

Université de Béjaïa
Faculté des Sciences et des Sciences de l'Ingénieur

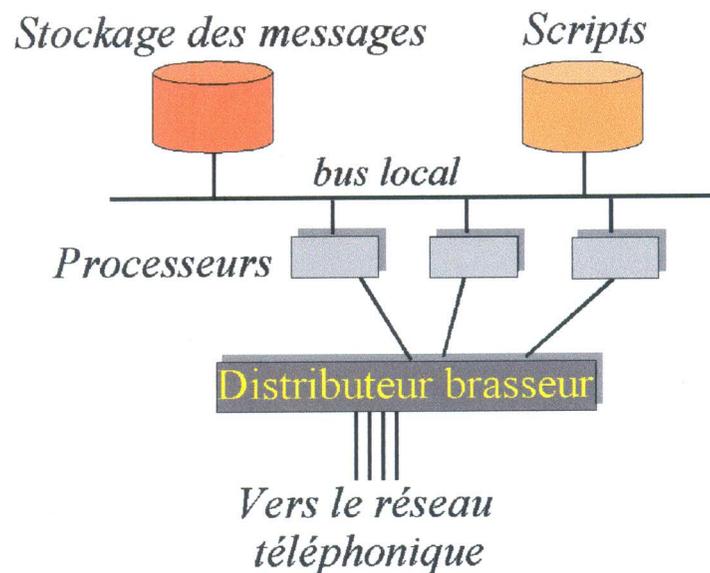
Département d'Informatique
Ecole Doctorale en Informatique

Djamil Aïssani

Amar Aïssani

La Théorie des Files d'Attente

*Fondements Historiques et Applications
à l'Evaluation des Performances*



Béjaïa – 2004

La Théorie des Files d'Attente

*Fondements Historiques et Applications
à l'Evaluation des Performances*

AVANT - PROPOS

Dans ce fascicule, nous présentons les éléments historiques liés à l'apparition, puis au développement de la Théorie des Files d'Attente. Ces fondements permettent de comprendre l'importance de cette théorie pour l'évaluation des performances des systèmes informatiques.

L'architecture d'un système informatique (ordinateurs et réseaux de communication) est composé de divers ensembles assurant des fonctions de transformation, stockage, diffusion et acquisition de l'information. La performance d'un tel système se définit comme l'adéquation entre la charge de travail imposée à un sous ensemble par les autres et sa capacité de traitement. Une mauvaise adéquation mettrait en évidence des modules qui agissent comme goulot d'étranglement de l'ensemble. Citons pour exemple : le bus, dans une architecture multiprocesseur à bus ; tout un système partagé, lorsque le temps de changement de contexte est trop lent ; un banc mémoire trop référencé dans une architecture à couplage fort (phénomène de « hot spot ») ; le maître dans toute architecture maître-esclave, le serveur dans une architecture client-serveur.

Il est établi que le problème de l'évaluation des performances d'un système doit être posé avant même sa conception effective (cf. Berkaoui et al. [5]). Plusieurs outils de modélisation, tous fondés sur le paradigme des systèmes à événements discrets coexistent aujourd'hui. La modélisation fondée sur la théorie des réseaux de files d'attente est certainement la plus utilisée. Jusqu'à ces dernières années, les deux termes *évaluation des performances* et *systèmes informatiques* étaient quasiment interchangeables [5].

Bien que la classe des réseaux à forme produit permette d'obtenir des résultats exacts et susceptibles de nombreuses généralisations, son pouvoir de modélisation reste limité. En effet, elle ne permet pas de prendre en considération des comportements répandus dans les systèmes informatiques de nouvelle génération (architecture multiprocesseur, systèmes répartis, calcul parallèle,...) tel que : la prise simultanée de plusieurs ressources par un même client, les files d'attente à capacité limitée et le blocage, la concurrence pour l'accès à des ressources partagées, les règles de routage incluant la synchronisation, la génération et le regroupement des clients. De nombreux travaux ont tenté de résoudre cet aspect dans les modèles de réseaux de files d'attente par des techniques approximatives [2], [5], [8]. Cependant aucune de ces techniques n'est universelle. Elles reposent généralement sur des

méthodes itératives dont il faut vérifier la convergence, ou sur des fortes contraintes d'application qui doivent être respectées pour assurer la validité de la solution.

D'autres approches de modélisation (Réseaux de Pétri Stochastiques, Réseaux de Neurone,...) sont bien adaptées pour décrire les phénomènes de concurrence, de conflit, de synchronisation prépondérants par les architectures des systèmes informatiques de nouvelle génération. Certaines d'entre elles ont été présentées dans le cours de Post-Graduation «*Modélisation des Systèmes Informatiques*» (Institut d'Informatique, U.S.T.H.B. Alger – 1988) [3], [4].

Le cours «*Evaluation des Performances des Systèmes Informatiques*» assuré cette année pour l'Ecole Doctorale d'Informatique de Béjaïa sera construit sur le même plan. Il prendra en compte les nouvelles perspectives dégagées lors du Colloque International MOAD'92 (*Méthodes et Outils d'Aide à la Décision*, Béjaïa – Novembre 1992) [5], [6] et dans la mise en œuvre de l'Accord – Programme Algéro – Français «*Evaluation des Performances des Systèmes Complexes : Applications aux Réseaux de Télécommunication*» [7]. Enfin, les auteurs tiennent à remercier Mademoiselle Karima Lagha pour avoir corrigé le manuscrit et formulé des suggestions constructives.

REFERENCES

- [1] Aïssani D. et Aïssani A., *Fiabilité des Systèmes et Systèmes de Files d'Attente non Fiabiles*, U.E.R. Mathématiques – Informatiques, ENITA, Bordj – el – Bahri, 1986, 90 pages.
- [2] Aïssani D., *Modélisation et Simulation des Systèmes Industriels*, Cours de Post-Graduation, Institut d'Informatique, Université d'Annaba, 1987.
- [3] Aïssani D., *Modélisation et Evaluation des Performances des Systèmes Informatiques*, Cours de Post-Graduation, Institut d'Informatique, U.S.T.H.B. Alger, 1988.
- [4] Iouallalen Malika, *Conception et Implémentation d'un Langage de Description, de Transformation et de Manipulation de Réseaux de Pétri : DEMA -RP*, Thèse de Magister en Informatique, Institut d'Informatique, U.S.T.H.B. Alger, Décembre 1989.
- [5] Actes du Colloque International MOAD'92 (*Méthodes et Outils d'Aide à la Décision, Béjaïa – Novembre 1992*), Béjaïa, Novembre 1992, 2 Volumes, 737 pages.
- [6] Aïssani D., *Méthodes et Outils d'Aide à la Décision*, Revue **Matapli** (Société Française de Mathématiques Appliquées et Industriels), Paris, 1993, pp. 71 – 72.
- [7] Aïssani D., Maurras J.F. et al., *Evaluation des Performances des Systèmes Complexes : Applications aux Réseaux de Télécommunication*. Accord – Programme Algéro – Français entre les laboratoires LAMOS Béjaïa et Laboratoire d'Informatique de Marseille (2000 – 2002).
- [8] Djellab Natalia, *Systèmes de Files d'Attente avec Rappels : Méthode d'Approximation pour un Système M/G/1 avec Rappels et Pannes*. Thèse de Doctorat d'Etat en Informatique, Faculté des Sciences, Université de Annaba, Décembre 2003.

Table des matières

1	Introduction	4
2	Fondements historiques	4
3	Description des modèles	7
4	Système de notations	11
5	Analyse des performances du service	12
6	Optimisation des performances du service ou contrôle du système	17
6.1	Choix du nombre de serveurs	17
6.2	Contrôle du taux de service	18
6.3	Stratégie de connexion des serveurs	18
6.4	Affectation des priorités ou ordonnancement	19
6.5	Problème d'affectation des réparateurs à la maintenance	19
6.6	Problèmes de routage dans les réseaux	20
6.7	Reconnaissance optimale de formes	20
7	Identification du modèle	21
7.1	Test du Khi-deux	22
7.2	Test de Kolmogorov-Smirnov	25
7.3	Test de Cramer-Von Mises	25
7.4	Autres tests	26
8	Conclusion et Perspectives	26
9	Bibliographie	27
10	Annexe : Sommaire de l'ouvrage MMAPA	39

1 Introduction

La théorie des files d'attente (Queueing Theory, ou Théorie des services des Masses comme on a prit l'habitude de l'appeler en Russie et dans les Pays de l'ex bloc de l'est), est née d'exigences pratiques liées à la nécessité de fournir aux décideurs des méthodes mathématiques d'aide à l'organisation rationnelle du service massif de clients, ou de phénomènes apparentés, et qui conduisent à la formation de files d'attente.

La notion de "files d'attente" peut-être évidemment comprise dans le sens du langage habituel, mais dans le contexte de ce fascicule, nous lui conférons un sens plus abstrait qui permet d'englober dans un même contexte un large champ d'applications pratiques. La terminologie adoptée s'inspire habituellement du schéma de service des clients dans un magasin ou des malades chez le médecin. Ces exemples de "files d'attente" de notre vie quotidienne peuvent être ainsi conceptualisés et offrent donc un vaste champ d'investigation dans divers secteurs stratégiques de l'industrie et de l'économie.

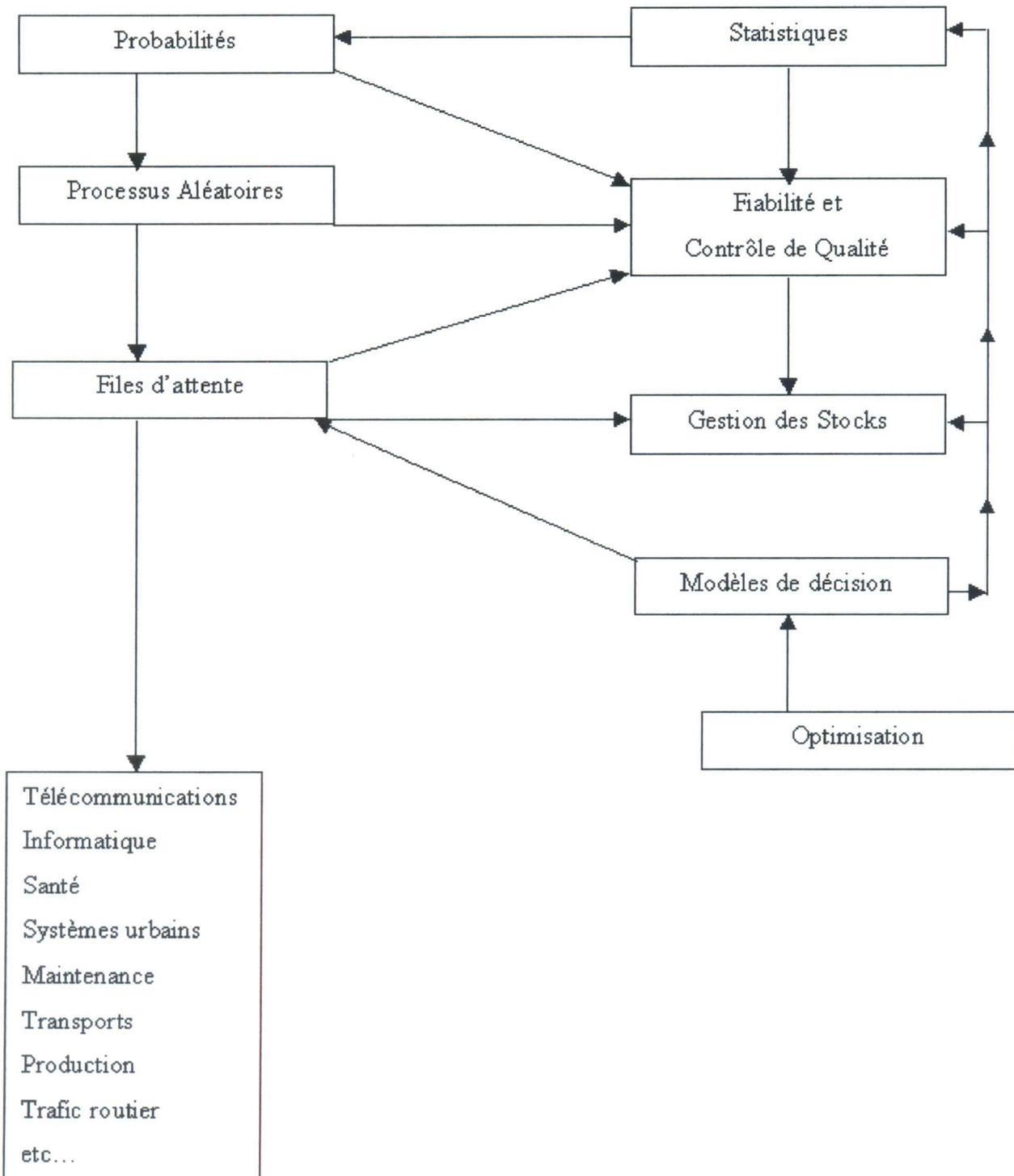
La théorie des files d'attente a pour objet l'analyse de l'évolution des "files d'attente", l'élaboration et l'optimisation des indicateurs de mesure des performances du service. Le caractère massif des demandes de service, ainsi que la diversité des facteurs d'influence externe conduisent naturellement à une formulation dans le langage de la théorie des probabilités.

2 Fondements historiques

La première publication sur le sujet est vraisemblablement l'article de Joannsen (1907) "*Waiting times and number of calls*", publié en 1910 dans "*Post Office Electrical Engineers Journal*". Cependant, les méthodes utilisées dans ce papier étaient dépourvues de rigueur mathématique. Il est couramment admis aujourd'hui que le fondateur de la théorie des files d'attente est le Danois Erlang A.K. (1878-1929), qui a collaboré de 1908 à 1922 avec la compagnie danoise de téléphone pour mener des études en télétrafic (débit de transmission des réseaux de communication téléphonique). Son article de 1909 ¹ est considéré comme la première contribution mathématique sérieuse à la théorie des files d'attente. L'influence de la mécanique statistique de l'époque l'a vraisemblablement conduit à introduire la notion d'équilibre statistique qui lui a permis d'obtenir la formule pour la probabilité de refus qui porte aujourd'hui son nom. La vie et l'oeuvre de A.K. Erlang ont fait l'objet de l'article de Brockmeyer E. & al. ². Ses travaux ont constitué le point de départ pour d'autres recherches s'articulant autour de la confirmation ou de l'infirmité de ses résultats en télétrafic, et du point de vue théorique, sur la généralisation de sa fameuse formule. La

¹Erlang A.K., *The theory of probabilities and telephone conversation*, Nyt-tidsskrift for matematik, B20, 33, (1909)[68].

