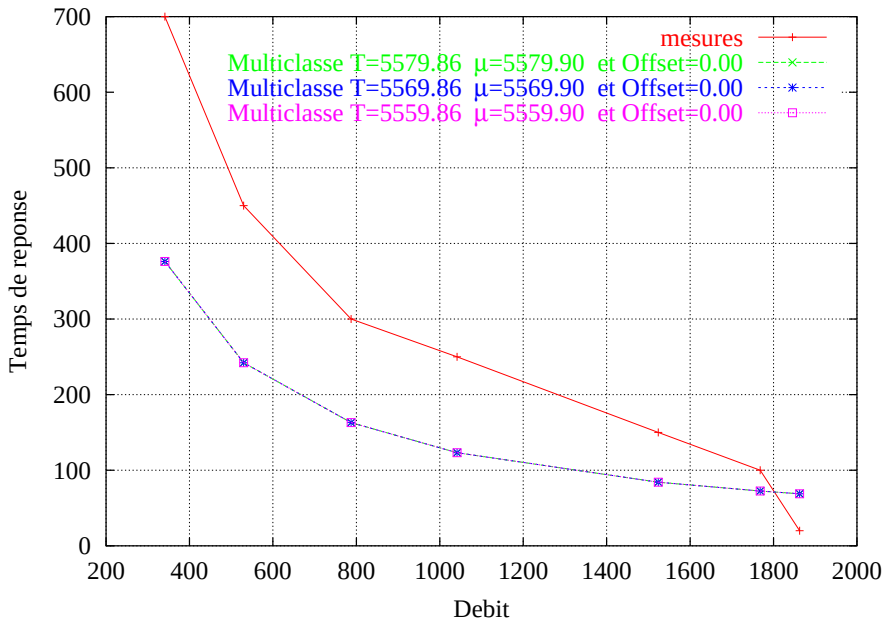


FIG. 43 – Courbes de performances du jeu de mesures 6 et des modèles imbriqués

Ordre	*Modèles*	*Distance*
1	Multiclasse avec T :5579.85, μ :5579.90 et Offset=0.000	0.254963
2	Multiclasse avec T :5569.85, μ :5569.90 et Offset=0.000	0.254963
3	Multiclasse avec T :5559.85, μ :5559.90 et Offset=0.000	: 0.254964

TAB. 22 – Classement des meilleurs modèles calibrés

On constate que les résultats ne sont très bons. Dans les deux exemples, la courbure de la courbe semble à peu près correcte mais il existe un décalage très marqué entre les courbes de performances des mesures et celles des modèles calibrés. On se peut se poser la question si ces résultats ne seraient pas bonifiés en présence d'un Offset. La réponse est non. Pas parce que l'Offset est inutile pour les modèles multiclasses mais simplement parce que la valeur de l'Offset doit être choisie entre $[0\,;\min_{i\in[1,n]} \bar{R}_i]$. Rapportée à nos jeux de mesures 5 et 6, cette contrainte se traduit par un Offset d'une valeur maximale de 40 ms. Ce qui correspond à une valeur très « petite » comparée à la dimension du décalage (de l'ordre de 100ms). On comprend donc pourquoi l'ajout d'un Offset sur les temps de séjour moyens calculés $\bar{R}_{i,anal}$ ne suffit pas à rendre les résultats bons. La conséquence de ces résultats peu convaincants est que pour mieux reproduire ces jeux de mesures, il nous faudra approfondir nos connaissances sur les modèles multiclasses afin de comprendre les raisons de ce décalage et d'y remédier. Certaines propositions sont énoncées dans la partie suivante de ce rapport. Toutefois afin de nous lancer dans des recherches, il faudra vérifier avec d'autres sources que ces jeux de mesures 5 et 6 sont bien représentatifs

d'un comportement fréquemment rencontré (comme TCP par exemple).

3 Les plus-values et les contributions apportées par nos travaux

Au delà des résultats expérimentaux sur les jeux de mesures présentées précédemment, les bénéfices occasionnés par nos travaux résident essentiellement dans :

- Une meilleure compréhension des modèles issus des briques. Cela se traduit notamment par une intuition acérée sur l'influence et le rôle des paramètres sur les performances d'un modèle. Cette connaissance sera indispensable pour mettre au point des méthodes de calibrage automatique qui permettraient d'éviter les « lenteurs » des approches systématiques.
- La validation (partielle) de la méthodologie du projet de génération de modèles calibrés automatiquement, et donc plus généralement d'une approche descriptive grâce à quelques exemples probants.

Sixième partie

Mise en oeuvre et détails techniques

1 Valoriser/déprécier certaines mesures - les poids sur les mesures

1.1 Ajuster la fonction distance

Au cours de la partie 3 (page 30, formule (1)), nous avons défini la fonction distance d. Nous avions noté sans l'expliquer la présence de coefficients pondérateurs w_i. Nous présentons à présent le rôle exact de ces coefficients qui permettent de paramétrer le processus de recherche du meilleur modèle.

Ces coefficients w_i officient plusieurs rôles qui impactent la prise en compte des mesures dans la recherche du meilleur modèle. Chronologiquement la première fonction de ces coefficients pondérateurs a été de forcer l'importance de certains points de mesure en leur affectant des coefficients plus élevés que la moyenne. Inversement pour rendre un point de mesure muet, il nous suffit d'attribuer à son coefficient une valeur très faible pour l'empêcher d'intervenir dans la fonction d. Cette relation causale se justifie facilement en observant l'expression de la fonction arbitre distance. De plus, afin de donner un sens « absolu » aux résultats de la fonction distance, la somme des coefficients pondérateurs est maintenue égale à 1. En pratique, les mesures créditées de peu de confiance ou marginales seront dotées d'un coefficient plus faible et en revanche les mesures appartenant à la zone de fonctionnement étudiée et considérées comme plus sûres seront affectées d'un coefficient plus élevé.

1.2 Relâcher les bornes trop restrictives

La deuxième fonction des coefficients w_i est plus subtile. Elle consiste à relâcher certaines bornes [21] rendues trop drastiques ou inaptes à cause de mesures biaisées. Jusqu'à présent, dans le cas extrême où l'on attache aucune importance ou aucune confiance à une mesure en particulier ($w_i = 0$), celle ci ne participe pas à établir le classement des modèles testés (par définition de la fonction distance) mais en revanche elle contribuera à évincer certains modèles de par les bornes qu'elle induit sur certains paramètres. C'est cet effet de bord que l'on veut endiguer. Pour cela nous avons défini une fonction dont l'ensemble de départ est l'intervalle [0 ;1] (relatif aux poids des mesures) et l'ensemble d'arrivée [0 ;0.5]. La définiton exacte de cette fonction qui régit la tolérance admise sur les points de mesures est la suivante :

$$\text{fonction relachement} : \begin{cases} [0;1] & \Longrightarrow & [0;0.5] \\ x & \longrightarrow & \begin{cases} 1-x & \text{si } x > 0.5 \\ 0.5 & \text{sinon} \end{cases} \end{cases}$$

Elle prend en entrée le poids d'une mesure, et retourne en pourcentage l'allongement maximal toléré (dans un sens comme dans l'autre). On note la

[21] Les bornes ont été présentées précédemment

présence d'un seuil maximum de relâchement à 50%. Ce mécanisme permet donc d'accroître l'espace de recherche sur les paramètres, un allongement dont on espère faire profiter la recherche d'un modèle calibré. Mais attention on ne peut pas toujours relâcher certaines bornes. Citons à titre d'exemple la borne μ_{min} d'une M/M/C ne peut pas être relachée car elle assure la stabilité du modèle.

