(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL
DE LA PROPRIÉTÉ INDUSTRIELLE

PARIS

(11) **N° de publication :** **2 740 579**
(à n'utiliser que pour les commandes de reproduction)

(21) **N° d'enregistrement national :** **96 12990**

(51) **Int Cl[6] : G 06 F 15/80**

(12) DEMANDE DE BREVET D'INVENTION A1

(22) **Date de dépôt** : 24.10.96.

(30) **Priorité** : 26.10.95 JP 27864795.

(43) **Date de la mise à disposition du public de la demande :** 30.04.97 Bulletin 97/18.

(56) **Liste des documents cités dans le rapport de recherche préliminaire :** *Ce dernier n'a pas été établi à la date de publication de la demande.*

(60) **Références à d'autres documents nationaux apparentés :**

(71) **Demandeur(s) :** *NEC CORPORATION* — JP.

(72) **Inventeur(s) :** KUBO HIDEHITO.

(73) **Titulaire(s) :**

(74) **Mandataire :** SOCIETE DE PROTECTION DES INVENTIONS.

(54) METHODE D'ORDONNANCEMENT D'UN TRAVAIL DANS UN SYSTEME INFORMATIQUE EN GRAPPE ET SON DISPOSITIF.

(57) Un système informatique et méthode d'ordonnancement d'un travail dans un système informatique en grappe ayant une pluralité de grappes et une mémoire globale stocke un travail entré dans une file d'attente de travaux allouée dans la mémoire globale, sélectionne un travail à exécuter, et exécute le travail sélectionné dans une grappe. La sélection du travail est activée par un achèvement d'un travail, une arrivée d'un travail, et un achèvement de mesure. Suite à la sélection du travail, si l'utilisation des ressources est faible, un nouveau travail est alors demandé. Cependant, si l'utilisation des ressources est élevée, un nouveau travail n'est alors pas demandé.

FR 2 740 579 - A1

METHODE D'ORDONNANCEMENT D'UN TRAVAIL DANS UN SYSTEME INFORMATIQUE EN GRAPPE ET SON DISPOSITIF

CONTEXTE DE L'INVENTION

La présente invention concerne une méthode d'ordonnancement d'un travail et son dispositif pour améliorer l'équilibre d'une charge entre les grappes respectives dans un système informatique en grappe.

Depuis peu, les processeurs parallèles sont de plus en plus utilisés pour la structure des systèmes informatiques. Même les ordinateurs non spécialisés ont une structure en grappe dans laquelle une pluralité de groupes de processeurs qui partagent une mémoire principale sont couplés à une mémoire partagée (c'est-à-dire une mémoire globale). Chacun des groupes de processeurs qui partagent la mémoire principale dans cette structure est appelé une « grappe ».

Dans un système informatique en grappe, l'équilibre d'une charge entre les grappes est nécessaire pour atteindre une performance de système satisfaisante. Pour un système de multiprocesseurs à couplage étroit, le partage d'une charge interne entre les processeurs est automatiquement maintenu à un niveau presque optimal. Ceci est dû au fait qu'une file d'attente de processus attendant un processeur est retenue dans la mémoire principale partagée et qu'un processeur inactif prend immédiatement un processus pour l'exécuter. En général, les processus individuels libèrent le processus toutes les millisecondes (ms) pour d'autres travaux, et l'opération de mise en file d'attente est répétée afin d'assurer à nouveau la productivité du système.

Cependant, dans un système informatique en grappe, en particulier dans un système de traitement par lots, déplacer un travail qui commence à être exécuté dans une grappe vers une autre grappe crée un grand temps

système. En conséquence, une unité d'affectation d'une charge à une grappe doit être un travail qui nécessite plusieurs minutes ou plusieurs dizaines de minutes de temps de traitement, afin de rendre une affectation réalisable. Plusieurs travaux ou dizaines de travaux sont exécutés sur chaque grappe simultanément. Ce groupe de travaux est la charge de travail à ce moment. La charge de travail doit être équilibrée entre les grappes respectives. Cependant, les caractéristiques des travaux individuels qui attendent d'être exécutés (par exemple, la durée d'un temps de traitement, le taux de charge des processeurs, etc.) ne sont pas connus à l'avance.