²Brockmeyer E., Halstrom H.L., Jensen A., *The life and works of Erlang A.K.*, Transactions of the Danish Academy Technical Sciences, N°2, Copenhagen, 1948 [37].



Légende : Situer le théorie des files d'attente dans le contexte des théories mathématiques ou branches annexes n'est pas chose aisée. Le diagramme ci-dessus nous semble assez représentatif.

première généralisation a été effectuée par Engset T. (1918) dans le cas d'une source finie³. D'autres modèles ont été étudiés à l'époque par O'Dell G. et Molina E. En 1928, Fry T.C. [76] publie un ouvrage sur la théorie des probabilités comprenant un chapitre sur les problèmes de télétrafic. Parmi les précurseurs de cette période, on citera Crommelin C.D. (1932), (1934), Wilkinson R.I. (1931), (1933), Vaultot E. (1924) et Khintchine A.Y. (1932).

L'introduction dans les années 30 des processus de naissance et de mort marque un tournant nouveau dans la description des phénomènes d'attente. Kolmogorov A.N. lui-même s'est intéressé à cette question en 1931. En 1933, il publie son ouvrage devenu classique où il propose son axiomatique de la théorie des probabilités qui est adoptée aujourd'hui de manière quasi-unanime⁴. L'idée de Erlang A.K. sur l'équilibre statistique est déjà assimilée à la notion de mesure stationnaire de processus markovien. Cependant, Pollaczek F. remarque (en 1934) que l'équilibre statistique ne permet pas toujours de décrire la micro-dynamique de tous les processus aléatoires en Théorie des Files d'Attente. A cette époque, Pollaczek F. et Khintchine A. s'intéressent au cas de durées de service arbitraires⁵. Les formules obtenues par Pollaczek F. vers 1950 sont déduites de manière plus simple par Kendall D.G. (1951) par la méthode de la chaîne de Markov immergée⁶.

Les travaux du suédois Palm C. (1936), (1938) et surtout (1943) ont été à la base de nombreuses études sur la variabilité des charges téléphoniques. Le flux de Palm, appelé également processus de renouvellement, a trouvé des applications intéressantes en théorie de fiabilité et de maintenance. Gnedenko B.V. a recueilli une série d'articles de Khintchine A.⁷ proposant diverses extensions, parmi lesquelles les formules connues aujourd'hui sous le nom de formules de Pollaczek-Khintchine.

L'introduction de nouvelles classes de processus aléatoires a permis d'élaborer de nouvelles approches intéressantes : les processus semi-markoviens [Smith W. (1953), (1958)], [Cox D.R. (1955)], les processus markoviens à trajectoires linéaires par morceaux [Beliaev Y. (1962)]⁸. Les processus de marquage [Franken P. & al. (1981)][75] constituent un cadre de formalisation plus élaboré. Les processus de diffusion conduisent à des méthodes approximatives très prisées par les praticiens [53, 167, 62].

Les ouvrages et monographies publiés dans les années soixante tentent de présenter

³La formule d'Erlang a été obtenue dans le cas d'une source infinie. Aussi paradoxal que cela puisse paraître, le cas d'une source finie est plus complexe et constitue aujourd'hui le centre d'intérêt des recherches.

⁴A l'exception de l'école des subjectivistes.

⁵La formule d'Erlang est obtenue dans le cas d'une distribution exponentielle.

⁶On utilise également dans la littérature Française le terme "incluse" ou "induite" (Imbedded Markov Chain).

⁷Après la mort de l'auteur, et dont il existe actuellement plusieurs traductions.

⁸Cette approche est similaire à celle utilisée par Takacs L. (1955) pour obtenir l'équation intégral-différentielle de la loi de la durée d'attente, et qui porte son nom.

une théorie unifiée : Riordan J. (1962)[176], Le Gall P. (1962)[138], Pollaczek F. (1957), Syski R. (1960)[205], Takacs L. (1962)[210], Cox D.R. et Smith W.L (1961)[55] et Benès V.E. (1963)[22]. La monographie de Saaty T.L. (1961)[184] comporte une bibliographie riche d'environ 1280 titres.

Jackson J.R (1957), (1963)[109] démontre les premières formules de décomposition pour les réseaux de files d'attente, et qui sont basées sur un résultat de Burke P.J. (1956) sur le flux des départs. Les méthodes de décomposition et leurs variantes sont actuellement très utilisées pour l'évaluation des performances des systèmes et ont conduit à l'élaboration d'une multitude de modèles qui offrent un large champ d'investigation, aussi bien pratique que théorique [127, 80, 78]. Le second volume de l'excellent ouvrage de Kleinrock L. (1976)[125] est recommandé pour comprendre l'apport de la théorie des files d'attente dans ce domaine. Dans cet ouvrage, Kleinrock L. élucide les diverses questions qui se sont posées lors de la conception des plus grands réseaux de l'époque (SITA, ARPANET, ...), ancêtres de l'actuel INTERNET.

Du point de vue de la formalisation mathématique, il est intéressant de citer la formule de conservation de Little J. (1961), qui a fourni matière à une longue polémique [Stidham (1974)] et [Whitt W. (1991)]. L'analyse opérationnelle, bien que souvent controversée, constitue une approche complémentaire intéressante à la théorie classique des files d'attente [122], [119]. Elle contribue à prouver la validité des résultats classiques (tels que les lois de conservation) pour des systèmes complexes lorsque les hypothèses conventionnelles ne peuvent être justifiées. Elle permet en outre d'expliquer la robustesse de ce type de résultats.

La méthode des événements fictifs, dont l'idée remonte à Van Dantzig D. (1948), constitue une autre approche non conventionnelle, appelée également "méthode des marques collectives" ou "méthode des catastrophes" [125, 126]. Elle permet de donner une interprétation probabiliste aux résultats complexes, souvent obtenus en termes de transformées de Laplace ou de fonctions génératrices. Cette méthode a été utilisée avec succès pour l'étude des systèmes avec priorités [38] et plus récemment, des systèmes avec vacances [Cong (1994)].

Les travaux de Lindley L., Kingman J. et Marshall K.T. sont à la base de nombreuses méthodes mathématiques plus complexes axées sur les questions d'approximation et de stabilité. On pourra se référer à ce sujet aux ouvrages de Borovkov A. (1976)[31], Kœnig & Stoyan (1976), (1988). Le lecteur trouvera un état de l'art actuel sur la question dans un mémoire élaboré par Gnedenko B. & Kœnig (1983), édité en langue allemande, et dont nous ne connaissons pas de traduction. L'élaboration de l'approche des opérateurs de la théorie de stabilité (connue également sous le nom de méthode de stabilité forte) permet d'obtenir les estimations qualitatives de stabilité, avec un calcul exact des constantes (cf. [Aïssani (1983)] [10] et [Kartashov (1996)] [116]).

L'émergence durant la seconde guerre mondiale de cette nouvelle science qu'est la recherche opérationnelle a contribué à donner à la théorie des files d'attente des développements spécifiques. Le collectif des 57 auteurs du mémoire (1978) *Handbook of Operations Research*[150], classe cette théorie parmi les modèles stochastiques de la recherche opérationnelle. La bibliographie de Bhat Y. sur le sujet fournit de nombreuses références sur le domaine du point de vue pratique. Pour les applications, nous conseillons les ouvrages de Lee A.M (1966)[139], Bhat U.N. (1972)[25], Gelenbe E. (1980)[78], Morse P.M (1958)[152], Newell G.F. (1971)[160], Prabhu N.U. (1965)[172], Hayes J. (1984)[97], Allen A. (1973) et Karakoviak (1987)[114].

Les ouvrages de recherche opérationnelle ou de théorie de télétrafic, comportent généralement un chapitre consacré à la théorie des files d'attente : Taha Hamdy A. (1982)[208], Howard (1961), Hillier & Lieberman (1967)[101], Schwartz (1981) et Sevcik K.C. (1983).

Malgré la complexité des modèles auxquels elle conduit, la théorie des files d'attente continue d'être un centre d'intérêt pour de nombreux chercheurs de par la diversité des applications qu'elle permet de traiter. La revue "Queueing systems" (QUESTA), éditée périodiquement depuis 1986, publie des contributions d'intérêt pratique et théorique sur le sujet. Parmi les volumes consacrés aux applications, citons entre autres : Systèmes de production (Vol. 12, 1992), Systèmes de communication (Vol. 9, 1991).

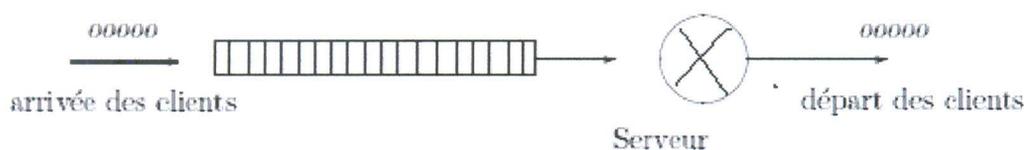


FIG. 1 - Système d'attente ouvert à un serveur.

3 Description des modèles

Dans le cas le plus simple, le flux des arrivées est décrit par la suite des dates successives de leurs occurrences dans le système

$$t_0(= 0) < t_1 < t_2 < \dots < t_n < t_{n+1} < \dots,$$

où t_n est la date d'arrivée du $n^{\text{ème}}$ client. Parfois, il est nécessaire de préciser cette information par d'autres indicateurs (lieu d'occurrence, catégorie d'importance ou priorité, ...).

Le système est dit *ouvert* si la source des clients est illimitée. Il est *fermé* si la source est finie (par exemple, dans le cas de service de n machines sujettes à des pannes).

La suite $\{t_n\}$ peut être définie en donnant la suite $\{\xi_n\}$, où $\xi_n = t_n - t_{n-1}$ représente l'intervalle entre les $(n-1)^{\text{ème}}$ et $n^{\text{ème}}$ arrivée.

Pour définir le flux, il suffit de donner toutes les fonctions de répartition finies des vecteurs aléatoires $(\xi_1, \xi_2, \dots, \xi_n)$ pour tout $n \geq 1, P\{\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_n < x_n\}$, $x_i \in \mathbf{R}, i = 1, \dots, n$.

Si les variables aléatoires $\{\xi_n\}$ sont indépendantes dans leur ensemble, le flux est à *post-action limitée*, et dans ce cas, il suffit de donner la famille de fonctions

$$A_n(x) = P\{\xi_n < x\}, n \geq 1.$$

Un flux à post-action limitée tel que $A_2(t) = A_3(t) = \dots = A(t), A_1(t) \neq A(t)$ est appelé *flux de Palm* ou *récurrent*, ou encore *processus de renouvellement attardé*⁹.

Si un flux de Palm simple est tel que $A_1(t) = A(t)$, on l'appelle flux récurrent simple ou *processus ordinaire de renouvellement*. Un flux de Palm simple vérifiant $A(t) = 1 - e^{-\lambda t}, \lambda > 0, t \geq 0$ est appelé *flux ordinaire de Poisson*, de paramètre λ .

Le flux de Poisson possède des propriétés mathématiques intéressantes ; il est ordinaire, stationnaire et sans mémoire. De plus, le processus de comptage $N(t) = \max\{n : t_n < t\}$ représentant le nombre d'arrivées dans la période $(0, t)$ obéit à une loi de Poisson, d'espérance mathématique λt ,

$$P\{N(t) = k\} = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

Le flux des arrivées est homogène si les clients possèdent des propriétés statistiques identiques. Lorsque le flux n'est pas homogène, les clients peuvent appartenir à des classes distinctes. Dans ce cas, la superposition des flux des différentes classes est appelé flux-somme. Le flux est ordinaire si chaque date t_n correspond à l'arrivée d'un seul client. Dans certains cas, les clients peuvent arriver par groupe de taille constante ou aléatoire.

Les modèles de files d'attente peuvent être également classifiés selon l'organisation du service. Le service est l'exécution d'un certain nombre d'opérations ou de tâches qui nécessitent une durée (ou une quantité de travail) pour satisfaire les demandes des clients. La situation la plus simple est celle où on ne s'intéresse qu'à la durée de service. Soit τ_n la durée de service du $n^{\text{ème}}$ client. Si $\{\tau_n\}$ est une suite de variables aléatoires indépendantes et identiquement distribuées (ce qui est généralement le cas), alors il suffit de donner la fonction de répartition :

$$H(x) = P\{\tau_n < x\}, \text{ pour tout } n \geq 1.$$

⁹L'origine de cette appellation provient de la théorie de fiabilité, la grandeur t_n représentant la date de la $n^{\text{ème}}$ panne d'un élément ; à cet instant même, on procède au renouvellement de cet élément et la durée de renouvellement est négligeable. Lorsque la durée de renouvellement n'est pas négligeable, t_n est un processus alterné de renouvellement [5, 16].

On utilisera parfois l'interprétation "énergétique" du service. La variable τ_n représente alors la quantité de travail nécessaire au serveur pour exécuter le service du $n^{\text{ème}}$ client. On doit préciser dans ce cas la vitesse α d'exécution de ce travail. Pour fixer les idées, supposons qu'un client arrive dans le système à la date t_n , et que son service nécessite une quantité de travail égale à x unités. Si à la date t_n le serveur est libre, alors le service débute à l'instant t_n même, et s'achève à l'instant $t_n + x/\alpha$.

En télécommunications, on représente la durée de service (qui est la durée de transmission du message) sous la forme $\tau_n = v_n/C$, où v_n est la taille du message (en unités d'information : bits, octets, ...), et C la capacité de transmission du canal de communication.