En résumé les coefficients pondérateurs traduisent :
- L'importance relative des mesures dans l'évaluation de la fonction distance
- Et le degré de tolérance ou d'imprécision sur les mesures en relâchant les bornes qu'elles induisent

Ces deux fonctionnalités sont régis par une seule variable, pour des raisons de simplicité. Ainsi une mesure créditée d'un poids de 1 imposera des bornes strictes et sera valorisée comparativement à une mesure associée à un poids inférieur à 1 qui engendrera des bornes plus lâches.

Exemple. Cet exemple doit permettre de mieux mesurer l'utilité du relâchement sur les mesures. Soit une série de i mesures $(\bar{X}_i, \bar{R}_i)$ que l'on suppose engendrée par un système assimilable parfaitement à une M/M/C/K (i.e. Arrivées Poissoniennes, loi de service exponentielle, C serveurs et une capacité de K). A présent on suppose que la dernière mesure n a été biaisée. Les valeurs $\bar{X}_n$ et $\bar{R}_n$ sont exagérées. Par conséquent la borne sur K présentée 29 risque d'imposer une borne inférieure sur K supérieure à la valeur réelle de K. Du coup, la présence de la dernière mesure empêche l'obtention du modèle optimal. or il est possible que l'analyste ait quelques réserves sur cette dernière mesure. Il lui suffira alors de créditer cette mesure douteuse d'un poids faible, mettons 0.5, pour s'assurer que la borne inférieure sur K apportée par la dernière mesure soit diminuée de 50%.

2 L'Offset

Nous apportons ici des informations très précises sur l'Offset.

2.1 Rappels

Nous avons vu que l'Offset consiste à supposer que dans les temps de séjour moyens $\bar{R}$ à calculer il existe une partie variable en fonction de la charge et une partie constante. Nous avons également vu que sur un plan physique, ce temps incompressible peut être perçu comme l'effet d'une M/M/∞ placée en série derrière le modèle principal. Le temps moyen de séjour dans une M/M/∞ est toujours le même quelque soit la charge en entrée et en régime permament la présence d'une M/M/∞ en série est transparente pour le débit calculé.

2.2 Un cas d'étude probant

Ce cas d'étude permet de mieux saisir l'apport de l'Offset pour la recherche du meilleur modele.
Supposons que nous souhaitons imposer à deux modèles M/M/C (M_1 et M_2),

avec un nombre de serveurs différent (C_1=1 et C_2=5), d'avoir un temps de séjour moyen $\bar{X}$ identique à faible charge tout en ayant le même seuil de saturation.

On démarre par imposer aux deux modèles de saturer dès que la charge atteint $\lambda = 1$. Or pour une M/M/C, c'est le résultat du produit $C\mu$ détermine le seuil de saturation. Par conséquent, cette condition équivaut à imposer μ_1=1 et μ_2=$\frac{1}{5}$. Mais puisque μ_1 et μ_2 sont fixés, les temps de réponse à charge quasi-vide sont connus : 1 pour le M_1 et 5 pour M_2. Il est donc impossible (sans Offset) d'avoir deux M/M/C ayant même temps de séjour moyen et même seuil de saturation. Pourtant il serait intéressant de pouvoir le faire. Cela permettrait de profiter de la différence de courbure qu'il y a entre deux M/M/C aux valeurs de C différentes (comme nous l'avons vu dans l'analyse détaillée des briques de base). Le rayon de courbure sera plus prononcé et plus tardif pour le modèle ayant le plus grand nombre de serveurs.

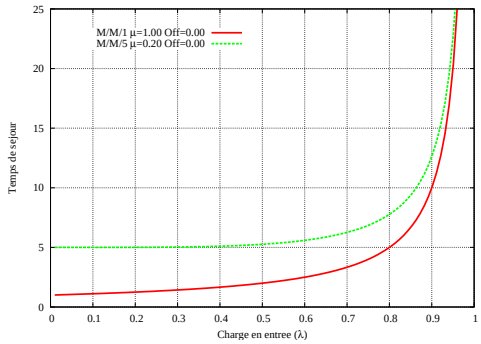


FIG. 44 – M/M/1 sans Offset et M/M/5 sans Offset

C'est pour combler ce manque que nous avons pensé à l'Offset. Dans notre exemple, il suffit d'ajouter un Offset de 4 aux temps de séjour moyen $\bar{R}_1$ du modèle M_1 pour assurer le même comportement aux deux modèles à faible charge. La figure 45 illustre la solution présentée : deux M/M/C aux valeurs de C différentes dont seul les rayons de courbure respectifs diffèrent.

En terme de modélisation, la solution consiste à placer une M/M/∞ en série de temps moyen de service de 4 avec le modèle M_1 .

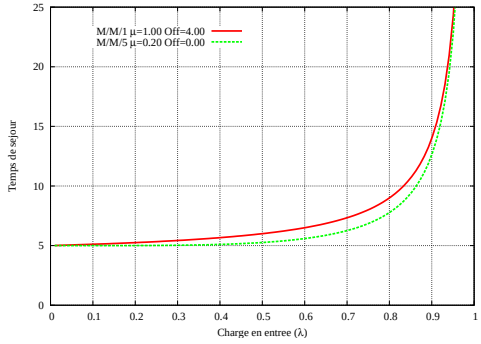


FIG. 45 – M/M/1 avec Offset et M/M/5 sans Offset

2.3 Analyse

L'exemple précédent peut être généralisé à toutes les briques de base. Nous avons vu au cours de la partie 4 que les courbes de performances de chaque brique de base avait un nombre d de degrés de libertés (3 pour les M/M/C, 4 pour les M/M/C/K et les FERME). Or il apparaît que dès que $d-1$ de ses degrés de libertés sont décidés, alors le d^{ieme} disparaît. L'exemple précédent l'a expliqué en détail pour les M/M/C. Cela signifie que sans Offset, on n'exploite pas totalement les capacités de nos briques de base.
En autorisant un Offset sur les temps de séjour moyens, le calibrage peut jouer librement sur les d degrés de libertés du modèle pour approcher au mieux la courbe de mesures proposée.

En pratique, on a observé au cours de la partie 5 de ce rapport que la présence de l'Offset permet d'obtenir des modèles bien meilleurs (la valeur de la fonction distance est réduite d'un facteur 6 environ sur les jeux de mesures les plus favorables). Par ailleurs on a remarqué également qu'avec l'Offset les modèles retenus se distinguent nettement des modèles retenus sans Offset par le nombre bien inférieurs de serveurs C dont ils disposent (ainsi que d'une réduction plus modérée de la taille des files K) pour les M/M/C/K. Cette nouvelle valeur de C paraît plus significative et nous expliquons cette baisse sur la valeur de C de la façon suivante : A présent la valeur de C (et dans une moindre mesure celle de K) est fonction uniquement par la courbure de la courbe à reproduire. Tandis que sans Offset le rayon de courbure du modèle est conditionné également par les choix exercés sur les autres degrés de libertés.