Garder l'utilisation de toutes les grappes à presque 100 % est relativement facile si la capacité de la mémoire principale est suffisante, dans la mesure où un nombre suffisamment grand de travaux peuvent être exécutés par toutes les grappes. Cependant, certains processus connectés sont souvent traités dans le même système, ou certains travaux exécutés en traitement par lots ayant une priorité de traitement sont traités simultanément avec les travaux exécutés en traitement par lots. Dans ce cas, une « politique de presque 100 % » est préjudiciable aux tâches de haute priorité. De ce fait, la méthode d'ordonnancement du travail exécuté en traitement par lots pour la grappe respective souffre d'un très sérieux et difficile problème dans un système informatique en grappe.

RESUME DE L'INVENTION

Etant donné le problème susmentionné du système classique, un objet de la présente invention est de fournir une méthode d'ordonnancement dynamique des travaux exécutés en traitement par lots et son dispositif, qui permettent aux charges se trouvant entre les grappes respectives d'être toujours

maintenues dans un état équilibré même à court terme, et à l'état équilibré d'être maintenu en moyenne proche d'une utilisation des ressources cible à long terme. Avec ces dispositions, la capacité de traitement maximale satisfaisant à une contrainte spécifiée peut être améliorée. Par ailleurs, chaque temps de traitement d'un travail peut être rendu égal. En outre, un travail de haute priorité (par exemple, un processus en ligne) peut être exécuté rapidement.

Pour résoudre le problème susmentionné, dans un système informatique ayant des grappes selon un premier aspect de la présente invention, chacune des grappes inclut au moins un processeur. Un mécanisme de mesure mesure une utilisation des grappes. Le preneur en charge d'un travail demande qu'un travail soit exécuté dans chacune des grappes. Un contrôleur de travail contrôle un travail en cours d'exécution dans chacune des grappes, et détecte l'achèvement du travail. Un contrôleur de travail demande une sélection de travail suite à l'achèvement du travail par le contrôleur de travail et suite à l'achèvement d'une mesure par le mécanisme de mesure. Un sélecteur de travail sélectionne un travail à exécuter dans chacune des grappes suite à l'une des demandes de sélection d'un travail par le contrôleur de travail et à la demande de travail du preneur en charge du travail.

Avec l'unique et non évidente structure de la présente invention, l'équilibre de la charge est toujours en moyenne maintenu d'une manière dynamique à court terme ainsi qu'à long terme. De plus, un travail de haute priorité est exécuté rapidement.

BREVE DESCRIPTION DES DESSINS

Les objets, caractéristiques et avantages de la présente invention susmentionnés et autres deviendront plus évidents par référence à la description détaillée de l'invention qui suit, prise conjointement avec les dessins d'accompagnement sur lesquels :

la figure 1 est un schéma fonctionnel montrant la configuration d'un système informatique en grappe selon un mode de réalisation de la présente invention ;

la figure 2 est un schéma fonctionnel montrant un mécanisme d'ordonnancement d'un travail dans le système informatique en grappe selon le mode de réalisation de la présente invention ;

la figure 3 est un organigramme montrant les opérations lorsqu'une mesure est terminée ;

la figure 4 est un organigramme montrant les opérations lorsqu'un travail est terminé ;

la figure 5 est un organigramme montrant les opérations permettant de déterminer si une utilisation d'une grappe est faible ou non ;

la figure 6 est un organigramme montrant les opérations lorsqu'un travail arrive ;

la figure 7 est un organigramme montrant les opérations de la sélection d'un travail pour une grappe (C_i) ;

la figure 8 est un schéma fonctionnel montrant un premier agencement du mécanisme d'ordonnancement et le système informatique en grappe ; et

la figure 9 est un schéma fonctionnel montrant un second agencement du mécanisme d'ordonnancement et le système informatique en grappe.