Le service peut-être assuré par plusieurs serveurs en parallèles. A chacun d'eux, on associera alors la caractéristique correspondante $H(x)$, qui peut-être différente pour chaque serveur. Le service peut-être également exécuté en série dans des "stations" comportant un ou plusieurs serveurs en parallèles.

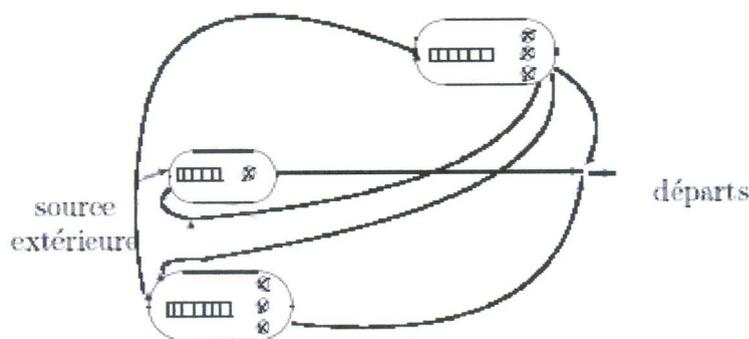


FIG. 2 - Réseau de files d'attente avec probabilités de transition.

On dit alors que le système est un réseau de files d'attente [QUESTA 1993, Vol. 13]. L'itinéraire de chaque client dans le réseau peut-être déterministe (par exemple une chaîne de fabrication), ou aléatoire (un grand réseau de transmission de données). Le réseau est ouvert si les clients proviennent d'une source externe; il est fermé s'il y a un nombre constant de clients qui circulent dans le réseau (par exemple, la réparation d'un groupe de machines par r brigades spécialisées : électriciens, mécaniciens, ...).

La majorité des modèles de files d'attente sont étudiés pour des lois de service exponentielle ou d'Erlang.

Les conflits qui se produisent lorsque plusieurs clients sollicitent un même serveur peuvent être résolus de diverses manières. Dans les systèmes avec refus, le client qui trouve les serveurs occupés quitte le système pour de bon ou momentanément. Dans le dernier

cas, on parle de système avec rappels [QUESTA - 1990]. Dans de nombreux modèles, les clients forment une file d'attente qui est contrôlée par une discipline de service donnée,

FIFO (first in-first out), premier arrivé - premier servi : C'est la discipline la plus courante.

LIFO (last in-first out), dernier arrivé - premier servi : C'est le cas lorsque la file est matérialisée par une mémoire d'ordinateur ou une pile de livres.

RANDOM, le client est choisi au hasard dans la file selon une loi de probabilité donnée.

PROCESSOR-SHARING, le client reçoit à chaque passage des quantums de temps. Si à l'issue d'un quantum son service ne s'est pas achevé, il rejoint la file d'attente (Modèle du Tourniquet ou Round Robin).

On peut imaginer des situations intermédiaires où le client est refusé lorsque la file d'attente est pleine (systèmes à capacité limitée). Le client peut-être impatient, auquel cas il quitte le système après une certaine durée d'attente. Dans certains systèmes, on peut instaurer des priorités entre classes de clients par catégorie d'importance : systèmes avec priorités ¹⁰ (Jaiswal [111]). La priorité est absolue si le client prioritaire peut interrompre le service du client en cours ; elle est relative dans le cas contraire. Les problèmes de priorité et leurs applications aux systèmes informatiques font l'objet de la monographie de Bronstein & Dukhovny (1976)[38], ainsi que QUESTA-1989.

Dans les systèmes manufacturiers et dans les systèmes informatiques actuels à architectures parallèles, on introduit des procédures de synchronisation décrites par des sémaphores : dédoublement (Fork) et "regroupement" (Join). Chaque client accédant au système génère un certain nombre de fils qui sont dirigés vers l'entrée du système. Les différents frères se synchronisent à la sortie de manière à quitter le système simultanément. Cela correspond aux actions de désassemblage / assemblage dans les systèmes manufacturiers [Fdida S. (1986) Dallery Y. (1991)].

Très souvent, il est nécessaire de prendre en considération les possibilités de pannes des serveurs (systèmes non fiables). La fiabilité d'un serveur est par définition sa probabilité de survie pendant une période $(0, t)$ *i.e.*, $R(t) = P(X > t)$, où X est la durée de vie (aléatoire) du serveur. Il est nécessaire alors d'introduire la loi de la réparation [86]. La motivation essentielle est l'étude de l'influence de la fiabilité sur les principales mesures de performance du système étudié.

Ces dernières années, l'attention de certaines communautés de chercheurs s'est focalisée sur l'étude des modèles d'attente avec répétition d'appels (Retrial Queues) qui occupent une situation intermédiaire entre les modèles FIFO et les modèles avec refus [72]. Ce

¹⁰Par exemple, dans un réseau téléinformatique, un télégramme est prioritaire par rapport à d'autres messages moins urgents.

type de situation est particulier aux systèmes téléphoniques, de réservation et protocoles CSMAKD. Les éléments décrits ci dessus permettent, comme nous l'avons vu, d'aboutir à une riche variété de modèles, en combinant les différentes descriptions. Ces situations ne sont pas exhaustives : on peut introduire l'usure du service¹¹, l'autonomie du serveur qui s'autorise des "promenades" ou vacances [QUESTA-1986].

Le flux des départs est un élément particulier qui dépend des précédents et qui peut être décrit à l'aide d'une suite $\{t_n'\}$ analogue à celle du flux des arrivées. Son importance est due au fait que dans les réseaux par exemple, le flux des départs d'une station joue le rôle d'un flux d'arrivées dans une ou plusieurs stations. Les flux de départs sont rarement à post-action limitée et cela complique l'étude des réseaux.

4 Système de notations

Le système de notations le plus répandu actuellement a été introduit par [Kendall D.G. (1953)], et complété par [Lee A.M. (1966)][139] :

$$a/b/m/K/M$$

a désigne le type de la loi des arrivées, *i.e.* de la fonction $A(t)$.

b celui de la loi de service, *i.e.* de la fonction $H(t)$.

Pour chaque modèle concret, on remplacera a et b par l'un des symboles standards suivants :

M (Markovian) : loi exponentielle, *i.e.* $A(t)$ (ou $H(t)$) est de la forme $1 - e^{-\theta x}, x \geq 0$;

D : loi déterministe ;

E_r : loi d'Erlang d'ordre r ;

GI : loi arbitraire, mais les variables constituant la suite $\{\xi_n\}$ (ou $\{\tau_n\}$)

sont indépendantes ;

G : loi arbitraire ;

H_k : loi hyper-exponentielle d'ordre k ;

PH : phase-type ;

IFR, NBU : distribution non paramétriques d'âge ou de vieillissement ;

SM : semi-markovien ;

Par ailleurs,

m : désigne le nombre de serveurs ;

K : la capacité du système (service + file d'attente) ;

M : nombre de sources de clients.

Par exemple, $M/E_r/3/10/20$ désigne le modèle ayant les éléments suivants :

Flux des arrivées Poissonien, *i.e.* $A(t) = 1 - e^{-\theta t}$.

¹¹Par exemple, si un message met un temps trop long pour être transmis, l'information qu'il comporte peut devenir désuète lorsque le serveur est humain. En effet, il peut être sujet à la fatigue, et la qualité du service s'en ressent.

Durée de service de loi d'Erlang d'ordre r .
 3 serveurs.
 10 places dans le système : 3 au service, 7 dans la file.
 La source peut émettre au plus 20 demandes.

Lorsqu'on ne le mentionne pas, la discipline de service est *FIFO*. Toute autre discipline peut-être mentionnée, soit en ajoutant un symbole supplémentaire $a/b/m/K/M(\text{FIFO})$, ou alors en précisant la nature de la discipline dans une description complémentaire. Pour les systèmes avec rappels [72], on peut ajouter la capacité de l'orbite au lieu de la file d'attente.

Lorsque $m = \infty$, on peut supprimer le symbole ∞ : $M/G/1$ désigne $M/G/1/\infty/\infty/\infty$. Il faut faire attention aux confusions lorsqu'il reste des symboles intermédiaires finis.

Les systèmes avec priorités sont désignés par des flèches et un indice indiquant le nombre de priorités $\vec{M}_N/\vec{M}_N/1$ (dans les notations de [Basharin G.P (1968)][18]). La notation d'un système en série comporte autant de quintuplés qu'il y a de stations dans le réseau $M/M/1 \rightarrow G/M/1 \rightarrow E_r/M/2$.

Dans les notations de [Borovkov A.A. (1972)][31], on indique en plus la loi de la taille du groupe d'arrivée, et celle des paquets servis e et f , respectivement. Le modèle est noté $\langle a, e, b, f \rangle$. Par exemple, $\langle G, G, M, G \rangle$. Lorsqu'il ne peut y avoir qu'une arrivée, et un service à la fois, on notera $\langle G, 1, M, 1 \rangle$. En indice, on peut ajouter le symbole A pour service autonome $\langle G, G, M, G \rangle_A$, ou R pour les systèmes avec refus.

Dans ce fascicule, nous utilisons les notations classiques de Kendall-Lee.

5 Analyse des performances du service

La quantité de service peut-être estimée à l'aide d'un ensemble fini de caractéristiques de performance $V = \{x_1, x_2, \dots, x_k\}$.

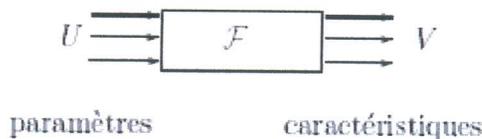


FIG. 3 - Modélisation d'un système d'attente

Ces grandeurs dépendent du modèle \mathcal{F} considéré, qui lui, est décrit à l'aide de la suite (finie) $U = \{u_1, u_2, \dots, u_l\}$ des éléments figurant par exemple, dans les notations de Kendall-Lee, et qu'on appelle "paramètres du système". Pour le système classique avec

attente, on peut s'intéresser à la durée moyenne d'attente ¹²

La caractéristique de base de la majorité des systèmes est le nombre de clients dans le système à l'instant t , noté $\{X(t), t \geq 0\}$. C'est un processus aléatoire défini sur un espace des états (fini ou dénombrable) $\Omega_n = \{0, 1, 2, \dots, N\}$, ($N \leq \infty$).

Ce processus peut être représenté sous la forme :

$$X(t) = S(t) + Q(t), t \geq 0,$$

où $S(t)$ est le nombre de clients en cours de service à l'instant t , et $Q(t)$ le nombre de clients dans la file à cet instant. Il faut noter que dans la littérature, les processus $\{X(t)\}$ et $\{Q(t)\}$ peuvent être désignés sous le même vocable : "file d'attente". Cette confusion ne présente pas d'inconvénients majeurs.

Le but de l'analyse consiste principalement à étudier les distributions de probabilités de ces processus aléatoires. En particulier, on s'intéresse à la distribution :

$$p_k(t) = P\{X(t) = k\}, k \in \Omega_N \quad (N \leq \infty).$$

L'analyse d'un tel processus est facilitée lorsqu'il possède la propriété de Markov, qui exprime l'indépendance du futur et du passé [125, 73]; par exemple pour les systèmes de type $M/M/m$, l'analyse se complique dès qu'on perturbe les hypothèses conduisant à cette propriété vitale.

On s'intéressera souvent au régime stationnaire (ou régime permanent) lorsque le système a fonctionné suffisamment longtemps. Cela revient à l'étude du processus $\{X(t)\}$ lorsque $t \rightarrow \infty$. Sous certaines conditions, les caractéristiques de la distribution $\{p_k(t)\}$ tendent à se stabiliser indépendamment de t , et de l'état initial du processus $X(t_0)$.

Les conditions d'existence du régime stationnaire résultent des théorèmes ergodiques pour les processus aléatoires. Le principe d'ergodicité de la physique théorique se formule sous une forme plus générale en théorie des processus aléatoires, et exprime justement cette invariance des distributions de leurs probabilités. Le régime stationnaire était appelé par Erlang, "équilibre statistique". En effet, il avait remarqué que les caractéristiques étudiées présentent un intérêt justement en régime stationnaire qui se manifeste, par exemple en téléphonie, aux heures "pics".

Notons $\sigma_k(t) = 1$, si $X(t) = k$, et $\sigma_k(t) = 0$, dans le cas contraire. Alors, la variable :

$$Y_T(k) = \frac{1}{T} \int_0^T \sigma_k(t) dt \quad (1.1)$$

¹²qui n'a pas de sens pour le modèle avec refus. On utilise pour ce dernier, la probabilité de refus qui est une caractéristique très importante en téléphonie.

représente la durée moyenne de séjour du processus $X(t)$ à l'état k dans l'intervalle $(0, T)$. Si le processus est ergodique, alors les limites

$$p_k = \lim_{T \rightarrow \infty} Y_T(k), \quad k \in \Omega_N \quad (1.2)$$

existent, et $\{p_k\}$ est la distribution stationnaire du processus $\{X(t)\}$, indépendante de t et de l'état initial.

Pour le système $M/M/1$, le régime stationnaire existe si le taux des arrivées λ ne dépasse pas le taux de service μ , ($\lambda < \mu$). C'est une condition "naturelle" de stabilité du système. En fait, la quantité $\rho = \frac{\lambda}{\mu} = 1 - p_0$ est la probabilité pour que le serveur soit actif à un instant arbitraire en régime stationnaire. En vertu de l'égalité (1.2), cette caractéristique peut-être estimée par $\hat{\rho} = T_0/T$ où T_0 est la durée de séjour cumulée à l'état actif pendant la période (suffisamment grande) d'observation $(0, T)$. On l'appelle charge du système ou taux d'utilisation du serveur. Il s'avère que la condition de stabilité $\rho < 1$ reste vraie dans le cadre plus général des modèles $G/G/1$. Il existe cependant des situations où même dans le cadre markovien la condition "naturelle" n'est pas forcément nécessaire.

Les caractéristiques stationnaires peuvent être généralisées à des fonctionnelles du processus $X(t)$. Ainsi :

$$\hat{f} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T f(X(t)) dt$$

pour des fonctionnelles f bornées réelles ou complexes.