2.4 Intégration au processsus de recherche du calibrage

Les conséquences de l'Offset pour le processus de recherche du calibrage sont très simples conceptuellement. Il suffit de soustraire la valeur de l'Offset à tous les temps de service moyen $\bar{R}_i$. Et naturellement les débits moyens ne sont pas modifiés.

En ce qui concerne l'algorithme de recherche du calibrage consiste à ajouter initialement une boucle sur le paramètre d'Offset, Off. Pour cahque valeur de Off testé, la recherche se déroule classiquement si ce n'est qu'elle repose à présent sur l'ensemble de mesures $(\bar{X}_i; \bar{R}_i - Off)$.

3 Inférer la charge de travail : λ

Pour évaluer les performances d'un modèle à l'essai (et ensuite les comparer à un point de mesure), il est indispensable de commencer par trouver le « bon » niveau de charge λ. Dans nos travaux, ce bon niveau signifie que : $\bar{X}_{i,anal} = \bar{X}_i$. Cette opération est très différente selon les types de modèles.

Pour les modèles ouverts sans perte, le calcul est très simple. Il suffit de choisir $\lambda_i = \bar{X}_i$.

Pour les modèles avec pertes ou avec rejets, la situation est plus compliquée. Toutefois il existe un procédé commun à tous nos briques pour trouver le taux de charge en entrée λ_i qui engendre exactement $\bar{X}_i$ en sortie. La solution s'appuie sur une proriété commune à toutes les briques (que nous avons vues dans la partie 4) qui dit que pour tout modèle (issu de nos briques et dont les paramètres intrinsèques sont fixés), plus λ_i augmente, plus $\bar{X}_{i,anal}$ augmente. Ce

comportement monotone est exploité pour inférer la cahrge λ correspondant à un débit moyen mesuré $\bar{X}$. A un niveau de débit en sortie correspond un seul niveau de charge en entrée.
On est donc en mesure pour n'importe quel modèle paramétré de déterminer, sans ambiguïté la valeur exacte à donner à λ pour reproduire en sortie un débit moyen égal à $\bar{X}_i$. Bien qu'il soit impossible d'avoir une expression littérale de la valeur à assigner à λ_i, la recherche peut être menée efficacement grâce à une dichotomie.

4 Agencement des briques lors d'une recherche

Nous avons expliqué les raisons qui nous ont menés à effectuer une recherche systématique sur toutes les briques de base. Les facteurs principaux sont la simplicité, le temps encore raisonnable de recherches (de quelques secondes à quelques minutes selon les briques) et la garantie d'obtenir le meilleur calibrage possible, une propriété particulièrement intéressante au stade exploratoire de ce projet. En revanche pour les modèles du second degré, le nombre exorbitant de combinaisons possibles entre les paramètres rend impossible une recherche purement systématique. C'est pourquoi on fait appel à une recherche plus orientée des paramètres. Autrement dit, pour ces briques-là, l'espace de recherche est inclus dans celui circonscrit par les bornes. Enfin notons qu'importe l'approche, les recherches s'exécutent par itération successives des combinaisons des paramètres.

Pour les courbes de performances croissantes, la recherche s'opère de la façon suivante. On parcourt l'ensemble des briques de base disponibles. Pour chaque brique, on retient les 3 meilleurs calibrages (qui correspondent à des modèles). On établit le classement des 3 meilleurs calibrages toutes briques confondues et un récapitulatif comprenant le meilleur modèle possible pour chaque famille de briques. Si les résultats sont jugés « bons », le processus de recerche est terminé. Dans le cas contraire, il nous reste la chance de trouver un « bon » calibrage avec les modèles imbriqués.

On note que nous n'avons pas suggéré l'utilisation des modèles multiclasses. Cette absence s'explique par le fait qu'ils constituent l'unique candidat pour les courbes de performance décroissantes et sont inaptes à reproduire une courbe croissante.

5 Optimisation

Les efforts d'optimisation sur la recherche du meilleur calibrage constitue un des objectifs des travaux futurs. Toutefois certaines formes d'optimisation sont d'ores et déjà implémentées. La principale, celle que nous présentons ici, s'appuie sur un heuristique se basant sur le calcul des μ locaux. Prenons l'exemple de la brique M/M/C.

5.1 M/M/C

Le processus de recherche démarre par une boucle sur C (éventuellement précédée d'une boucle sur Off). On sait qu'il suffit de choisir pour tout i, $\lambda_i = \bar{X}_i$ pour s'assurer que $\bar{X}_{i,anal} = \bar{X}_i$. Une fois la valeur de C fixée, on

recherche pour chaque niveau de charge λ_i en entrée, la valeur de μ qui permettrait de reproduire exactement le temps de réponse moyen mesuré $\bar{R}_i$. Il n'est pas toujours possible de trouver une valeur de μ qui convienne mais en revanche une propriété remarquable, commune à toutes les briques, assure son unicité si elle existe. Cette propriété de monotonie : plus μ augmente, plus le temps moyen de séjour $\bar{R}_{anal}$ d'un client diminue, autorise de surcroît une recherche efficace, par dichotomie, de la valeur de μ_i. Par conséquent après avoir mené ce calcul sur μ_i pour chaque point de mesures i, on dispose de n valeurs différentes de μ_i (sauf si les mesures ont vraiment été générées à partir d'une M/M/C). On nomme cet ensemble de n valeurs, l'ensemble des μ locauwx.

C'est là qu'intervient notre heuristique. Si les mesures avaient été générées à partir d'une M/M/C, même légèrement bruitée, les valeurs des μ locaux seraient (quasi-) identiques. Par conséquent, on supposera que si la dispersion (on calcule l'écart-type) entre les μ locaux est trop « grande » alors c'est que nos mesures n'ont pas été générées par un système assimilable à une M/M/C avec la valeur de C fixée précédemment. On décide donc d'écarter toute tentative d'essais cette valeur du paramètre C. Au final on a donc pu écarter un grand nombre de modèles sans pour autant les éprouver réellement (calcul des temps de séjour moyens, de la fonction distance, ...). De plus si l'indice de dispersion est favorable, les μ locaux obtenus sont utilisés pour affiner les bornes sur le paramètre μ. En pratique ces nouvelles bornes se révèlent souvent plus restrictives, donc plus efficaces que les précédantes.
Une fois l'optimisation réalisée, il reste à parcourir la boucle sur μ et à déterminer l'ensemble des λ_i qui, injectés dans ledit modèle reproduisent exactement les débits mesurés. Enfin il suffit de calculer le temps de séjour $\bar{R}_{i,anal}$ associé à chaque niveau de charge λ_i et d'estimer la qualité de chaque modèle à l'aide de la fonction distance.