DESCRIPTION DETAILLEE DES MODES DE REALISATION PREFERES

Une méthode d'ordonnancement d'un travail dans un système informatique en grappe selon les modes de réalisation préférés de la présente invention sera

décrite en détail avec référence aux dessins d'accompagnement.

En référence à la figure 1, la structure d'un système informatique en grappe auquel s'applique la présente invention a une pluralité de processeurs et une mémoire principale. Par exemple, « m » processeurs P_{11} à P_{1m} partagent une mémoire principale S_1, formant ainsi une grappe C_i, où « m » est un nombre entier. Ainsi « n » grappes C_1 à C_n couplées à une mémoire globale G forment un système informatique en grappe, dans lequel « n » est un nombre entier.

La mémoire globale G est, par exemple, constituée d'une mémoire à semi-conducteurs et d'une mémoire secondaire telle qu'un ensemble de disques magnétiques, et est utilisée pour stocker les informations partagées du système entier ou l'échange d'informations entre les grappes.

Afin de permettre la communication directe entre les processeurs arbitraires, des lignes de signalisation L_{11} à L_{nm} peuvent être fournies couplées aux processeurs respectifs, et à un commutateur S qui échange les signaux sur ces lignes. Un système d'exploitation (OS) peut exister dans chacune des grappes à contrôler indépendamment. Une grappe peut être logiquement divisée en une pluralité de sections ayant chacune un autre système d'exploitation.

En référence à la figure 2, un mécanisme d'ordonnancement d'un travail dans un système informatique en grappe inclut un mécanisme de mesure 1 pour mesurer une utilisation d'une grappe, une mémoire de mesure 2 pour stocker un résultat de mesure, un contrôleur de demande 3 pour demander la sélection d'un travail, un sélecteur de travail 4 pour sélectionner un travail, une file d'attente de travaux 5 pour stocker une ou plusieurs demandes de travail, un preneur en charge d'un travail 6 pour prendre en charge un

travail, et un contrôleur de travail 7 pour contrôler un travail. Le sélecteur de travail 4 a une mémoire d'état des grappes 40 pour stocker les identificateurs de grappes qui ont de la place (espace) pour accepter et exécuter de nouveaux travaux.

Lorsqu'un travail est entré pour un traitement par lots, le travail entré est reçu par le preneur en charge du travail 6 (contenu dans une, certaines ou toutes les grappes) par l'intermédiaire d'une ligne 601, et enregistré dans la file d'attente des travaux 5 dans la mémoire globale G. Le mécanisme de mesure 1 existe dans chacune des grappes, et est activé à tous les temps prédéterminés (par exemple, « 1 » seconde) afin de mesurer l'utilisation des ressources dans la grappe après la mesure précédente. L'utilisation des ressources est, par exemple, le taux d'utilisation des processeurs, le taux d'utilisation des canaux, le taux d'utilisation d'une zone dans la mémoire principale, une fréquence de saisie des pages, une fréquence d'opérations d'entrée / de sortie, le taux d'utilisation d'une ressource de logiciel (par exemple, une table de commande) dans la grappe et similaire.

Le contrôleur de demande 3 est activé non seulement par l'exécution d'un travail mais également par l'achèvement d'une mesure, comme cela est décrit ci-après en référence aux figures 3 et 4. Ensuite, le contrôleur de demande 3 active le sélecteur de travail 4 par l'intermédiaire d'une ligne 301. Le sélecteur de travail 4 est activé même lorsqu'un nouveau travail y est enregistré par le preneur en charge d'un travail 6 par l'intermédiaire d'une ligne 602. En conséquence, un travail sera sélectionné à trois différentes synchronisations (par exemple, par l'intermédiaire de la ligne 301 activée par la ligne 101, par l'intermédiaire de la ligne 301 activée par la ligne 701, et l'intermédiaire la ligne 302).