La connaissance de la distribution stationnaire $\{p_k\}$ permet l'obtention d'une variété de caractéristiques de performance (stationnaires) non moins importantes. Par exemple, pour le système à m serveurs avec attente,

- Le nombre moyen de clients dans le système ($N \leq \infty$) :

$$\bar{N} = \sum_{k=0}^N k p_k.$$

- Le nombre moyen de serveurs occupés :

$$\bar{N}_1 = \sum_{k=1}^m k p_k \quad (m < N).$$

En télécommunications, cette caractéristique s'appelle également trafic moyen écoulé. Il existe diverses définitions de la notion de trafic selon la forme de sa mise en œuvre. Le trafic offert est le nombre de demandes arrivant pendant un service; le trafic avec refus est le nombre de demandes refusées pendant un service; le trafic d'attente est le nombre de demandes en attente à un instant donné. Les trafics moyens se définissent à partir des trafics instantanés correspondants. Une discussion des autres formes de trafic peut-être trouvée dans le livre de Doyon (1989)[67].

- Le nombre moyen de serveurs libres :

$$\bar{N}_0 = m - \bar{N}_1.$$

- Le coefficient d'inactivité :

$$K_0 = \bar{N}_0/m.$$

- Le coefficient d'activité :

$$K_1 = \bar{N}_1/m.$$

- La taille moyenne de la file :

$$\bar{Q} = \sum_{k=m}^N (k-m)p_k \quad (N \leq \infty).$$

- La probabilité pour qu'il y ait plus de l clients en attente :

$$P_{>l} = \sum_{k=l+1}^N p_k.$$

Pour les systèmes avec refus, on s'intéressera à la probabilité de refus P (P est la probabilité du refus, de rejet ou de perte). Par exemple, pour le système $M/M/m/m$, $P = p_m$ la probabilité pour que tous les serveurs soient occupés, et par conséquent, le client arrivant dans le système est refusé. Cette caractéristique est très utilisée en téléphonie et en théorie de trafic ; elle est donnée par les formules d'Erlang et d'Engset. Ses valeurs sont tabulées pour un usage plus facile [Palm 1947]. Les formules permettaient de résoudre le problème de dimensionnement de réseaux téléphoniques évoqué au paragraphe 2 [6].

Si R_k désigne le nombre de clients perdus parmi les k premières demandes, alors R_k/k est la fréquence de refus, et en vertu de la loi des grands nombres, $R_k/k \xrightarrow{\text{loi}} P$ sur une période d'observation suffisamment grande. Dans le cas général, et contrairement à l'exemple ci-dessus, la probabilité d'encombrement p_m (que tous les serveurs soient occupés) diffère de la probabilité de refus [67].

Si C_n désigne le $n^{\text{ème}}$ client arrivé dans le système, on peut lui associer une variable aléatoire $W_n \geq 0$ égale à sa durée d'attente dans la file avant le début de service. La durée virtuelle d'attente (pour la discipline *FIFO*),

$$w(t) = \begin{cases} 0, & \text{si } W_n \leq t - t_n \\ W_n - (t - t_n), & \text{si } W_n \geq t - t_n. \end{cases}$$

Pour $t_n < t < t_{n+1}$, $w(t)$ représente la durée d'attente d'un client qui serait arrivé dans le système à l'instant t . Il est clair que $W_n = w(t_n)$, où t_n est la date d'arrivée de C_n .

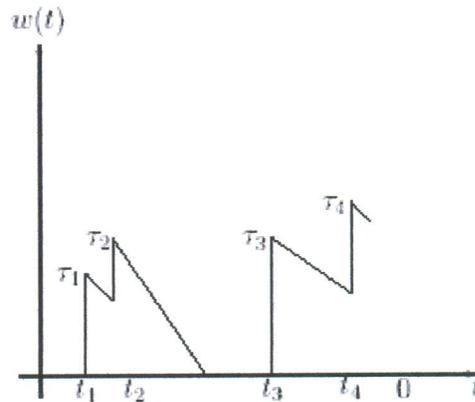


FIG. 4 - Evolution de la durée d'attente dans le temps.

Les spécificités des systèmes concrets considérés conduisent à l'introduction de paramètres supplémentaires. La fonction principale d'un réseau de transmission de données consiste, par exemple, à transmettre le message d'un point du réseau à un autre¹³. Cependant, le message peut "vieillir" durant la période consacrée à cet effet : au sens où sa valeur (exprimée en termes d'efficacité, de rentabilité économique, ...) décroît au cours du temps. L'information apportée par le message peut devenir inutile passé un certain délai. On peut même envisager des situations où cette information serait néfaste. On peut introduire la fonction $\varphi(t)$ exprimant la valeur de l'information en fonction du temps [Zakharov G.P. (1982)][226], ($\varphi(t) = \varphi(t_A, t_L, t_{AD}, t_{cr})$, où t_A est la durée aléatoire d'usure du message; t_{AD} la valeur admissible, t_{cr} la valeur critique, et t_L la valeur limite).

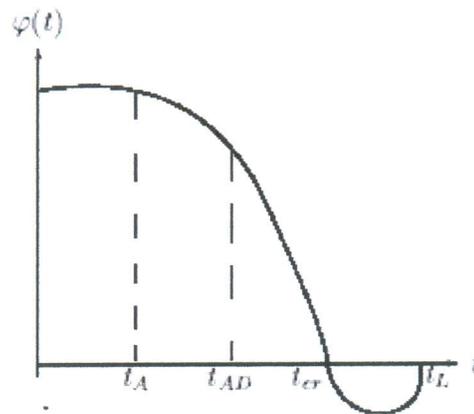


FIG. 5 - Valeur de l'information en fonction du temps .

La valeur de l'information décroît lentement dans l'intervalle $(0, t_A)$, et plus rapidement au cours de la période (t_{AD}, t_{cr}) . Dans l'intervalle (t_{cr}, t_L) , la valeur de l'information est négative. Toujours dans les réseaux de transmission de données, on peut introduire le

¹³Par analogie aux réseaux classiques de transports.

paramètre indicateur de confiance de la réception de l'information. Lorsque le réseau n'est pas automatisé l'incertitude q est assimilée à l'erreur humaine lors de la lecture d'un texte ¹⁴.

6 Optimisation des performances du service ou contrôle du système

L'organisation rationnelle du service pose aux décideurs de nombreux problèmes liés à l'amélioration de la qualité du service. De manière générale, le problème sera formulé mathématiquement comme suit. Trouver une stratégie de contrôle des paramètres du système (ou seulement certains d'entre eux) qui garantisse les "meilleures" performances, dans un sens donné.

La fonction objective à optimiser peut être :

- L'une des caractéristiques de performance décrites dans le paragraphe précédent.
- Les coûts globaux d'exploitation du système que l'on exprimera comme une fonctionnelle dépendant de U , ou seulement de certains des éléments de cet ensemble.

Ci-dessous, nous tentons de passer en revue succinctement les questions typiques auxquelles on est le plus souvent confronté. Cette liste n'est pas exhaustive et on pourra se référer à des ouvrages cités en référence [38, 51, 100].

6.1 Choix du nombre de serveurs

En guise de critère économique, on peut choisir la fonctionnelle

$$F(m) = \bar{S}C_s + C_w\bar{Q} \longrightarrow \min_m$$

où

- C_s , est le coût unitaire d'exploitation d'un serveur,
- \bar{S} , le nombre moyen de serveurs actifs,
- C_w , le coût unitaire d'attente par client,
- \bar{Q} , la taille moyenne de la file d'attente.

Ce problème a été étudié par [Morse P.M (1958)][152] et [Bhat U.N. (1972)][25] pour le système $M/M/m$. Pour le système avec refus, le critère est la minimisation de la probabilité de refus. C'est le problème de dimensionnement d'un réseau téléphonique auquel nous avons fait allusion précédemment ¹⁵

¹⁴Il est admis par exemple que cette incertitude $q \geq 10^{-3}$ par caractère lors d'une transmission rapide et unique. Elle oscille entre 10^{-6} et 10^{-12} pour les stations automatiques [Zakharov (1982)]

¹⁵Erlang A.K. (1948), On the Rational Determination of the Number of Circuits (voir Brockmeyer & Ciel).

Le contrôle optimal de la taille de la file $M/M/m$ a été formulé par Moder J.J. [150] et Phillips C.R. de la manière suivante. Le service débute avec un nombre minimal de serveurs $m = m_0$. Dès que la file atteint une valeur fixée N , on met en marche un certain nombre de serveurs (i.e. on augmente m). Le nombre maximal de serveurs pouvant fonctionner simultanément est connu m_1 , ($m_0 \leq m \leq m_1$). Si à un instant donné la taille de la file devient inférieure au seuil critique N , les serveurs supplémentaires sont "déconnectés".

Le problème de Morse P.M. peut-être formulé pour les réseaux : il s'agit d'une affectation optimale des serveurs par station selon un critère économique analogue à celui de Morse.

6.2 Contrôle du taux de service

Lorsque le nombre de serveurs est fixé, on peut penser à réguler certains paramètres tels que le flux des arrivées ou le processus de service. Généralement, le contrôle de flux des arrivées est difficile. L'ordonnancement des arrivées est étudié par exemple par Healy K. (1992). Voyons comment on pourrait contrôler les mécanismes de service.

Pour le système $M/M/1$, on peut considérer le critère économique précédent

$$F = F(\mu) \longrightarrow \min_{\mu}.$$

Le taux optimal de service est de la forme [Morse (1958)][151] :

$$\mu^* = \lambda + \sqrt{\lambda C_w / C_s}, \text{ où } \lambda \text{ est le taux d'entrée.}$$

La stratégie de service du système $M/M/1$ proposée par [Yadin M. et Naor P. (1963)] consiste en ce qui suit. Soit $\{\mu_k, k \geq 0\}$ la suite des taux de service admissibles ($\mu_0 = 0, \mu_{k+1} > \mu_k$), et soient les suites $\{R_k, k \geq 1\}, \{S_k, k \geq 0\}$ telles que :

$$R_{k+1} > R_k, S_{k+1} > S_k, R_{k+1} > S_k, S_0 = 0.$$

A chaque fois que la taille de la file atteint la valeur R_k , on augmente le taux de service de μ_{k-1} à μ_k . Si par contre la taille de la file diminue jusqu'à la valeur S_k , alors on diminue le taux de service de μ_{k+1} à μ_k .

L'optimalité pour le système $M/M/1$ de ce type de stratégie a été établie par [Grabill T.B. (1972)].

6.3 Stratégie de connexion des serveurs

C'est un procédé particulier de régulation du service étudié par [Heyman D.P. (1968)]. Le serveur est connecté dès que le nombre de clients dans le système atteint la valeur k , et il est déconnecté lorsque le système se vide. Le critère économique dans ce cas tient compte des coûts d'exploitation des serveurs, des coûts d'attente des clients, ainsi que des

coûts liés aux opérations de connexion et de déconnexion. La méthode de résolution est basée sur la programmation dynamique [181].

La stratégie de Sobel M.J. (1962), ou $(N - m)$ -stratégie, consiste en ce qui suit. Si la taille de la file est inférieure à m , alors le service débute à l'instant où la file atteint la valeur N . Le service se prolonge tant que la taille de la file ne descend pas jusqu'à la valeur m . Pour $m = 0$ on obtient la stratégie de Heyman. La stratégie de [Bobrovitch O., Medvediev G. (1985)] est randomisée au sens où le taux de service est choisi selon une distribution p_k . Pour les systèmes avec service par groupe, on peut formuler le problème du choix de la taille optimale du groupe [Deb. R, Serfozo (1973)]. Pour plus de détails sur cette question, on pourra consulter la synthèse de Teghem (1986).

6.4 Affectation des priorités ou ordonnancement

Considérons le cas d'un serveur qui assure le service de n types de clients. Le client de type i a la priorité sur le client de type j si $i < j$. Le problème consiste alors à déterminer la variance optimale d'affectation des priorités aux différentes classes de clients, qui minimise les coûts moyens :

$$F = \sum_{i=1}^n C_i \bar{N}_i$$

où C_i est le coût unitaire de séjour d'un client de type i dans le système, et \bar{N}_i , le nombre moyen de clients de type i dans le système. Les flux d'arrivées des clients doivent être numérotés dans l'ordre de décroissance des quantités $C_j/\bar{\tau}_j$, où $\bar{\tau}_j$ est la durée moyenne de service d'un client de type j . Cette règle est valable aussi bien pour les priorités relatives, qu'absolues (lorsqu'il y a possibilité d'interruption de service du client en cours). D'autres types de priorités sont décrites dans les monographies de [Jaiswal N.K. (1968)][111], [Bronstein O.I., Dukhovny I.M. (1976)][38], [Kleinrock L. (1976) (vol 2)[125], Conway Maxwell, Miller (1967)[51]]. Les problèmes d'ordonnancement font l'objet de notations et typologie spécifique. Par exemple, ajouter dans les notations de Kendall-Lee la fonction objectif au critère à optimiser.

6.5 Problème d'affectation des réparateurs à la maintenance

Le problème classique de la maintenance de N machines par une brigade de m techniciens (analogue à celui du paragraphe 6.1) peut être envisagé du point de vue suivant : Est-il préférable d'affecter l'ensemble des techniciens à la maintenance de toutes les machines ou alors faut-il procéder à une décentralisation du type $m/2$ techniciens pour $m/2$ machines ? [(Gnedenko (1969)[87]). Un problème analogue est étudié par [Stidham (1970)].

Le problème d'affectation peut-être formulé de manière plus adéquate à l'aide des réseaux de files d'attente [Medvediev (1978)]. Chaque station S_1, S_2, \dots, S_n représente un

type d'intervention (électrique, mécanique ...). Il s'agit de trouver une affectation optimale (m_1, m_2, \dots, m_n) des serveurs par station (m_i est le nombre de techniciens de spécialité i , $i = \overline{1, n}$) minimisant les coûts moyens.