5.2 M/M/C/K

Pour la M/M/C/K, la situation est plus compliquée. On pourrait pour des valeurs de C et K fixées, chercher les niveaux de charge λ_i permettant de reproduire exactement les débits moyens mesurés $\bar{X}_i$ ($\bar{X}_{i,anal} = \bar{X}_i$), et puis trouver les valeurs des μ_i locaux qui permettent d'avoir également ($\bar{R}_{i,anal} = \bar{R}_i$). On calculerait alors l'indice de dispersion entre les μ locaux et on reprendrait le procédé exposé pour la M/M/C. Cette façon de faire fonctionnerait mais une solution plus efficace permet d'arriver aux mêmes résultats.
Cette alternative s'appuie sur la variable $\rho = \frac{\lambda}{\mu}$ et sur le nombre moyen de clients mesuré $\bar{Q}_i = \bar{R}_i\bar{X}_i$. Une des propriétés remarquables aux M/M/C/K [22] est que plus ρ augmente, plus $\bar{Q}$ augmente. Nous sommes donc en mesure de trouver pour chaque point de mesures i, l'unique valeur de ρ_i qui engendre un nombre de clients moyens $\bar{Q}_{i,anal}$ égal à celui mesuré $\bar{Q}_i$. Une fois les n valeurs de ρ_i obtenues, il nous est possible de calculer la probabilité de la file d'avoir K clients pour chacune des valeurs de ρ_i. En effet les probabilités stationnaires d'une M/M/C/K dépendent uniquement du quotient $\frac{\lambda}{\mu} = \rho$.
Puisqu'on dispose des $p(K)_i$ et des ρ_i et que la définition des pertes sur une M/M/C/K se traduit par : $\lambda_i = \frac{\bar{X}_i}{1-p(K)_i}$, on est en mesure d'obtenir pour

[22] Se reporter à la partie 4

chaque point de mesures i, un couple (λ_i, μ_i) qui permet de reproduire pour la M/M/CK à l'essai les valeurs exactes de $\bar{X}_i$ et $\bar{R}_i$ mesurées.
Naturellement ces λ_i et ces μ_i correspondent à des paramètres locaux, propres à un point de mesure, et ne constituent pas de résultats définitifs. En revanche comme pour l'exemple de la M/M/C précédent, la connaissance des μ_i permet d'améliorer très sensiblement les bornes de l'espace de recherche sur μ (grâce à notre heuristique sur les μ locaux). Une fois l'optimisation réalisée, il reste à parcourir la boucle sur μ et à déterminer l'ensemble des λ_i qui, injectés dans ledit modèle reproduisent exactement les débits mesurés. Enfin il suffit de calculer le temps de séjour $\bar{R}_{i,anal}$ associé à chaque niveau de charge λ_i et d'estimer la qualité de chaque modèle à l'aide de la fonction distance.

5.3 Généralisation

Les mécanismes d'optimisation exposés précédemment peuvent être répliquées sur chacune des briques de nos travaux. En effet, toutes exhibent des comportements remarquables de monotonie vis-à-vis de μ et de λ. Plus formellement, l'optimisation sur les μ locaux consiste à vérifier si une fois les $p-1$ premiers paramètres d'un modèle fixés, il est possible de trouver un compromis sur la valeur du paramètre restant qui permette d'avoir à la fois d'avoir pour tout i, $\bar{X}_{i,anal}$ égal à $\bar{X}_i$ et $\bar{R}_{i,anal}$ proche de $\bar{R}_i$. L'heuristique pilier de cette optimisation est que si le système est asimmilable au modèle à l'essai complété d'une valeur pour le dernier paramètre, alors ce compromis existe. En revanche, si ce compromis n'existe pas, nous supposerons que quelque soit la valeur du dernier paramètre, le modèle à l'essai ne sera pas « bon ».

Enfin l'apport de cette optimisation dans le processus de recherche est difficile à quantifier mais nos tests répétés ont mis en évidence de nets gains en performance sans détériorations des résultats.

Septième partie

Améliorations à apporter et travaux futurs

1 Obtenir de nouveaux jeux de mesures

Un de nos objectifs est de parvenir à se constituer un répertoire de jeux de mesures riche et varié. Ce répertoire comportera des jeux de mesures artificielles et autant que possible des jeux de mesures réelles provenant de domaines divers. Dans un premier temps, nous allons rechercher prioritairement des jeux mesures recensant des temps de séjour en fonction du débit. Nous privilégierons également les mesures relatives au domaine des réseaux informatique car c'est dans ce secteur que les compétences globales de l'équipe sont le plus affutées et surtout car nous ne disposons jusqu'à présent d'aucun jeux de mesures dans ce secteur. La confection de ce répertoire servira plusieurs intérêts comme nous le verrons dans la suite de cette partie.

2 Eprouver les briques actuelles

Afin de vérifier la robustesse de nos briques, de circonscrire l'étendue de la grammaire que forme leur ensemble et d'orienter nos recherches futures, nous allons avoir besoin d'un répertoire de jeux de mesures assez conséquent. L'analyse des résultats nous conduira peut être à réviser l'ensemble des briques disponibles soit par l'intégration de nouvelles briques, soit par l'éviction de certains briques jugées redondantes avec d'autres modèles.

3 Evincer certaines briques

Cette étape consistera à vérifier la pertinence de la sélection du modèle fermé avec rejet dans notre ensemble de briques. Par exemple le comportement d'un modèle fermé avec rejet semble être reproductible aussi finement que souhaité par un simple modèle ferme. Si tel est le cas, alors le modèle fermé avec rejet sera retiré de notre ensemble de briques car son faisceau de courbes de performances est inclus dans celui du modèle fermé, un modèle plus simple donc meilleur.

4 Définir de nouvelles briques (avec précaution)

Le pendant de la tâche précédente est de définir de nouvelles briques. La définition d'un nouveau modèle s'impose uniquement lorsque le comportement d'un système (jugé représentatif) ne peut fondamentalement pas être correctement reproduits par les briques existantes. La sélection des modèles candidats à l'ensemble des briques qui répondent au besoin seront sélectionnés sur leur simplicité (nombre de paramètres et analyse), sur la complémentarité de leur faisceau de courbes de performances avec les modèles existants et sur l'in-

terprétation physique que l'on peut en faire. A ce sujet nous avons déjà quelques pistes dont certaines ont été suggérées au cours de ce rapport.

4.1 Expérimenter une nouvelle forme de modèles imbriqués

Les résultats obtenus par les modèles imbriqués sont bons mais doivent pouvoir être perfectionnés [23]. La solution préconsiée consiste à modifier le modèle interne afin de borner le temps d'attente subi par un client au sein de ce dernier. Cela peut se faire en remplaçant la M/M/1 par une M/M/1/K ou bien par un modèle fermé. La solution exposée se justifie de la façon suivante.
Dans les modèles imbriqués (tels que nous les avons implémentés), pour une charge λ qui croît et demeure éloignée du seuil de saturation, le modèle interne voit son temps d'attente augmenter très progressivement (ce qui permet d'avoir un début de courbe de performances « pentu » comme le sont ceux des jeux de mesures à reproduire), puis lorsque λ se rapproche des taux de charge maximums alors intervient la saturation du bloc interne qui impactera et conduira à la saturation du bloc principal. Le problème tient dans le fait que plus λ croît, plus le temps d'attente dans le bloc interne devient long et donc plus le taux de service du bloc principal devient petit : $\mu_1 = \frac{1}{S_1} = \frac{1}{S_0 + a \times W_2}$. Or puisque le modèle interne doit commencer à saturer « tôt », arrivé à des taux de charge élevés, le temps d'attente dans la file 2 devient démesuré. La reaction obligatoire mais néfaste de notre programme est la suivante : il accroit considérablement la valeur de C_1 afin de maintenir le produit $C_1\mu_1$ au dessus du seuil de stabilité du modèle.
Voilà pourquoi on espère endiguer ce phénomème d'accroissement sur C_1 en suppléant la M/M/1 à capacité illimitée par un modèle à perte ou par un modèle fermé dont les temps de séjour et donc d'attente sont bornés.