En référence aux figures 2 et 3, le mécanisme de mesure 1 stocke un résultat mesuré C_i dans la mémoire de mesure 2 par l'intermédiaire d'une ligne 102, et informe le contrôleur de mesure 3 par l'intermédiaire d'une ligne 101. Le contrôleur de demande 3 qui a reçu la notification juge si l'utilisation de la grappe C_i est faible en fonction du résultat mesuré stocké (S31) par comparaison à un taux d'utilisation prédéterminé. Si l'utilisation est faible dans la grappe C_i, le contrôleur de demande 3 requiert une demande de sélection de travail au sélecteur de travail 4 par l'intermédiaire d'une ligne 301 afin de commencer un nouveau travail dans la grappe C_i (S32). Si l'utilisation de la grappe C_i n'est pas faible, alors la sélection d'un travail n'est pas nécessaire.

Le sélecteur de travail 4 qui reçoit la demande sélectionne un travail approprié à une grappe spécifiée dans la file d'attente des travaux 5 à partir d'une ligne 501, et informe le contrôleur de travail 7 dans la grappe spécifiée du travail sélectionné par l'intermédiaire une ligne 402. Si aucun travail approprié n'y existe, aucune notification n'est nécessaire.

Suite à la réception de la notification du travail sélectionné, le contrôleur de travail 7 extrait le travail de la file d'attente des travaux 5, commence à exécuter le travail, et contrôle l'exécution jusqu'à son achèvement. Après l'achèvement, le contrôleur de travail 7 informe le contrôleur de demande 3 de l'achèvement de l'exécution par l'intermédiaire d'une ligne 701.

En référence aux figures 2 et 4, le contrôleur de demande 3, qui a reçu la notification du contrôleur de travail 7, juge si l'utilisation des ressources de la grappe qui achève l'exécution du travail est élevée, en

fonction du résultat mesuré stocké dans la mémoire de mesure 2 (S33).

Si l'utilisation n'est pas élevée (par exemple, par comparaison à un taux d'utilisation prédéterminé établi par l'opérateur ou le concepteur), une demande est transmise à partir du contrôleur de demande 3 jusqu'au sélecteur de travail 4 par l'intermédiaire d'une ligne 301 afin de commencer l'ordonnancement d'un nouveau travail pour la grappe.

Si l'utilisation est élevée, la sélection du travail n'est pas requise. Dans la mesure où le nombre de travaux qui sont exécutés simultanément dans la grappe est déterminé sur la base de l'utilisation des ressources, la limite supérieure du nombre de travaux simultanément exécutés (ce que l'on appelle un « nombre initiateur ») doit être déterminée à une valeur légèrement plus grande que le degré pour lequel aucune limite n'est réellement donnée.

Afin de déterminer si l'utilisation des ressources est élevée, des valeurs de premier et second seuils sont fournies. De préférence, la valeur du premier seuil est déterminée à une valeur proche du taux d'utilisation des ressources cible et la valeur du second seuil est déterminée à une valeur plus importante que la valeur du premier seuil. Si la valeur mesurée est supérieure à la valeur du second seuil, alors l'utilisation des ressources est jugée trop élevée, et dans le cas contraire, l'utilisation est alors jugée n'être pas élevée.

Par ailleurs, afin de juger si l'utilisation des ressources d'une grappe est faible, une valeur de troisième seuil est fournie ainsi qu'une valeur de premier seuil. De préférence, la valeur du troisième seuil est déterminée à une valeur d'environ 80 % de la valeur du premier seuil. Bien sûr, cette valeur dépend des spécifications du concepteur. Une variable C est

prévue qui compte le nombre de fois que l'utilisation des ressources de la grappe est continuellement plus faible que la valeur du premier seuil, et a une limite supérieure N.