6.6 Problèmes de routage dans les réseaux

Ce type de problème a été en particulier posé lors de la conception de certains réseaux informatiques tels que ARPA ou ARPANET. [Kleinrock L. (1976) (vol. 2)[125]] formule les quatre problèmes d'optimisation suivants.

a. Problème du choix de la capacité de transmission (CCT).

Données : - Paramètres des flux d'arrivées dans chaque station.
- Topologie du réseau.

Objectif : Minimiser la durée de séjour du client (le message) dans le réseau.

Variables : débit des canaux.

Contraintes : Coûts des canaux de communication.

b. Problème de distribution des flux (DF).

Données : - Capacités de transmission des canaux.
- Topologie du réseau.

Objectif : Idem.

Variables : Flux des arrivées dans chaque station.

c. Problème mixte (CCT) et (DF).

Données : Topologie du réseau.

Objectif : Idem.

Variables : Capacités des canaux et flux d'arrivées aux stations.

Contraintes : Coûts des canaux de communication.

d. Choix de la topologie (CCT) et (DF).

Objectif : Idem.

Variables : Variantes topologiques.

Capacités de transmission.

Flux d'arrivées aux stations.

Contraintes : Idem.

Ces questions sont particulièrement discutées dans la monographie de [Schwartz (1981)][190] et de [Doyon (1989)][67]. Des problèmes similaires sont traités en relation avec les problèmes de transmission de l'information sur la "toile".

6.7 Reconnaissance optimale de formes

La reconnaissance de formes est une étape importante en analyse d'image en vue de diverses applications : contrôle de qualité au niveau d'une chaîne de production, analyse d'image par satellites, en météorologie ou géographie, ... On désire classer des formes d'un ensemble \hat{f} partitionné au préalable en k classes $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_k$. Une classification $d(x)$ de la forme x , $x \in \hat{f}$ est une application de \hat{f} dans $\{1, 2, \dots, k\}$.

La classification $d(x)$ est correcte si $d(x) = i$ et $x \in \hat{f}_i$; dans le cas contraire elle est erronée.

Un discriminateur séquentiel de formes (Σ) est un dispositif qui associe à chaque forme x une séquence de classifications $\{d_i(x)\}_{i \geq 1}$ (voir par exemple l'ouvrage classique de Pattern Decognition. Academic press 1968). Dans le but d'évaluer les performances d'un tel discriminateur, Viscolani B. (1984) définit le problème d'attente suivant : On suppose que les formes arrivent au classifieur à des instants aléatoires selon un flux de Poisson et qu'elles peuvent constituer une file d'attente. Une discipline de service détermine la classification de chaque forme par un ou plusieurs stades consécutifs. Après avoir calculé $d_m(x)$, ce qui demande un temps aléatoire τ_m , le discriminateur commence à calculer la $(m + 1)^{\text{ème}}$ classification $d_{m+1}(x)$ de la même forme x avec une probabilité α_m , ou bien stoppe la séquence courante de la classification avec $1 - \alpha_m$.

Dans le dernier cas, Σ est prêt à commencer une séquence de classifications pour une éventuelle nouvelle forme. De plus, le discriminateur peut être interrompu à chaque instant durant son opération, à l'exception du moment où il calcule la première classification d_1 . On peut associer un système de files d'attente à cette méthode de classification, l'objectif étant de garder la taille de la file la plus petite possible, et de minimiser la probabilité d'erreur de classification.

Un autre problème consiste à minimiser le temps de gaspillage, c'est-à-dire la partie du temps de travail que le discriminateur passe à calculer toute classification sans finir l'opération à cause d'une nouvelle forme arrivée. Ce type de protocoles est destiné à exploiter les périodes d'inactivité du serveur. Ces modèles sont similaires aux modèles avec vacances développés pour d'autres applications dans les systèmes de Télécommunications ou de protection.

7 Identification du modèle

L'observation d'un phénomène concret permet en général de définir immédiatement certains paramètres tels que le nombre de serveurs, la capacité de la file et la discipline de service, pour un système d'attente.

Afin d'estimer les caractéristiques probabilistes du flux des arrivées $\{\xi_n\}$ ou du processus de service $\{\tau_n\}$ (cela revient à estimer les fonctions de répartition $A(x) = P(\xi_n < x)$ ou $H(x) = P(\tau_n < x)$), on peut faire appel aux méthodes statistiques usuelles.

Tester l'hypothèse $H_0 : F(x) \equiv F_0(x)$, où $F(x)$ est la fonction de répartition qu'on cherche à estimer, et $F_0(x)$ est la fonction de répartition théorique choisie parmi les classes de distributions usuelles (exponentielle, Erlang, Cox, ...).

Soit $\tilde{X} = \{X_1, \dots, X_n\}$ un n -échantillon de la variable aléatoire X , de fonction de répartition $F(x) = P(X < x)$, et qui constitue l'ensemble des observations. On peut assimiler \tilde{X} à un point de l'espace Euclidien R^n . Un test d'ajustement de la fonction de répartition $F(x)$ (inconnue) par la fonction de répartition $F_0(x)$ (connue) consiste à construire un domaine S_α de l'espace R^n tel que $P(\tilde{X} \in S_\alpha) = \alpha$, où α est le niveau du test fixé généralement à 1%, 5% ou 10%.

Pour une réalisation donnée de l'échantillon, on rejette l'hypothèse H_0 si $\tilde{X} \in S_\alpha$, on l'accepte si $\tilde{X} \notin S_\alpha$. Le domaine S_α est appelé région critique (de rejet), et α la probabilité d'erreur de première espèce (de rejet de l'hypothèse H_0 alors qu'elle est vraie).

Par définition, la fonction de répartition empirique.

$$\hat{F}(x) = \begin{cases} 0 & \text{si } x < X_{(1)} \\ i/n & \text{si } X_{(i)} < x < X_{(i+1)} \\ 1 & \text{si } x \geq X_{(n)} \end{cases}$$

où $\{X_{(1)}, \dots, X_{(n)}\}$ est la statistique d'ordre de l'échantillon (la suite ordonnée dans le sens croissant). Il est connu que $\hat{F}(x)$ est un estimateur sans biais et convergent de la fonction de répartition $F(x)$ inconnue :

$$E\{\hat{F}(x)\} = F(x) \quad \text{et} \quad P(\lim_{n \rightarrow \infty} \hat{F}(x) = F(x)) = 1.$$

De plus, la distance entre ces deux fonctions :

$$P(\lim_{n \rightarrow \infty} |\hat{F}(x) - F(x)| = 0) = 1 \quad (\text{Théorème de Glivenko})$$

L'idée des tests d'ajustement consiste ainsi à comparer l'écart entre les distributions empiriques et théoriques (sous H_0). Considérons quelques types de mesures de cet écart, conduisant à différents critères.

7.1 Test du Khi-deux

Les observations portent sur la variable aléatoire X (par exemple ξ_n ou τ_n) et forment l'échantillon de base ¹⁶ $\tilde{X} = \{x_1, \dots, x_n\}$ qui permet de construire la fonction de répartition empirique $\hat{F}(x)$.

La distribution empirique doit cependant être représentée sous une forme quelque peu différente. L'espace des valeurs possibles de X , (R^+ en général) est partagé en k intervalles disjoints :

$$I_1 = [a_0, a_1), I_2 = [a_1, a_2), \dots, I_k = [a_{k-1}, a_k) \quad \text{avec } (a_0 = 0, a_k = \infty).$$

Soit $p_j = F_0(a_j) - F_0(a_{j-1})$, la probabilité théorique (si H_0 est vraie) pour que la variable X soit dans l'intervalle I_j .

¹⁶Lorsqu'elles forment des suites de variables aléatoires indépendantes, de même loi.

Le test est basé sur la statistique de décision :

$$T = \sum_{j=1}^k \frac{(np_j - m_j)^2}{np_j}$$

où m_j est le nombre d'observations dans l'intervalle I_j , et np_j le nombre théorique (si H_0 est vraie) correspondant. Cette statistique sert de mesure de l'écart entre les distributions empirique et théorique. On rejettera l'hypothèse H_0 si la valeur observée de T est trop importante.

Lorsque H_0 est vraie, la statistique T obéit asymptotiquement à une loi de Khi-deux ξ ($n \rightarrow \infty$) à $k - 1$ degrés de liberté. On rejette donc H_0 au seuil α si $T > \chi_{1-\alpha}^2(k - 1)$ où $\chi_{1-\alpha}^2(k - 1)$ est le $(1 - \alpha)$ quantile de la distribution de χ^2 à $k - 1$ degrés de liberté i.e. $P(T > \chi_{1-\alpha}^2(k - 1)) = \alpha$.

Si la loi hypothétique $F_0(x)$ comporte m paramètres $(\alpha_1, \dots, \alpha_m)$, on peut les estimer par la méthode du minimum de Khi-deux, et on utilise la règle de Fisher stipulant que T obéit également à une loi de Khi-deux à $k - m - 1$ degrés de liberté.

Dans la règle de Fisher, les paramètres sont choisis tels que :

$$T(\alpha_1, \dots, \alpha_m) = \min_{\alpha'_1, \dots, \alpha'_m} T(\alpha'_1, \dots, \alpha'_m)$$

où

$$T(\alpha'_1, \dots, \alpha'_m) = \sum_{j=1}^k \frac{(m_j - np_j(\alpha'_1, \dots, \alpha'_m))^2}{np_j(\alpha'_1, \dots, \alpha'_m)}$$

Les paramètres $\alpha'_1, \dots, \alpha'_m$ peuvent être estimés par la méthode du maximum de vraisemblance :

$$\max_{\alpha'_1, \dots, \alpha'_m} \prod_{i=1}^n f(X_i, \alpha'_1, \dots, \alpha'_m) = \prod_{i=1}^n f(X_i, \alpha_1, \dots, \alpha_m)$$

où $f(x, \alpha_1, \dots, \alpha_m)$ est la fonction de densité de X .

Cette règle conduit à un test moins bon, mais acceptable.

Exemple 1.

Les données ci-dessous portent sur $n = 42$ observations répartis sur k intervalles $\xi_n = t_n - t_{n-1}$, entre arrivées successives de clients dans un système d'attente. On cherche à tester si le flux $\{t_n\}$ est Poissonien. Cela revient à tester l'exponentialité des intervalles ξ_n . D'où l'hypothèse nulle $H_0 : A(x) = 1 - e^{-\lambda x}$, où $A(x) = P(\xi_n < x)$.

Pour $k = 6$, $n = \sum_{i=1}^6 m_i = 42$.

La moyenne empirique est estimée par $\bar{\xi} = \sum_{i=1}^k \left\{ \frac{a_{i-1} + a_i}{2} \right\} \frac{m_i}{n} = 150,5$ heures et le

N° de la classe i	Classe $I_i = [a_{i-1}, a_i)$	Nombre n_i de valeurs observées (ξ_n) dans I_i .	Nombre théorique np_i de valeurs de ξ_n dans I_i .
1	0 - 50	5	11,76
2	50 - 100	6	8,4
3	100 - 150	11	6,3
4	150 - 200	8	4,2
5	200 - 250	7	3,36
6	250 - 300	5	8,06

paramètre $\lambda = 1/\bar{\xi} \simeq 0,0066$.

Les probabilités théoriques (si H_0 est vraie) sont

$$p_i = A(a_i) - A(a_{i-1}) = (1 - e^{-\lambda a_i}) - (1 - e^{-\lambda a_{i-1}}) = e^{-0,0066 a_{i-1}} - e^{-0,0066 a_i}.$$

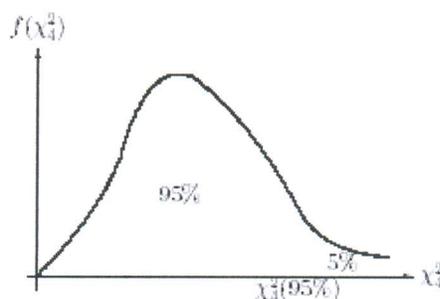


FIG. 6 - Loi du Khi-deux à 4 degrés de liberté

$T \simeq 18$ et $\chi_4^2(95\%) = 9,49$, il faut donc rejeter H_0 .

Le test de Khi-deux est intéressant pour l'ajustement des distributions discrètes (Poisson, par exemple) : il s'agit de tester l'hypothèse d'un flux d'arrivées Poissonnien. Les observations portent sur le nombre d'arrivées par tranches horaires.

Exemple 2.

On a observé le nombre d'arrivées dans chacun des 2068 intervalles de temps de longueur 7,5 secondes.

On a $n = \sum_{i=0}^{10} m_i$, avec m_i = nombre d'intervalles où on a enregistré i arrivées. L'hypothèse nulle est $H_0 : F(x) = \sum_{j=0}^{x-1} e^{-\lambda} \frac{\lambda^j}{j!}$.

L'estimateur du paramètre λ est $\hat{\lambda} = \sum i m_i / n = 3,870$ et les $p_i = e^{-\hat{\lambda}} \frac{\hat{\lambda}^i}{i!}$, $i = \overline{0,10}$. $T \simeq 13,05$, pour un degré de liberté $11 - 1 - 1 = 9$, on trouve dans la table de Khi-deux ($\alpha = 5\%$), $\chi_{1-\alpha}^2(9) = 16,919 > 13,05$. On accepte donc H_0 .

i Nombre d'arrivées	m_i Nombre d'intervalles	p_i	np_i
0	57	0,021	054,8
1	208	0,081	211,2
2	383	0,156	406,8
3	525	0,201	524,2
4	532	0,195	508,6
5	408	0,151	393,2
6	273	0,097	253,0
7	139	0,054	140,8
8	45	0,026	067,8
9	27	0,011	028,7
10	16	0,070	018,3

7.2 Test de Kolmogorov-Smirnov

L'écart entre les distributions empirique et théorique est mesuré à l'aide de la statistique :

$$D_n = \sup_{-\infty < x < \infty} |\hat{F}(x) - F_0(x)| = \max_{1 \leq i \leq n} \left\{ \left| \frac{i}{n} - F_0(X_{(i)}) \right|, \left| \frac{i-1}{n} - F_0(X_{(i)}) \right| \right\}.$$

La distribution de D_n est tabulée et on peut déterminer la valeur $d_{n,\alpha}$ (pour α donné), telle que $P(D_n > d_{n,\alpha}) = \alpha$. On rejette l'hypothèse nulle si la valeur observée de D_n dépasse $d_{n,\alpha}$.