4.2 Expérimenter des modèles multiclasse plus sophistiqués

De nouveaux degrés de liberté (jusqu'à présent l'implémentation décidée n'en dispose que de deux) seront probablement nécessaires pour élargir le spectre de courbes de performances que peut reproduire un modèle multiclasse. Cela peut passer par la différenciation des taux de services moyen du serveur selon la classe ($\mu_1 et \mu_2$), en introduisant un cv^2 différent de 1 (dont la valeur serait comprise approximativement dans l'intervalle [0.5 ;2]), et éventuellement un cv_1^2 et cv_2^2 .

5 Estimer le pouvoir prédicitf des modèles retenus

Il s'agirait de tester la capacité des modèles retenus à prédire les performances du système. Cette exercice peut s'opérer trivialement sur un jeu de mesures volontairement incomplet pour lequel on a retiré un point de mesure. Ou plus probant encore, après un changement de la charge qui l'éloigne du comportement observé sur l'ensemble des mesures. Et enfin bien plus fort mais bien plus dur, après un changement structurel du système. Hormis pour le premier cas, le verrou est l'obtention d'un double jeu de mesures (avant et après) qui

[23] Une des interrogations majeures concernent la valeur excessivement élevée de C_1

permet de mettre en évidence la qualité prédictive des modèles retenus par nos travaux.

6 Autoriser un Offset sur les débits ?

L'idée est d'étudier la faisabilité et l'intérêt de transposer l'idée de l'Offset sur les temps de séjour aux mesures du débit. Autrement dit, il s'agit de permettre au programme de chercher le meilleur modèle pour les mesures de débit $\overline{X}$ à une constante négative près (prise dans l'intervalle $[0\,;\min(\overline{X_i})]$). L'interprétation d'un tel Offset pourrait se rapprocher d'un modèle dans lequel le débit mesuré en sortie est la résultante (avec pertes éventuellement) de la somme de deux débits en entrée : celui régi par la variable λ et une charge constante exogène.

Pour les modèles ouverts sans perte (type M/M/C), cet Offset officierait le rôle d'un trafic exogène qui se mèlerait en entrée du modèle au trafic de mesures. En sortie, tout le trafic quelque soit son origine est mesuré. Bien entendu, l'intensité du trafic exogène sera à prendre en considération pour le calcul de la contrainte de stabilité du modèle. On envisage de confronter les performances de ce type de modèles sur les jeux de mesures 2 et 4 qui avaient amené à l'idée des modèles imbriqués.
Pour les modèles ouverts à perte (type M/M/C/K), la démarche de l'Offset sur le débit risque d'être plus compliquée car même si la charge exogène est constante en entrée du modèle, son débit en sortie varie en fonction du niveau de λ. Plus λ augmente, plus la source exogène risque d'essuyer des rejets.

7 Rechercher intelligement les paramètres - Optimiser

Il s'agit d'un effort d'optimisation. Ce volet du projet devrait être mené en partenariat avec l'équipe RO dans le cadre d'un projet LIP6. Le but consiste à remplacer la recherche itérative par une recherche orientée pour gagner en rapidité tout en maintenant la précision des résultats.
L'approche itérative simple à implémenter, exhaustive, sûre et générant des temps de calcul raisonnables constitue une solution adaptée à nos exigences et à nos besoins à ce stade du projet, c'est-à-dire pour l'amorce du projet mais pas pour la solution finale.

Une première étape intermédiaire, pourrait être de rechercher le calibrage par degré de précision successifs. Le meilleur intervalle trouvé à une granularité modérée serait conservé et hébergerait en son sein une nouvelle recherche avec une granularité plus fine. Si ce mécanisme récursif est appliqué par exemple 3 fois en découpant l'intervalle en 10 sous-intervalle, on aurait pour un coût de 30 essais, une précision aussi fine que si l'on avait exécuté 10^3 évaluations de modèles (sous réserve que le meilleur calibrage appartienne bien à chaque tour au sous-intervalle retenu).

Plus fort, on a déjà observé une rupture dans la méthodologie de recherche pour la classe des modèles imbriqués. On ne parcourt pas l'ensemble de l'espace circonscrit par les bornes mais le voisinage des valeurs probables. Ainsi ce type d'approche nécessite plus d'intuition sur les résultats, plus de compréhension sur les modèles pour orienter habilement la recherche. Dans ce sens, les influences

des paramètres sur les courbes de performances représentent une information capitale. On pourrait par exemple définir la méthode de calibrage automatique pour un modèle FERME de la façon suivante (en admettant qu'en entrée, on dispose d'une courbe de mesures bimodale) :

- Exécuter une boucle sur le paramètre d'Offset entre $[0\,;\min_{i\in[1;n]}(\bar{R}_i)]$
- Fixer C en fonction du rayon de courbure observé sur la courbe de performances des mesures
- Fixer μ afin de situer le résultat du produit de $C\mu$ près (mais au dessus) du seuil de saturation enregistré par les mesures
- Fixer Off afin d'avoir le bon temps de réponse à charge quasi-vide (encore faut il le savoir que la mesure a été prise dans ces conditions de charge)
- Fixer N de manière à ce que le résultat de $\frac{N}{C\mu}$ soit égale au point de saturation observé sur la courbe de mesures (encore faut il savoir quelles mesures correspondent à des points de saturation du système)

On remarque que l'ordre d'apparition des paramètres est modifié par rapport à la recherche exhaustive et que les bornes des paramètres (excepté pour Off) n'officient plus le rôle de délimitateur de l'espace de recherche mais se contentent d'assurer la cohérence des modèles testés.

Une fois ce premier vecteur de paramètres connus, des techniques mathématiques peuvent venir assister la recherche du meilleur calibrage. Citons à titre d'exemple les approches par voisinage, par gradient, ...

8 Etendre nos travaux à d'autres ensemble de mesures

Ici l'objectif serait de dépasser le contexte restreint du stage. En effet comme nous l'avons dit dans la deuxième partie de ce rapport, l'objectif originel et ultime de ce projet consiste à générer automatiquement un modèle simple et calibré qui reproduit « au mieux » le comportement d'un système quelconque avec pour seule source d'infirmation un n-uplets de mesures sur des paramètres de performance variés. On pourrait par exemple supposer devoir mener cette étude à partir d'un jeu de mesures composé de probabilités de rejet P_r et du nombre de clients Q mesurés sur le système.

Conclusion

Le sujet de ce stage avait véritablement un caractère exploratoire. La justesse de la logique sous-jacente bien que partiellement validée demeure à vérifier (ou à réfuter) dans son intégralité. Par essence donc ce sujet offrait beaucoup de libertés dans sa réalisation et il revenait aux protagonistes de ce projet de s'accorder sur le choix des orientations notamment. Cette propriété prospective inhérente à nos travaux rend ce sujet particulièrement attractif et intéressant mais en revanche participe à sa complexité.

Au cours de ce rapport, nous nous sommes efforcés de véhiculer l'idée que les directions empruntées par nos travaux n'ont pas toutes un caractère définitif. En effet il va de soi que certains choix seront appelés à être révisés et corrigés dans les semaines, dans les mois à venir. La raison est que nos avancées sont le fruit d'une conjugaison étroite d'intuition, de savoir-faire et de connaissances empiriques (occasionnées au cours de ces cinq mois). A ce titre soulignons que ce sont essentiellement les résultats (même intermédiaires, et surtout infructueux) qui ont conditionné et conditionneront les recherches à venir.