En référence à la figure 5, lorsque la valeur mesurée est inférieure à la valeur du troisième seuil (« OUI » dans S311), le processus continue à faire progresser S315 dans lequel C est déterminée à « 0 » et ensuite il est inconditionnellement jugé que l'utilisation de la grappe est faible (S316). Même lorsque la valeur mesurée n'est pas inférieure à la valeur du troisième seuil (« NON » dans S311), si la valeur mesurée est continuellement (par exemple, consécutivement) plus faible que la valeur du premier seuil par N fois (déterminée par les étapes S312, S313 et S314), il est jugé que l'utilisation de la grappe est faible (S316). Dans d'autres cas (par exemple, un « NON » dans l'étape S312 et détermination de C à « 0 » dans l'étape S317), il est jugé que l'utilisation n'est pas faible (S318).

Lorsque la grappe est jugée avoir une utilisation faible, la variable C de la grappe est déterminée à « 0 » (S315), remettant ainsi à « 0 » le nombre de fois où l'utilisation des ressources est continuellement plus faible que la valeur du premier seuil. La remise à zéro est nécessaire lorsque l'utilisation des ressources est plus faible que la valeur du troisième seuil car il existe une forte probabilité que l'utilisation des ressources soit successivement plus faible que la valeur du premier seuil, à moins que la remise à zéro ne soit effectuée, et une forte probabilité qu'une demande d'ordonnancement soit à nouveau émise, et que le nombre de travaux d'exécution devienne excessif.

Dans une autre méthode permettant de juger si l'utilisation des ressources est faible, l'utilisation

des ressources actuelles peut être estimée en fonction de la valeur mesurée obtenue historiquement. Pour chaque mesure de l'utilisation des ressources par le mécanisme de mesure 1, l'utilisation en cours est estimée en utilisant l'expression suivante (1). En supposant que « m » est la valeur mesurée, « e » est la valeur estimée, « T » est le temps présent, et « t » est l'intervalle de mesure,

$$e(T) = a \times m(T) + (1 - a) \times e(T - t) \quad (1)$$

où « a » est un paramètre satisfaisant à « $0 < a \leq 1$ ». De plus, la valeur initiale de e(T) est déterminée à $e(T_0) = m(T_0 + t)$.

Ce qui signifie que la somme de ce qui est obtenu en multipliant la présente valeur mesurée « m(T) » par « a » et ce qui est obtenu en multipliant la valeur précédente estimée « e(T) » par « (1 - a) » est considérée comme la valeur présente estimée « e(T) ». Cette expression est développée comme suit :

$$e(T) = a \times m(T) + a(1 - a) \times m(T - t) + a(1 - a)^2 \times m(T - 2t) + a(1 - a)^3 \times m(T - 3t) + \ldots\ldots$$

Cette expression montre que la valeur estimée tient compte de toutes les valeurs passées mesurées afin que les valeurs passées mesurées soient aussi importantes que les nouvelles valeurs. Dans la mesure où « a » est grand (presque « 1 »), le degré consistant à rendre la valeur la plus récente importante devient plus élevé. Pour juger si l'utilisation des ressources est faible ou non en utilisant la valeur estimée, la valeur estimée est simplement comparée à la valeur du premier seuil. Si elle est plus faible que la valeur du premier seuil, on juge alors que l'utilisation est faible. Lorsque la valeur estimée est utilisée, il est préférable que la valeur du troisième seuil soit également utilisée avec elle (par exemple, en prenant en considération le premier seuil).

En référence à la figure 6, lorsque le preneur en charge d'un travail 6 informe le sélecteur de travail 4 de l'enregistrement d'un nouveau travail (par exemple, une arrivée d'un travail), le sélecteur de travail 4 sélectionne un travail pour la grappe C_i si la mémoire d'état 40 de la grappe C_i indique que C_i est dans un état non complet (S41).

Ensuite, le sélecteur de travail 4 sélectionne un travail optimal sur la base, par exemple, de l'utilisation des ressources de la grappe, la priorité et la catégorie des travaux respectifs en cours d'exécution et en attente d'exécution, les contraintes imposées au système entier, les contraintes imposées à une grappe spécifiée et chacune des catégories de travail (S42).