7.3 Test de Cramer-Von Mises

$$\begin{aligned} w_n^2 &= \int_{R^+} [\hat{F}(x) - F_0(x)]^2 dF_0(x) \\ &= \frac{1}{12n} + \sum_{i=1}^n \left\{ F_0(X_{(i)}) - \frac{(2i-1)}{2n} \right\}^2. \end{aligned}$$

Cette statistique est tabulée.

Les avantages des critères basés sur les statistiques D_n et w_n^2 par rapport au critère du Khi-deux sont les suivants :

- Les critères ne dépendent pas de la distribution de l'échantillon, et ils ne nécessitent pas un groupement des observations en classes. D'autre part, le critère D_n est aussi simple que celui du Khi-deux (du point de vue calcul).
- Pour les grandes valeurs de n , le critère D_n est plus puissant que celui du Khi-deux. La comparaison des puissances est toutefois délicate.

7.4 Autres tests

Test d'Anderson-Darling [Cox D.R., Lewis P (1966)][54]

$$A_n = n \int_R \left(\frac{\hat{F}(x) - F_0(x)}{\sqrt{F_0(x)(1 - F_0(x))}} \right)^2 dF_0(x).$$

Test de Kuiper N. (1960)

$$V_n = \sup_{-\infty < x < \infty} |\hat{F}(x) - F_0(x)| - \inf_{-\infty < x < \infty} |\hat{F}(x) - F_0(x)|.$$

Test de Watson G.S. [Watson G.S.] (1961)

$$U_n = n \int_{R^+} (\hat{F}(x) - F_0(x)) - \int_R ((\hat{F}(x) - F_0(x)) dF_0(x))^2 dF_0(x).$$

Test de Finkelstein-Shafer [Finkelstein-Shafer(1971)]

$$K_n = \sum_{i=1}^n \max\left\{ \left| \frac{i}{n} - F_0(X_{(i)}) \right|, \left| \frac{i-1}{n} - F_0(X_{(i)}) \right| \right\}.$$

Il existe des tests spécifiques à la loi exponentielle, notamment en présence de données censurées, caractéristiques des problèmes de fiabilité et d'analyse de survie [5, 87]. Dans le livre de Cox D.R., Lewis P. (1966)[54] on peut trouver des tests liés aux problèmes de stationnarité et d'indépendance.

D'autres problèmes d'identification sont décrits dans l'ouvrage de Kalashnikov V.V., Rachev S. (1988)[113].

8 Conclusion et Perspectives

Le développement des Systèmes Informatiques et des Réseaux de Télécommunication se caractérise par un accroissement de la complexité dû à l'interfonctionnement de réseaux hétérogènes et à la diversification des services et des natures de trafic. Le développement de l'intelligence dans les réseaux et la diversification des protocoles contribuent à cette complexité.

Les besoins sont donc de modélisation des comportements de l'utilisateur, de simulation du fonctionnement de systèmes complexes et évolutifs et d'évaluation de leurs performances. Il faut donc pouvoir assurer une gestion dynamique des ressources, en optimisant le dimensionnement des réseaux et en réalisant de bonnes prédictions du trafic. Les nombreuses approches développées ces dernières années (voir la bibliographie) permettent d'assurer la disponibilité des réseaux et l'adaptation de leur qualité aux services attendus, dans un contexte d'interfonctionnement [7, 6, 9, 8, 33, 34, 35, 59, 65, 92, 94, 107, 106, 140, 207].

9 Bibliographie

Références

- [1] V. Ahuja. *Design and Analysis of Computer Communication Networks*. McGraw Hill, New York, 1992.
- [2] A.O. Allen. *Probability Statistics and Queueing Theory with Computer Science Applications*. Academic Press, New York, 1978.
- [3] V. Anissimov and A. Voina. *Processus markoviens et semi-markoviens*. Université de Kiev., 1986.
- [4] W.D. Ashton. *The Theory of Road Traffic Flow*. Methuen London, 1966.
- [5] A. Aïssani. *Modèles Stochastiques de la Théorie de Fiabilité*. Office des Publications Universitaires, Alger, 1992.
- [6] A. Aïssani and D. Aïssani. Méthodes mathématiques d'analyse des phénomènes d'attente. (*à paraître aux Editions Scientifika - Paris*).
- [7] A. Aïssani and D. Aïssani. Réseaux de files d'attente. *Institut d'Informatique, U.S.T.H.B Alger*, pages 1–60, 1988.
- [8] D. Aïssani. Analyse, optimisation et modèles stochastiques. *Recueil, Ed. LAMOS Béjaïa, 2-ème édition augmentée, Béjaïa*, pages 1–200, Avril 1994, 2004.
- [9] D. Aïssani and al. Evaluation des performances des réseaux de télécommunication. *Séminaire du Laboratoire d'Informatique de Marseille, Marseille*, Juin 2001, 2002.
- [10] D. Aïssani and N.V. Kartashov. Ergodicity and stability of markov chains with respect to operator topology in the space of transition kernels. *Comptes Rendus Acad. Sciences U.S.S.R, serie A*, 11 :3–5, 1983.
- [11] F. Bacceli and P. Brémaude. *Palm Probabilities and Stationary Queues*. Springer, New York, 1987.
- [12] Bachelier. *Calcul des Probabilités*. Gauthier-Villars, Paris, 1912.
- [13] T.L. Bagchi and J.G.C. Templeton. *Numerical Methods in Markov chains and Bulk queues*. Springer-Verlag, Berlin, 1972.
- [14] K.R. Baker. *Introduction to Sequencing and Scheduling*. John Wiley, New York, 1974.
- [15] R. Barlow and F. Proshan. *Mathematical Theory of Reliability*. John Wiley, New York, 1965.
- [16] R. Barlow and F. Proshan. *Statistical Theory of Reliability and Life Testing*. Rinehart & Winston, New York, 1975.
- [17] R.E. Barlow, D.J. Babartholomey, J.M Brenner, and H.D. Brunk. *Statistical inference under order restrictions*. Wiley and Sons, New York, 1972.
- [18] G.P. Basharin, A.D. Kharkevitch, and M.A Shneps. *Queueing Theory in Teletraffic*. Nauka, Moscow, 1968.

- [19] D. Bear. *Pinciples of Telecommunication Traffic Engineering*. Peregrinus, London, 1976.
- [20] P. Beckman. *Introduction to Elementary Queueing Theory and Telephone Traffic*. Golem Press, Boulder, 1968.
- [21] F. Beichelt and P. Franken. *Zwver Lassigket un instandhaltung Matematishe methoden*. Veb-Verlag Technik, Berlin (Ed. Nauka. Moscow), 1983.
- [22] V.W. Benès. *General Stochastic Processes in the Theory of Queues*. Addison-Wesley Reading, Massachusetts, 1963.
- [23] V.W. Benès. *Mathematical Theory of Connection and Telephone Traffic*. Academic Press, New York, 1965.
- [24] D. Bertsekas and R. Gallager. *Data Networks*. Prentice-Hall, Englewood Cliffs, 1987.
- [25] U.N.A. Bhat. *Elements of Applied Stochastic Processes*. John Wiley, New York, 1972, (2ème ed. 1984).
- [26] U.N.A. Bhat. *Study of the Queueing Systems M/G/1 and GI/M/1*. Springer-Verlag, Berlin, 1989.
- [27] J.P.C. Blanc. *Application of the Theory of Boundary Value Problem in the Analysis of a Queueing Model with Paired Services*. Mathematics Centrum, Amesterdam, 1982.
- [28] R.P. Blanc and I.W. Cotton. *Computer Network*. Ed. IEEE Press, New York, 1976.
- [29] G. Bloch, H. Greiner, H. de Meer, and K.S. Trivedi. *Queueing Networks and Markov Chains Modeling and Performance Evaluation with Computer Science Applications*. Wiley, 1998.
- [30] F.T. Boesch. *Large Scale Networks : Theory and Desing*. Ed. IEEE Press, New York, 1976.
- [31] A. Borovkov. *Stochastic Processes in Queueing Theory*. Springer-Verlag, Berlin, 1976, (Ed. Nauka, 1972).
- [32] A. Borovkov. *Asymptotic Methods in Queueing Theory*. Nauka, Moscow, 1980, (Ed. Wiley, 1984).
- [33] L. Bouallouche and D. Aïssani. Performance evaluation of an sw communication protocol (send and wait). *Proceedings of the MCQT'02 (First Madrid International Conference on Queueing Theory), Madrid (Spain)*, pages 18–26, July 2002.
- [34] L. Bouallouche, S. Taghzouti, and D. Aïssani. Calculation of the optimal degree of multiprogramming of a central server with two processors. *Proceedings of the International Conference CCCT04 (Computing, Communications and Control Technologies), Austin (Texas - U.S.A)*, 2004.
- [35] K. Boukhetala and D. Aïssani. Les projets sisa (systèmes d'information statistiques en algérie). *Proceedings de la Journée d'Etudes Nationale BDTS03 (Banques de Données et Traitement Statistique), Béjaia*, pages 05 – 10, Mars 2003.

- [36] P. Brémaud. *Point Processes and Queues : Martingale Dynamics*. Springer-Verlag, New-York, 1980.
- [37] E. Brockmeyer, H.L. Halstrom, and A. Jensen. *The life and works of Erlang A.K.* Transactions of the Danish Academy Technical Sciences, N°2, Copenhagen, 1948.
- [38] O.I. Bronstein and I. Dukhovny. *Priority Queueing Models in Computer Systems*. Nauka, Moscow, 1976.
- [39] M.L. Chandhry and J.G.C. Templeton. *A First Course in Bulk Queues*. John Wiley, New York, 1983.
- [40] K.M. Chandy and M. Reiser. *Computer Performance*. North-Holland, Amesterdam, 1977.
- [41] X. Chao, M. Miyazawa, and M. Pinedo. *Queueing Networks : Customers, Signal and Product Form Solutions*. Wiley, 1999.
- [42] W. Chou. *Computer Communications*. Ed.Vol.1/1983 Vol.2/1985 Prentice-Hall, Englewood Cliffs, 1983.
- [43] E. Cinlar. *Introduction to Stochastic Processes*. Prentice Hall : Englewood Cliffs, New York, 1975.
- [44] A.B. Clarke. *Mathematical Models in Queueing Theory*. Springer-Velag, Berlin, 1974.
- [45] E.G.Jr. Coffman. *Computer and Job Shop Scheduling Theory*. John Wiley, New York, 1976.
- [46] E.G.Jr. Coffman and P.J. Denning. *Operating Systems Theory*. Prentice-Hall, Englewood Cliffs, 1973.
- [47] W.J. Cohen. *The Single-Sever Queue*. North-Holland, Amesterdam, 1969, (2nd Ed. 1982).
- [48] W.J. Cohen. *On Regenerative Processes in Queueing Theory, Lecture Notes in Mathematical Economics and Mathematical Systems*. Vol.121, Springer-Velag, Berlin, 1976.
- [49] W.J. Cohen and O.J. Boxma. *Boundary Value Problems in Queueing System Analysis*. North-Holland, Amesterdam, 1983.
- [50] B.W. Conolly. *Queueing System*. Ellis Horwood, Chichester, 1975.
- [51] R.W. Conway, W.L. Maxwell, and L.W. Miller. *The Theory of Scheduling*. Addison-Wesly Reading, Massassuchetts, 1967.
- [52] R.B. Cooper. *Introduction to Queueing Theory*. Edwar Arnod London : 2-nd Elsevier North Holland, Amesterdam, Mc Millan 1972. N.Y., 1981.
- [53] P.J. Courtois. *Decomposability Queueing and Computer System Applications*. Academic Press, London, 1977.
- [54] D.R. Cox and P.A. Lewis. *The Statistical Analysis of Series of Events*. Methuen and Co, London, 1966.
- [55] D.R. Cox and W.L. Smith. *Queueing Systems*. Methuen, London, 1961.

- [56] M.A. Crane and A.J. Lemoine. *An Introduction to Regenerative Method for Simulation Analysis*. Springer-Verlag, New York, 1977.
- [57] R. Cruon. *Queueing Theory, Recent Developments and Applications*. Elsevier, New York, 1967.
- [58] G. Cullmann. *Recherche opérationnelle Théorie et Pratique*. Masson, Paris, 1970.
- [59] N. Dehas, S. Adjabi, and D. Aïssani. Discrete events simulation model for reliability problem in telecommunication network. *Proceedings of the International Conference CITSA 2004 - Cybernetics and Information Technologies System and Applications - Orlando (U.S.A)*,, 2004.
- [60] G. Desbazzelle. *Exercices et Problèmes de Recherche Opérationnelle*. Dunod, Paris, 1972.
- [61] A. Descloux. *Delay Tables for Finite and Infinite Source Systems*. Mc Graw-Hill, New York, 1962.
- [62] R.L. Disney, R.L. Farell, and P.R. De Morais. *A Characterization of M/G/1/N Queues with Renewal Departure Processes*, volume 19 of 11. Management Sci, 1973.
- [63] R.L. Disney and P.C. Kiessler. *Traffic Processes in Queueing Networks : a Markov Renewal Approach*. Johns Hopkins, University Press, Baltimore, 1987.
- [64] R.L. Disney and T.J. Ott. *Applied Probability and Computer Science* : The Interface Ed. Volumes I et II Brikhauser, Boston, 1982.
- [65] N. Djellab. Systèmes de files d'attente avec rappels : Méthode d'approximation pour un système m/g/1 avec rappels et pannes. *Thèse de Doctorat d'Etat en Informatique, Faculté des Sciences, Université de Annaba*, 2003.
- [66] D.R. Doll. *Data Communications : Facilities Networks and system Design*. John Wiley, New York, 1980.
- [67] G. Doyon. *Systèmes et Réseaux de Télécommunication en Régime Stochastique*. Masson, Paris, 1989.
- [68] A.K. Erlang. *The Theory of Probability and Telephone Conversation*. Nyt-tidsskrift for matematik, B20, 33, 1909.
- [69] G.I. Falin and J.G.C. Templeton. *Retrial Queues*. Chapman and Hall, 1997.
- [70] R. Faure. *Précis de recherche opérationnelle*. Dunod, Paris 4ème Ed., 1979.
- [71] S. Fdida and G. Pujolle. *Modèles de systèmes et de réseaux : Performances (tome1), Files d'attente (tome2)*. Eyrolles, 1989.
- [72] A. Federgruen and L. Green. *Queueing Systems with Service Interruptions*. Operations Research. Vol. (34) N° (5), 1985.
- [73] W. Feller. *An Introduction to Probability Theory and its Applications*. Wiley, New York, 1968.
- [74] D. Ferrari. *Computer System Performance Evaluation*. Prentice-Hall, Englewood Cliffs, 1978.