Les travaux engagés et les résultats obtenus laissent entrevoir des perspectives prometteuses. Toutefois de nombreux travaux restent à faire afin. La première grande étape consistant à étendre le champ d'application de nos travaux (à des ensembles plus variés de paramètres de performances) et à poursuivre la validation de la méthodologie par des expérimentations sur des systèmes divers et variés.

Références

[1] P. Heidelberger and S. Lavenberg. Computer performance evaluation methodology. *IEEE Trans. Computers*, 33(12) :1195–1220, 1984.

[2] B. Baynat. *Theorie des files d attente*. Hermes, 2000.

[3] S. Fdida and G. Hebuterne. *Methodes exactes d'analyse de performance des reseaux*. Hermes, 2004.

[4] K. Salamatian and S. Fdida. A framework for interpreting measurement over internet. In *MoMeTools '03 : Proceedings of the ACM SIGCOMM workshop on Models, methods and tools for reproducible network research*, pages 87–94, New York, NY, USA, 2003. ACM Press.

[5] L. Kleinrock. *Queueing Systems*, volume I : Theory. Wiley-Interscience, 1975.

[6] L. Kleinrock. *Queueing Systems*, volume II : Computer Applications. Wiley-Interscience, New York, 1976.

[7] K. Salamatian and S. Vaton. Hidden markov modeling for network communication channels. In *SIGMETRICS/Performance*, pages 92–101, 2001.

[8] R. Jain. *The Art of Computer Systems Performance Analysis*. John Wiley and Sons, Inc., 1991.

[9] M. Crovella, C. Lindemann, and M. Reiser. Internet performance modeling : the state of the art at the turn of the century. *Perform. Eval.*, 42(2-3) :91–108, 2000.

[10] D. Schiller. System capacity and performance evaluation. *j-IBM-SYS-J*, 19(1) :46–67, 1980.

[11] S. Alouf, P. Nain, and D. Towsley. Inferring network characteristics via moment-based estimators. In *INFOCOM*, pages 1045–1054, 2001.

Huitième partie
Annexes

A Implémentation en C

A.1 Fonctionnalités

Pour pouvoir mener à bien nos recherches, nous avons développé 3 programmes distincts, tous écrits en C. Les fonctionnalités de chacun de ces programmes sont décrites ci-dessous :

- Générer des jeux de mesures bruitées en instanciant un modèle de nos briques.
- Tracer des courbes de performances en instanciant un ou plusieurs modèles de nos briques.
- Rechercher pour un jeu de mesures les meilleurs calibrages de modèles parmi les briques, c'est l'application principale.

A.2 Exigences et difficultés

La programmation de ce logiciel ne présente pas véritablement d'exigences particulières. Toutefois la dialectique du projet, qui s'appuie sur un ensemble de briques que l'on parcourt successivement, induit un code fonctionnant par modules. A chaque module logiciel correspond une brique de notre réservoir.
La difficulté la plus significative du développement de ce programme provient du risque des débordements numériques (valeur de la variable dépasse la taille maximale autorisée). Les calculs des performances de tous les modèles présentés, exceptées celles du modèle multiclasse, sont éffectués à partir de la chaîne de Markov associée au modèle. Le procédé de calcul, simple et efficace, se réalise récursivement. Toutefois, en démarrant le calcul aveuglement, on risque, d'obtenir en bout de chaîne (par récursion successive), des valeurs de probabilités d'états infinitésimalement grandes. Ce problème ne peut avoir lieu que dans les modèles limités soumis à une forte charge. L'astuce pour le contourner est d'attaquer la chaîne par le bon côté. Si $\rho > C$, on s'attend à ce que l'état p(K) soit beaucoup plus probable que l'état 0. On fixe alors p(K)=1 provisoirement, et on calcule par récurisivité toutes les probabilités d'état (le calcul des probabilités s'arrête dès qu'une probabilité passe sous une valeur seuil). Puis suit logiquement l'étape de normalisation. Inversement, si $\rho < C$, alors le processus de calcul démarrera par le calcul de l'état 1, on fixe p(0)=1.

A.3 Présentation des résultats

L'affichage des résultats requiert l'utilisation d'un éditeur de graphes. Nous avons opté pour Gnuplot, dont l'utilisation est très répandue et les fonctionnalités très nombreuses. Ainsi notre logiciel édite des fichiers textes en code ASCII, contenant des couples de débit et de temps de séjour, conformes à la syntaxe Gnuplot.

B Calcul du temps de séjour dans une M/G/1 à ordonnancement HOL

Cette section est très largement inspirée par [5] [24] et apporte les explications suffisantes pour comprendre les formules des temps d'attente présentées dans la section se rapportant aux modèles multiclasses. On rappelle que l'on suppose deux classes de clients, avec priorité sans préemption de la classe 2 sur la classe 1.

Terminologie et variables

On nomme W_i l'attente subie par les clients de la classe i dans la file.
Le temps total passé dans la file pour un client de la classe i est défini par $T_i = W_i + \frac{1}{\mu}$. Pour chaque classe i, le temps d'attente dans la file d'un client W_i peut être décomposé en 3 parties :

- Le délai dû au client en cours de traitement lors de son arrivée.
- Le délai dû aux clients de sa classe ou d'une classe plus prioritaire déjà en attente.
- Le délai dû aux clients d'une classe plus prioritaire qui entreront dans la file entre son instant d'arrivée et le début de son temps de service.

On définit la variable $N_{i,p}$ comme le nombre moyen de clients de la classe i déjà présents dans la file d'attente et dont l'ordre de service ne sera pas perturbé lorsqu'un un client (de classe p) arrive. Puisque notre modèle comporte seulement deux classes, il existe en tout quatre combinaisons de $N_{i,p}$ à définir. On profite également du fait que les processus d'arrivées des clients sont Poissonniens ce qui permet de connaître facilement le nombre moyen de clients vu par un client lorsqu'il entre dans la file (ASTA).

- $N_{2,1} = \lambda_2 \times W_2$ (en appliquant Little et sachant que X_i=λ_i car le modèle est sans perte, ni saturé)
- $N_{1,1} = \lambda_1 \times W_1$
- $N_{2,2} = \lambda_2 \times W_2$
- $N_{1,2} = 0$

On définit également $M_{i,p}$ comme le nombre moyen de clients de la classe i qui arriveront après un client de classe p et qui seront servis avant.

- $M_{2,1} = \lambda_2 \times W_1$ (d'après la loi du robinet)
- $M_{1,1} = 0$
- $M_{2,2} = 0$
- $M_{1,2} = 0$

Enfin ρ_i est défini comme le ratio de λ_i et de μ_i. Pour donner un sens physique à ρ_i, il correspond à la fraction de temps où le serveur est occupé par un client de classe i (à condition que ρ_i soit inférieur à 1). Quant à $\rho = \frac{\lambda}{\mu}$ avec $\lambda = \sum_{p=1}^{2} \lambda_p$, il reflète la proportion de temps où le serveur est occupé à traiter un client quelle que soit sa classe (à condition que $\rho < 1$).

Calcul des temps d'attente W_p dans la file [6]

Le temps d'attente d'un client de classe p est égal à la somme du nombre de clients placés avant lui lors de son entrée dans la file et du nombre de client

[24] page 169

s'insérant devant lui au cours de son attente rapportée à la vitesse d'éxécution du serveur. A ce temps d'attente s'ajoute un autre temps d'attente W_0 qui correspond au délai moyen pour terminer le service du client en cours (s'il existe).

$W_p = W_0 + \sum_{i=1}^{2}(N_{i,p} + M_{i,p}) \times \frac{1}{\mu}$ pour tout p $\in$ [1 ;2] avec W_0 le temps résiduel de l'éventuel client en cours de service.

ainsi : $W_p = W_0 + \sum_{i=p}^{2} \frac{\lambda_i \times W_i}{\mu} + \sum_{i=p+1}^{2} \frac{\lambda_i \times W_p}{\mu}$ pour tout p $\in$ [1 ;2]

soit : $W_p \times (1 - \frac{\lambda_p}{\mu} - \sum_{i=p+1}^{2} \frac{\lambda_i}{\mu}) = W_0 + \sum_{i=p+1}^{2} \frac{\lambda_i \times W_i}{\mu}$

soit : pour tout $p \in [1; 2]$,

$$W_p = \frac{W_0 + \sum_{i=p+1}^{2} \rho_i \times W_i}{1 - \sum_{i=p}^{2} \rho_i} \tag{3}$$

Le calcul complet des W_p s'opère en deux temps distincts : le calcul de W_0 et le calcul récursif des W_p. Nous traiterons d'abord le cas de W_0 puis nous présenterons le procédé général pour calculer les W_p.

1 - Calcul de W_0 - le paradoxe du temps de vie résiduel

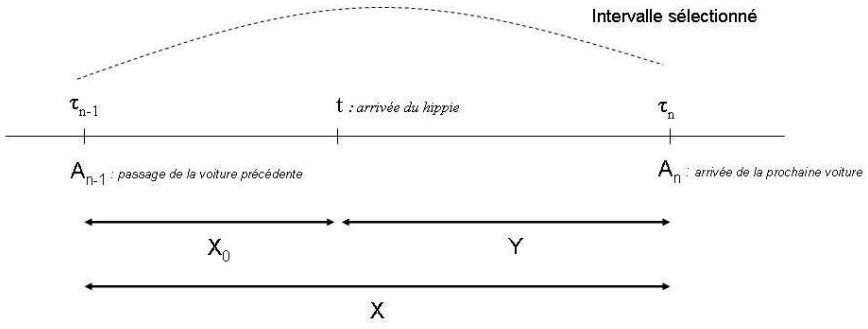
Ce sujet est largement inspiré de [5] [25]. La situation est la suivante : un client arrive dans une file et trouve en entrant dans la file un client partiellement servi par le serveur. Combien de temps va prendre le temps de service résiduel ?
Exemple symptomatique : un hippie arrive à un instant arbitraire sur le bord d'une route et se met à faire du stop. Les voitures passent devant lui selon un processus que l'on supposera Poissonnien de taux moyen λ voitures par minute. On suppose que la première voiture qui passe décide de le faire monter. Combien de temps doit en moyenne attendre le hippie pour qu'une voiture passe ?
Il existe deux façons de mener le calcul :

- Le temps de moyen entre deux passages successifs de voitures est égal à $\frac{1}{\lambda}$ minutes. Et puisque l'instant d'arrivée du hippie est aléatoire, il attendra en moyenne $\frac{1}{2\times\lambda}$ minutes.
- Puisque le processus d'interarrivée est sans mémoire, l'instant d'arrivée de la prochaine voiture est totalement indépendant du temps écoulé depuis le passage de la voiture précédente. Du coup le hippie devra attendre $\frac{1}{\lambda}$ minutes. Or le même raisonnement peut être répété pour estimer le temps écoulé entre le passage de la voiture précédente et l'arrivée du hippie. Ainsi l'intervalle de temps entre le passage des 2 voitures précédant et suivant respectivement l'arrivée du hippie est de $\frac{2}{\lambda}$ minutes. Ce résultat est le double de celui attendu pour un processus de Poisson[26].

Définition des variables - Soit F(x)=P$[\tau_{k+1} - \tau_k \leq x]$, la distribution des intervalles $[\tau_{k+1} - \tau_k]$ que l'on suppose indépendants et identiquement distribués. La fonction de densité de probabilité aossicée (pdf) est définie elle comme f(x)=$\frac{dF(x)}{dx}$.
On aura pour but de rechercher une expression des pdf de X et Y, et il sera surprenant de constater que X n'est pas distribué selon F(X) !
Pour Y, on définit :

* $\hat{F}(x) = \text{P}[Y \leq x]$

[25] page 169
[26] il s'avérera que c'est cette solution-là qui est juste



X: temps de vie
X_0: âge
Y: temps de vie résiduel

FIG. 46 – Temps de vie, âge et temps de vie résiduel

* $\hat{f}(x) = \frac{d\hat{F}(x)}{dx}$

Pour X, on définit :

* $F_X(x) = \mathrm{P}[X \leq x]$
* $f_X(x) = ???$

Première partie - Or $(F_X(x+dx) - F_X(x)) = f_X(x)dx = Kxf(x)dx$ puisque la probabilité qu'un intervalle de taille x soit choisi est proportionnelle à la taille x de cet intervalle ainsi qu'à sa fréquence d'occurence : $\mathrm{P}[x < X \leq x + dx] = F(x + dx) - F(x) = f(x)dx$. K est une constante réelle positive. En intégrant cette relation, on obtient l'égalité suivante : K$\times m_1 = 1$. Soit K=$\frac{1}{m_1}$ avec $m_1 = \mathrm{E}[\tau_k - \tau_{k-1}]$, soit le temps moyen entre deux renouvellements.
Au final on a donc montré que la pdf associée à l'intervalle choisie s'exprime comme : $f_X(x) = \frac{xf(x)}{m_1}$.

Deuxième partie - On cherche une expression pour $\hat{f}(x)$.
On sait que $\mathrm{P}[Y \leq y \mid X = x] = \frac{y}{x}$ pour tout y $\in [0; x]$. Cette égalité vient du fait que le point correspondant à l'instant d'arrivée du client est sélectionné aléatoirement dans l'intervalle retenu et que donc sa distribution y est uniformément répartie.
On en déduit la distribution jointe de X et de Y :
$\mathrm{P}[y < Y \leq y + dy \ , x < X \leq x + dx] = (F[y + dy] - F[y]) \times (F_X[x + dx] - F_X[x])$
$\mathrm{P}[y < Y \leq y + dy \ , x < X \leq x + dx] = (\frac{y+dy}{x} - \frac{y}{x}) \times (\frac{xf(x)dx}{m_1})$
$\mathrm{P}[y < Y \leq y + dy \ , x < X \leq x + dx] = (\frac{dy}{x})(\frac{xf(x)dx}{m_1})$
Donc $\mathrm{P}[y < Y \leq y + dy \ , x < X \leq x + dx] = \frac{f(x)dydx}{m_1}$ pour tout y $\in [0; x]$.
En intégrant cette expression sur x, on obtient : $\hat{f}(y)dy = \int_{x=y}^{\infty} \frac{f(x)dydx}{m_1}$

Soit : $\hat{f}(y) = \frac{1-F(y)}{m_1}$. Cette formule donne une expression simple pour la pdf du temps de vie résiduel de l'intervalle fonction de la taille de l'intervalle et de sa moyenne.