- [75] P. Franken, D. Konig, U. Arndt, and V. Schmidt. *Queues and Point Processes*. Academie-Verlag, Berlin, 1981.
- [76] T.C. Fry. *Probability and its Engineering Uses*. Van Nostrand Reinhold, New York, 1928.
- [77] E. Gelenbe. *Multiprocessor performances*. Jhon Wiley Sons, 1989.
- [78] E. Gelenbe and al. *Réseaux de Files d'Attente, Modelisation et Traitement Numérique*. Ed. Hommes et Techniques, 1980.
- [79] E. Gelenbe and I. Mitrani. *Analysis and Synthesis of Computer Systems*. Academic Press, New York, 1980.
- [80] E. Gelenbe and G. Pujolle. *Introduction aux Réseaux de Files d'Attente*. Eyrolles Paris, (Engl. Transl. Introduction to Queueing Networks, Chichester), 1982, (Ed. John Wiley 1986).
- [81] F. Georges and H. Gerard. *Trafic et Performances des réseaux de télécoms (coll. Technique et scientifique des télécommunications)*. Andrew Tanenbaum, réseaux, Dunod, 1996.
- [82] D.L. Gerlough and D.G. Cappelle. *An Introduction to Traffic Flow Theory*. Highway Research Board, Washington D.C., 1964.
- [83] A. Ghosal. *Some Aspects of Queueing and Storage Systems*. Springer-Verlag, Berlin, 1970.
- [84] W.C. Giffin. *Queueing : Basic Theory and Applications*. Grid, Columbus, 1978.
- [85] B. V. Gnedenko, G.P. Klimov, and E. Danielan. *Priority Queues*. Moscow University, 1973.
- [86] B. V. Gnedenko and I.N Kovalenko. *Introduction to Queueing Theory*. Nauka, Moscow, 1966, (2nd Ed. Birkhauser Boston/basel/Berlin 1989).
- [87] B.V. Gnedenko, Y.K. Beliaev, and A.D. Soloviev. *Mathematical Methods of Reliability Theory*. Academic Press, New York (Traduction Francaise Ed. Mir 1972), 1969.
- [88] B. Gopinath. *Computer Communications Proc. Sympusia in Applied Mathematics*. Ed. Volume 31 American Mathematical Society, Providence, 1985.
- [89] Y. Grenander and V. Frayberger. *Short Course of Numerical Probability and Statistics*. Nauka, Moscow, 1978.
- [90] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory*. John Wiley, New York (1974), 1985.
- [91] I.I. Guikhmann and A. Shorokhod. *Introduction à la Théorie des Processus Aléatoires*. Ed Mir, Moscou, 1986.
- [92] M. Hadji, N. Touati, D. Aïssani, and F. Laïb. Routage et contrôle de congestion dans les réseaux de télécommunications. (à paraître).
- [93] F.A. Haight. *Mathematical Theory of Traffic Flow*. Academic Press, New York, 1963.

- [94] D. Hamiti, T.R. Atmani, S. Adjabi, and D. Aïssani. Study of the congestion for the dimensioning of a telephone network. *Proceedings of the XXth European Conference on Operational Research EUROXX "O.R. and the Management of Electronic Services", Rhodes - Greece, 2004.*
- [95] M.J. Harrison. *Brownian Motion and Stochastic Flow Systems.* John Wiley, New York, 1985.
- [96] B.R. Haverkort. *Performnace of Computer Communication Systems : A Model Based Approach to Performance Evaluation.* Jhon. Wiley Sons, 1998.
- [97] J.F. Hayes. *Modeling and Analysis of Computer Communications Networks.* Plenum Press, New York, 1984.
- [98] J. Héлары and R. Pedrono. *Recherche Opérationnelle Travaux Dirigés Modèles Stochastiques.* Hermann Collection Méthodes, Paris, 1983.
- [99] D.P. Heyman and M.J. Sobel. *Stochastic Models in Operations Research Vol.1 Stochastic Processes and Operating Characteristics.* Mc Graw Hill, New York, 1971.
- [100] F.S. Hillier and G.J. Liberman. *Introduction to Operations Research.* Holden-Day, San Francisco, 1967.
- [101] F.S. Hillier and O.S. Yu. *Queueing Tables and Graphs.* Elsevier/North-Holland, Amesterdam, 1981.
- [102] R.A. Howard. *Dynamic Probabilistic Systems Vol.1 / Markov Models - Vol.2 / Semi-Markov and Decision Processes.* Wiley, New York, 1971.
- [103] G. Iazeolla, P.J. Courtois, and A. Hordijk. *Mathematical Computer Performance and Reliability.* Ed. Elsevier, North-Holland, Amesterdam, 1984.
- [104] D.L. Ighlehart. Limiting diffusion approximations for the many server queue and the repairman problems. *J. Appl. Prob.*, 2 :429–441, 1965.
- [105] D.L. Iglehart and G.S. Shedler. *Regenerative Simulation of Response Times in Networks of Queues.* Springer-Verlag, New York, 1976.
- [106] S. Imloul. Systèmes avec rappels dans les réseaux de télécommunication. *Thèse de Magister en Mathématiques Appliquées, Département : Informatique et Recherche Opérationnelle, Université de Béjaia, Décembre 2003.*
- [107] S. Imloul and A. Fardjallah. Evaluation des performances de réseaux de télécommunications : cas d'un simulateur d'un réseau de files d'attente. *Mémoire d'Ingéniorat, Département : Informatique et Recherche Opérationnelle, Université de Béjaia, Septembre 2001.*
- [108] G.I. Ivchenko, V.A. Kachtanov, and I.N. Kovalenko. *Queueing Theory.* Vishaia Shkola, Moscow, 1970.
- [109] J.R. Jackson. Jobshop-like queueing systems. *Managment, Science*, 10 :131–142, 1963.
- [110] R. Jain. *The Art of Computer Systems Performance Analysis Techniques for Experimental Design, Measurment, Simulation and Modeling.* Jhon. Wiley Sons, 1991.

- [111] N.K. Jaiswal. *Priority Queues*. Academic Press, New York, 1968.
- [112] Kalashnikov. *Qualitative Analysis of the Behaviour of Complicated Systems using the Method of Trial Functions*. Nauka, Moscow, 1978.
- [113] V. Kalashnikov and S.T. Rachev. *Mathematical Methods for Constuction of Stochastic Queueing Models*. Nauka, Moscow (English Translation : Wadsworth & Brooks/Cole Advenced Books & Softwar Pacific Grove CA.1990), 1988.
- [114] S. Karakoviak. *Principes des Systèmes d'Exploitation des Ordinateurs*. Dunod Informatique, Paris, 1987.
- [115] S. Karlin. *Introduction to the Theory of Stochastic Processes*. Academic Press, New York, 1968.
- [116] N.V. Kartachov. *Strong stable Markov chains*. VSP Utrecht, TbiMC Scientific Publishers, 1996.
- [117] A. Kaufman and R. Cruon. *Les Phénomènes d'Attente : Théorie et Applications*. Dunod, Paris, 1972.
- [118] J. Keilson. *Green's Function Methods in Probability Theory*. Charles Griffin, London, 1965.
- [119] J. Keilson. *Markov Chain Models : Rarity and Exponentiality*. Springer-Verlag, New York, 1979.
- [120] F.P. Kelly. *Reversibility and Stochastic Networks*. John Wiley, 1979.
- [121] A.Y. Khinchine. *Mathematical Methods in the Theory of Queueing*. Charles Griffin, London, 1969.
- [122] R.A. King. The covariance structure of the departure process from M/G/1 queues with finite waiting lines. *J. Royal Statisti. Soc*, 33(3), 1977. Ser. B.
- [123] J.F.C. Kingman. *On the Algebra of Queues, Methuen's Series in Applied Probability*. Methuen, London, 1961.
- [124] L. Kleinrock. *Communication Nets Stochastic Message Flow and Delay*. Dover, New York, 1963.
- [125] L. Kleinrock. *Queueing Systems*. Vol.I / Theory Vol.II / Computer Applications, John Wiley, New York (1975), 1976.
- [126] G.P. Klimov. *Stochastic Service Systems*. Nauka, Moscow, 1966.
- [127] H. Kobayashi. *Modelling and Analysis : An Introduction to System Performance Evaluation Methodology*. Addison-Wesley, Reading, Massachusetts, 1981.
- [128] L. Kosten. *Stochastic Theory of Service System*. Pergamon Press, Oxford, 1973.
- [129] L.S. Kostevitch and A.A. Lapko. *Theory des Jeux, Recherche Opérationnelle*. Ed. Vishania. Kiev, 1982.
- [130] I.N. Kovalenko and I. N. Kouatsus. *Calcul des Indices Probabilistes des Systèmes Complexes*. Technics, University of Stuttgart, 1976.

- [131] P. Kuehn. *Tables on Delay Systems*. Institute of Switching and Data Technics, University of Stuttgart, 1976.
- [132] L.T. Kuzin. *Bases de la Cybernétique*. Vol.1/Vol2/ Energuia, Moscow, (1973), 1979.
- [133] I. Kuzmin. *Simulation des Systèmes Complexes*. Vishaia Schkola, Kiev, 1981.
- [134] LAMOS. *Actes du Colloque International "Méthodes et Outils d'Aide à la Décision"*. Béjaia 12-17 déc. Editions LAMOS, 1992.
- [135] S. Lavenberg. *Computer Performance Modelling Handbook*. Academic-Press, New York, 1983.
- [136] E.D. Lazowska, J. Zahorian, G.S. Graham, and K.C. Sevcik. *Quantitative System Performance : Computer System Analysis Using Queueing Network Theory*. Prentice-Hall, Englewood Cliffs, 1984.
- [137] P. Le Gall. *Introduction to Congestion Theory in Telephone Systems*. Oliver and Boy, London, 2nd Ed. Elsevier (North-Holland Amsterdam 1986), 1960.
- [138] P. Le Gall. *Les Systèmes avec ou sans Attente et les Processus Stochastiques*. Dunod, Paris, 1962.
- [139] A.M. Lee. *Applied Queueing Theory*. Mac Millia, New York, 1966.
- [140] O. Lekadir and D. Aïssani. Strong stability in a jackson network with two tandem stations. *Proceedings of the XXII International Seminar on Stability Problems for Stochastics Models, Varna (Bulgaria)*, May 2002.
- [141] P. Lèvy. *Processus Stochastiques et Mouvement Brownien*. Paris (1948), 1965.
- [142] A.Z.L. Livschitz and E.A. Malts. *Statistical Simulation of Queueing Systems*. Sov. Radio, Moscow, 1978.
- [143] B.S. Livschitz, A.P. Pshenishnikov, and A.D. Kharkevish. *Teletraffic Theory*. Sviaz, Moscow, 1979.
- [144] J. Maurin. *Simulation Déterministe du Hasard*. Masson, Paris, 1975.
- [145] J. Meddhi. *Recent Developments in Bulk Queueing Models*. Wiley Eastern, New Delhi, 1984.
- [146] J. Medhi. *Stochastics Models in Queueing Theory*. Academic Press, Inc., 1991.
- [147] D.A. Menascé and V.A.F. Almeida. *Capacity Planning for Web Performance : Metrics, Models, and Methods*. Prentice-Hall, 1989.
- [148] C.D. Meyer and R. J. Plemmons. *Linear Algebra, Markov Chains and Queueing Models, The IMA Volumes in Mathematics and its Applications*. Ed. A. Friedman W. Miller, Vol.48, XVI. Springer-Verlag, Berlin, 1993.
- [149] I. Mitrani. *Modelling of Computer and Communications Systems*. Cambridge Computer Sciences Texts, Cambridge University Press, 1987.
- [150] J.J. Moder and S.E. El-Maghreby. *Handbook of Operations Research*. Ed. Van Nostrand Reinhold Company, New York, Vol.1, 1978.