Transformation de Laplace - Puisque toutes les fonctions considérées sont nulles avant 0 et sont à valeurs positives, $\hat{f}(y)$ peut s'écrire $\frac{1}{m_1} \times step(t) - \frac{1}{m_1}\int_0^y f(t)dt$ on a alors : $\hat{F}^*(s) = \frac{1-F^*(s)}{s \times m_1}$

Calcul général des moments du temps de vie résiduel en fonction des moments du temps de vie lui-même - Soit m_n le moment d'ordre n du temps de vie et r_n celui d'ordre n pour le temps de vie résiduel. Par définition, $m_n = \mathrm{E}[(\tau_k - \tau_{k-1})^n]$ et $r_n = \mathrm{E}[Y^n]$
La loi de l'Hôpital permet d'obtenir la relation suivante : $r_n = \frac{m_{n+1}}{(n+1) \times m_1}$
Soit pour le premier moment, $r_1 = \frac{m_2}{2 \times m_1}$ ou encore $\frac{m_1}{2} + \frac{\sigma^2}{2 \times m_1}$ où $\sigma^2 = m_2 - m_1^2$ représente la variance.

Interprétation des résultats - Il apparait clairement que la première solution trouvée pour le calcul du temps d'attente pour le hippie ($\frac{m_1}{2}$) est juste si la variance des intervalles est égale à 0, c'est-à-dire si les arrivées sont régulières dans le temps.Or pour un processus Poissonnien de taux λ, $m_1 = \frac{1}{\lambda}$ et $\sigma^2 = \frac{1}{\lambda^2}$. D'où $r_1 = \frac{1}{\lambda} = m_1$.

Calcul du W_0 rapporté à notre problème - Ainsi, le temps moyen du temps de vie résiduel qui correspond au moment d'ordre 1 du temps de service restant , W_0 ici égale : $\frac{m_2}{2 \times m_1}$. Or puisque tous les processus d'arrivées sont poissonniens, la probabilité de chance que le client en cours de service soit de la classe i est égale à la proportion de temps où le serveur traite un client de classe i.
Soit $W_0 = \sum_{i=1}^{2} \rho_i \times \frac{m_{2i}}{2 \times m_{1i}}$
avec $m_{2i} = \frac{2}{\mu^2} = 1$ (car le temps de service est supposé exponentiel) et $m_{1i} = \frac{1}{\mu}$. (On vérifie que le $cv^2 = \frac{m_2 - m_1^2}{m_1^2}$ est bien égal à 1.)
Ainsi

$$W_0 = \sum_{i=1}^{2} \rho_i \times \frac{m_{2i}}{2 \times m_{1i}} = \sum_{i=1}^{2} \frac{\rho_i}{\mu_i} \tag{4}$$

2 - Calcul récursif des W_p pour p > 0

A partir de la formule (3), on peut calculer tous les W_p récursivement. Il suffit de calculer d'abord W_P et puis d'en déduire W_{P-1} et ainsi de suite.
D'après Kleinrock, W_p peut alors s'exprimer de la façon suivante : $\frac{W_0}{(1-\sigma_p) \times (1-\sigma_{p+1})}$
pour tout p $\in$ [1 ;2] avec $\sigma_p = \sum_{i=p}^{P} \rho_i$
On remarque entre autres qu'il est possible d'avoir des W_p finis pour certaines classes d'indices p supérieurs à un seuil critique tandis que les autres classes de priorités inférieures peuvent présenter des temps d'attente instables (non bornés).

C Jeux de mesures

C.1 Sur les contrôleurs disques

On dispose de 4 jeux de mesures réelles fournis par Alexandre. Les mesures ont été réalisées sur des contrôleurs d'E/S de disques durs. Les débits sont exprimés en nombre de requêtes par milli-seconde et les temps de séjour en seconde. Chaque tableau ci-dessous relate les mesures associées à un scénario.

C.1.1 Jeu de mesures 1

Débit mesuré (req/sec)	*Temps de séjour mesuré (msec)*
14.851	0.64
18.182	0.65
21.275	0.68
24.129	0.70
25.471	0.72
26.712	0.74
27.972	0.77

TAB. 23 – Jeu de mesures 1

C.1.2 Jeu de mesures 2

Débit mesuré (req/sec)	*Temps de séjour mesuré (msec)*
3.875	6.52
9.526	8.73
10.438	9.08
10.781	9.42
11.156	9.78
11.337	9.97
11.492	10.18

TAB. 24 – Jeu de mesures 2

C.1.3 Jeu de mesures 3

Débit mesuré (req/sec)	*Temps de séjour mesuré (msec)*
14.847	0.64
18.131	0.65
21.199	0.68
24.011	0.71
25.330	0.72
26.623	0.74
27.766	0.76

TAB. 25 – Jeu de mesures 3

C.1.4 Jeu de mesures 4

Débit mesuré (req/sec)	*Temps de séjour mesuré (msec)*
1.943	11.89
3.873	14.03
4.256	15.01
4.449	15.28
5.021	17.59
5.402	20.80
5.756	28.19

TAB. 26 – Jeu de mesures 4

C.2 Sur les réseaux TCP/IP

Nous disposons également de deux jeux de mesures, 5 et 6, qui correspondent à des relevés de débits et de temps de réponse sur un réseau TCP/IP. Ces jeux de mesures mettent en évidence le comportement de TCP qui vise à partager équitablement la bande passante entre les flots sui rivalisent pour accéder aux ressources du réseau. Le protocole TCP fixe une borne supérieure sur le débit moyen qu'un flux peut émettre dans le réseau. Cette borne est fonction de son taux de perte et de son RTT comme l'indique la formule $\bar{X}_{max} = \frac{1.22MTU}{\sqrt{L} \times RTT}$.

Le procédé opératoire pour récolter les mesures s'est appuyé sur une plate-forme de test. Le lien réseau qui servira pour les mesures est soumis à un trafic TCP qui occupe toute la bande passante disponible lorsque cela lui est autorisé. Un logiciel permet de contrôler le temps de traversée de ce lien. Puis un trafic de test (lui aussi TCP) dont le débit autorisé peut dépasser la bande passante du lien est envoyé sur ce même lien. Les deux flux rivalisent donc et le protocole TCP va fixer le partage des ressources entre eux. L'étape des mesures va consister à faire varier (artificiellement) le RTT du flux de test et d'observer alors quel est le débit qui s'ensuit. Voici comme on obtient des couples de valeurs $(\bar{R}; \bar{X})$

Le jeu de mesures 5 a été généré pour une bande passante du lien à 1000 kbps tandis que le jeu de mesures 6 se rapporte à une bande passante à 2000 kbps. Les temps de séjour ont été estimés en divisant le RTT par 2.

Les tableaux ci-dessous présentent les valeurs de ces jeux de mesures.

Débit mesuré (kbits/sec)	*Temps de séjour mesuré (msec)*
333.8	700
518.9	450
775.1	300
887.9	250
910.8	150
925.5	100
945.7	40

TAB. 27 – Jeu de mesures 5

Débit mesuré (kbits/sec)	*Temps de séjour mesuré (msec)*
341	700
529.9	450
787.3	300
1041.3	250
1523.6	150
1768	100
1826.6	40

TAB. 28 – Jeu de mesures 6