- [152] P.M. Morse. *Queues Inventories and Maintenance*. John Wiley, New York (Files d'Attente Stocks et Entretien, Dunod, Paris, 1960)., 1958.
- [153] K. Mosler and M. Scarsini. *Stochastic Orders and Applications Lecture Notes in Economics and Mathematical Systems*. Vol.401, VI Springer-Verlag, Berlin, 1993.
- [154] R.H. Myers, K.L. Wong, and H-M. Gordy. *Reliability Engineering for Electronic Systems*. Wiley, 1964.
- [155] M.F. Neuts. *Algorithmic Methods in Probability Studies in Management Science*. Ed. Volumes 7, North-Holland, Amsterdam, 1977.
- [156] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models : an Algorithmic Approach*. the Johns Hopkins Univ. Press Baltimore, 1981.
- [157] J. Neveu. *Construction de Files d'Attente Stationnaires Lect. Notes on Control and Information Sciences 60*. Springer-Verlag, 1983.
- [158] G.F. Newell. *Approximate Stochastic Behaviour of n - Server Systems with Large n* . Springer-Verlag, Berlin, 1973.
- [159] G.F. Newell. *Approximate Behaviour of Tandem Queues*. Springer-Verlag, Berlin, 1979.
- [160] G.F. Newell. *Application of Queueing Theory*. Chapman and Hall, London, 2nd ed., 1971, 1982.
- [161] G.F. Newell. *The $M/M/\infty$ Service System With Ranked Servers in Heavy Traffic*. North-Holland, Amsterdam, 1984.
- [162] E. Page. *Queueing Theory in OR*. Grane Russek & Company, London, 1972.
- [163] J.A. Panico. *Queueing Theory*. Prentice-Hall Englewood Cliffs, 1979.
- [164] L.G. Peck and R.N. Halzelwood. *Finite Queueing Theory Tables*. John Wiley, New York, 1958.
- [165] J. Pellaumail. *Probabilités, Statistiques et Files d'attente*. Dunod, Paris, 1986.
- [166] J. Pellaumail. *Graphes Simulation L-Matrices*. Hermes, Paris, 1992.
- [167] H.G. Perros. Queueing networks with blocking : A bibliography. *Performance Evaluation Review*, 1984.
- [168] F. Pollaczek. *Problèmes Stochastiques Posés par le Phénomène de Formation d'une Queue d'Attente à Guichet et par des Phénomènes Apparentés*. Gauthier-Villars, Paris, 1961.
- [169] F. Pollaczek. *Théorie Analytique des Problèmes Stochastiques Relatifs à un Groupe de Lignes Téléphoniques avec Dispositif d'Attente*. Gauthier-Villars, Paris, 1961.
- [170] D. Potier. *Modeling Techniques and Tools for Performance Analysis*. Ed. Elsevier North-Holland, Amstardam, 1985.
- [171] N.U. Prabhu. *Stochastics Storage Processes : Queues, Insurance Risk and Dams*. Springer-Verlag, New York, 1980.

- [172] U.N. Prabhu. *Queues and Inventories : A Study of Their Basic Stochastic Processes*. Wiley, New York, 1965.
- [173] G. Pujolle. *Les réseaux*. Eyrolles, 1995.
- [174] G. Pujolle and E. Horlait. *Architecture des réseaux informatiques (tome1) : Les outils de communication*. Eyrolles, 1990.
- [175] P. Quittard. *Eléments de Storage Processus Aléatoires et Files d'Attente*. OPU, Alger, 1983.
- [176] J. Riordan. *Stochastic Service Systems*. John Wiley, New York, 1962.
- [177] B.D. Ripley. *Stochastic Simulation*. Wiley, New York, 1982.
- [178] Ph. Robert. *Réseaux et Files d'attente : Méthodes Probabilistes*. Springer-Verlag, Berlin Heidelberg, New York, 2000.
- [179] T. Rolski. *Stationary Random Processes Associated with Point Processes Lecture Notes in Statistics*. Springer-Verlag, New York, 1981.
- [180] Roseaux. *Exercice et Problèmes Résolus de Recherche Opérationnelle*. Masson, Paris, 1987.
- [181] S.M. Ross. *Applied Probability Models with Optimization Applications*. Holden-Day, San-Francisco, 1970.
- [182] R.Y. Rubinstein. *Monte-Carlo Optimization, Simulation and Sensitivity of Queueing Methods*. John Wiley, New York, 1986.
- [183] E. Ruiz-Pala, C. Avial-Beloso, and W.W. Hines. *Waiting Lines Models*. Reinhold, New York, 1967.
- [184] T.L. Saaty. *Elements of Queueing Theory*. Mc Graw- Hill, New York, 1961.
- [185] T.L. Saaty. *Mathematical Methods of Operations Research*. Mac Graw-Hill, Reinhold, New York, 1969.
- [186] G.V. Sarma and M. Babes. Out-patient queues at the ibn-rochd health center. *J. Op. Res. Soc.*, 42(10) :845–855, 1991.
- [187] C. H. Sauer and K.M. Chand. *Computer Systems Performance Modelling*. Prentice-Hall Englewood Cliffs, 1981.
- [188] R. Schassberger. *Wartechlangen*. Springer-Verlag, Vienna, 1983.
- [189] M.A. Schnepps. *Numerical Methods of Télétrafic Theory*. Moscow, Sviaz, 1974.
- [190] M. Schwartz. *Computer Communication Networks, Design and Analysis*. Prentice-Hall Englewood Cliffs, New Jersey, 1977.
- [191] M. Schwartz. *Telecommunication Network*. Prentice-Hall Englewood Cliffs, New Jersey, 1987.
- [192] L.P. Seelen, H.C. Tijms, and M.H. Van Hoorn. *Tables for Multi-Server Queueing*. Elsevier/North-Holland, Amsterdam, 1985.
- [193] G.S. Shalder. *Regeneration and Networks of Queues*. Springer-Verlag, New York, 1987.

- [193] G.S. Shalder. *Regeneration and Networks of Queues*. Springer-Verlag, New York, 1987.
- [194] Shanthikumar. *Stochastic Orders and Their Applications*. Lennigrad, 1973.
- [195] A. Shiriaev. *Probabilité*. Moscow Nauka (Wiley, New York 1988), 1980.
- [196] W.L. Smith and W.E. Wilkinson. *Proc. Symp. on Congestion Theory*. Ed. University of North Press, Chapel Hill, 1965.
- [197] S. Solomon. *Simulation of Waiting-Line Systems*. Englewood-Cliffs, 1983.
- [198] A.D. Soloviev. *Computation and Estimation of Reliability Coefficients*. Moscow, Znanié, 1978.
- [199] H.M. Srivastava and B.R. Kashyap. *Special Functions in Queueing Theory and Related Stochastic Processes*. Academic Press, New York, 1982.
- [200] W. Stallings. *Local Networks : An Introduction*. Macmillan, New York, 1984.
- [201] S.N. Stepanov. *Numerical Methods of Calculation for Systems with Repeated Calls*. Moscow, Nauka, 1983.
- [202] H.E. Stormer, E. Behlendorff, N. Bininda, G. Bretschneider, E. Hoffman, and H. Suchaland. *Teletraffic Theory*. In German (Oldenbourg Munchen), 1966.
- [203] D. Stoyan. *Comparison Methods of Queues and Other Stochastic Models*. Wileys, New York, 1983.
- [204] D.C. Sylvestrov. *Semi-Markovian Discrete Processes*. Sov. Radio, Moscow (Biblioth. of Ingen. on Reliability), 1980.
- [205] R. Syski. *Introduction to Congestion Theory in Telephone Systems*. Oliver & Boyd (Ed. Elesvier/North Holland. Amesterdam 1986)., 1960.
- [206] J. Taghem, J. Loris-Teghem, and J.P. Lambotte. *Modèles d'Attente M/G/1 à Arrivées et Services en Groupes*. Springer-Verlag, Berlin, 1969.
- [207] S. Taghzouti, L. Bouallouche, L. Mekaouche, and D. Aïssani. Degré de multiprogrammation optimal et evaluation de performances d'un serveur central à deux processeurs. *Proceedings du VIII-ème Atelier d'Evaluation des Performances (J.M. Fourneau et P. Moreaux Ed .), Reims (France)*, pages 07 – 08, Mai 2003.
- [208] H.A. Taha. *Introduction to Operations Research*. Wiley, 1982.
- [209] I. Takacs. *Investigation of Waiting Time Problems by Reduction to Markov Processes*. Acta Math. Acad. Scie. Hungar.6, 101-129, 1955.
- [210] I. Takacs. *Introduction to the Theory of Queues*. Oxford University Press, New York, 1962.
- [211] I. Takacs. *Combinatorial Methods in the Theory of Stochastic Processes*. Wiley, New York, 1967.
- [212] H. Takagi. *Analysis of Polling Systems*. The MIT Press Cambridge, 1986.
- [213] A. Tannenbaum. *Computer Networks*. Prentice-Hall Englewood Cliffs, 1981.

- [216] E. Van Doorn. *Stochastic Monotonicity and Queueing Applications of Birth-Death Processes Lecture Notes in Statistics 4*. Springer-Verlag, New York, 1980.
- [217] E.S. Ventsel and L.A. Ovtcharov. *Problèmes Appliqués de Théorie des Probabilités*. Ed. Radio & Sviaz, Moscou, 1983.
- [218] H.M. Wagner. *Principles of Operations Research*. Prentice-Hall Englewood Cliffs, New Jersey, 1975.
- [219] J. Walrand. *Introduction of Queueing Networks*. Prentice-Hall Englewood Cliffs, 1987.
- [220] J. Walrand. *Introduction to Queueing Networks*. Prentice-Hall, 1989.
- [221] J. Walrand and P. Varaiya. *Hign-Performance Communication Networks*. The Morgan Kaufmann Series in Networking, 1996.
- [222] J.A. White, J.W. Schmidt, and G.K. Bennett. *Analysis of Queueing Systems*. Academic Press, New York, 1975.
- [223] P. White. *Systems in Stochastic Equilibrium*. John Wiley, Chichester, 1986.
- [224] M.E. Woodward. *Communication and Computer Networks, Modelling with discrete-time queues*. IEEE Computer Society Press, 1994.
- [225] Y. Zaitchenko. *Recherche Opérationnelle*. Vishaia Schkola, Kiev, 1975.
- [226] G.P. Zakharov. *Methods for Communication System's Studies*. Moscow, Radio & Sviaz, 1982.

10 Annexe : Sommaire de l'ouvrage MMAPA

Méthodes Mathématiques d'Analyse des Phénomènes d'Attente
(Sommaire)

Table of Contents

1	GENERALITES	40
1.1	Introduction et bref historique	40
1.2	Classification des modèles	40
1.3	Systèmes de notation	40
1.4	Analyse des performances du système	40
1.5	Optimisation des performances	40
1.6	Identification du modèle	40
2	METHODES DE DESCRIPTION DES ELEMENTS D'UN MODELE DE FILES D'ATTENTE	41
2.1	La distribution exponentielle et ses propriétés	41
2.2	Distribution à taux de hasard monotone	41
2.3	Le flux de Poisson et ses caractérisations	41
2.4	Eléments de théorie de renouvellement	41
2.5	Autres classes particulières de flux d'évènements	41
3	MODELES MARKOVIENS	42
3.1	Les processus Markoviens et leurs propriétés	42
3.2	Processus de naissance et de mort	42
3.3	Modèle M/M/1	42
3.4	Modèles à plusieurs serveurs M/M/m	42
3.5	Autres modèles Markoviens	42
3.6	Exemples	42
4	MODELES NON MARKOVIENS	43
4.1	Introduction	43
4.2	Méthode d'Erlang	43
4.3	Méthode de Kendall (M/G/1)	43
4.4	Etude de la période d'activité	43
4.5	Durée virtuelle d'attente	43
4.6	Méthode de la variable auxiliaire	43

4.7	Système G/M/m	43
4.8	Méthode des événements fictifs	43
5	PROCESSUS SEMI - MARKOVIENS	44
5.1	Définition	44
5.2	Noyaux semi - Markoviens	44
5.3	Chaîne de Markov induite	44
5.4	Modèles d'attente semi - Markoviens	44
5.5	Processus à évolution semi - Markovienne	44
6	METHODES D'APPROXIMATION	45
6.1	Introduction	45
6.2	Equation intégrale de Lindley	45
6.3	Résolution spectrale de l'équation de Lindley	45
6.4	Approximation en régime chargé	45
6.5	Méthode du petit paramètre	45
6.6	Bornes pour la durée d'attente	45
7	SYSTEMES AVEC RAPPELS	46
7.1	Formule de Polaczek - Khintchine Alexandrov	46
7.2	Etude de la durée d'attente	46
7.3	Etude de la période d'activité	46
7.4	Approximation de diffusion	46
7.5	Bornes	46
8	SYSTEMES AVEC PRIORITES	47
8.1	Introduction	47
8.2	Système $M_2/G_2/1$ (priorité absolue)	47
8.3	Optimisation des priorités relatives et absolues	47
8.4	Priorités mixtes avec zones d'interruption	47
9	RESEAUX DE FILES D'ATTENTE	48
9.1	Introduction	48
9.2	Réseaux Markoviens ouverts	48
9.3	Réseaux Markoviens fermés	48
9.4	Réseaux multi-classes	48
9.5	Approximation de diffusion	48
9.6	Autres approximations	48

10 SIMULATION DES FILES D'ATTENTE	49
10.1 Principe de la simulation	49
10.2 Génération de nombres aléatoires	49
10.3 Simulation d'évènements et de variables aléatoires	49
10.4 Simulation de processus aléatoires	49
10.5 Application aux phénomènes d'attente	49

BIBLIOGRAPHIE

ANNEXES

INDEX

L'ouvrage :

Il est couramment admis aujourd'hui que le fondateur de la théorie des files d'attente est le Danois Erlang E.K. (1878 – 1929), qui a collaboré de 1908 à 1922 avec la Compagnie Danoise de Téléphone pour mener des études en Télétrafic (débit de transmission des réseaux de communication téléphonique).

Dans ce fascicule, nous présentons les éléments historiques liés à l'apparition, puis au développement de la Théorie des Files d'Attente. Ces fondements permettent de comprendre l'importance de cette théorie pour l'évaluation des performances des systèmes informatiques.

Les Auteurs :

* Djamil Aïssani a assuré les premiers cours de Files d'Attente et de Sûreté de Fonctionnement (Fiabilité) pour la première promotion des ingénieurs en informatique de l'ENITA (Ecole Nationale des Ingénieurs et Techniciens d'Algérie, Bordj – el – Bahri, Alger – 1985 et 1986). Il a été responsable des modules de Post-Graduation « *Modélisation et Simulation des Systèmes Industriels* » (P.G. Institut d'Informatique – Université d'Annaba, 1987) et « *Modélisation des Systèmes Informatiques* » (P.G. Institut d'Informatique – U.S.T.H.B. Alger, 1988).

* Amar Aïssani est l'auteur des ouvrages « *Phénomènes d'Attente dans les Systèmes Informatiques* » (ENITA Ed., 1987) et « *Modèles Stochastiques de la Théorie de Fiabilité* » (O.P.U., 1992). Il est actuellement Doyen de la Faculté de Génie Electrique et Informatique de l'U.S.T.H.B. Alger.

Lamos Editions, 2004