En référence à la figure 7, le sélecteur de travail 4 est activé par une demande de sélection de travail sur la ligne 301. Ensuite, le sélecteur de travail 4 sélectionne un travail approprié à la grappe spécifiée C_i (S421). S'il n'existe pas de travail approprié (S422), un état de la grappe C_i est stocké dans la mémoire d'état de la grappe 40 en tant que grappe non complète (S425). Si un travail approprié existe, le sélecteur de travail 4 informe alors le contrôleur de travail 7 de la grappe C_i du travail sélectionné (S423). Si la grappe C_i a été enregistrée dans la mémoire d'état de la grappe 40 en tant que grappe non complète, l'enregistrement est alors remis à l'état initial (S424).

Ensuite, une correspondance entre un système informatique montré sur la figure 1 et un mécanisme d'ordonnancement montré sur la figure 2 est décrite. Dans la mesure où le mécanisme de mesure 1 mesure les grappes respectives, et que le contrôleur de travail 7 gère minutieusement l'exécution sur les grappes respectives, le mécanisme de mesure 1 et le contrôleur

de travail 7 sont nécessairement fournis dans les grappes C_i à C_n.

La file d'attente des travaux 5 doit être incluse dans la mémoire globale G pour la flexibilité de la sélection du travail. Le preneur en charge du travail 6 peut exister dans toutes les grappes ou seulement dans certaines grappes. Le sélecteur de travail 4 peut être situé dans une seule grappe C_i pour un contrôle centralisé, ou peut être situé dans les grappes respectives pour un contrôle réparti. Ainsi, il y a deux manières de positionner le sélecteur de travail 4 en fonction du type de contrôle désiré. Le contrôle centralisé et le contrôle réparti ont des avantages / inconvénients respectifs, et doivent être choisis en fonction de l'exigence du système ou de l'objectif du système.

En référence à la figure 8, un mode de réalisation dans lequel les sélecteurs de travail 4 sont répartis dans les grappes respectives C_1 à C_n sera décrit. Dans cet exemple, le mécanisme de mesure 1, la mémoire de mesure 2, le contrôleur de demande 3, le sélecteur de travail 4 et le contrôleur de travail 7 ayant chacun la même fonction sont répartis dans toutes les grappes.

Le travail entré est reçu par un preneur en charge du travail 6i prévu dans une grappe spécifique ou dans toutes les grappes, et est ensuite enregistré dans une file d'attente de travaux 5 de la mémoire globale G. Le mécanisme de mesure 1_j mesure l'utilisation des ressources dans la grappe C_j à chaque période de temps prédéterminée (par exemple, à chaque « 1 » seconde »), stocke le résultat de la mesure dans la mémoire de mesure 2_j, et informe le contrôleur de demande 3_j de l'achèvement de la mesure.

Le contrôleur de demande 3_j qui reçoit la notification, juge si l'utilisation de la grappe C_j est faible, sur la base du résultat de la mesure stocké

dans la mémoire de mesure 2_j. Si l'utilisation est faible, la demande de sélection de travail est transmise au sélecteur de travail 4_j afin que la grappe C_j commence un nouveau travail.

Le mécanisme de sélection du travail 4_j est initialisé même lorsqu'un nouveau travail est enregistré par un preneur en charge de travail arbitraire 6_i. Le preneur en charge du travail 6_i informe toutes les grappes C_1 à C_n par l'intermédiaire d'une ligne de signalisation L jusqu'au commutateur X et par l'intermédiaire d'une autre ligne de signalisation L traversant le commutateur X jusqu'aux autres grappes du travail. S'il n'existe pas de commutateur X ni de ligne de signalisation L (par exemple si aucun commutateur et aucune ligne de signalisation ne sont disponibles), le preneur en charge du travail 6_i inscrit la notification dans la file d'attente des travaux 5, et ensuite le sélecteur de travail 4_j effectue périodiquement une recherche dans la file d'attente des travaux 5. Le mécanisme de sélection d'un travail 4_j, qui reçoit la demande, sélectionne le travail approprié à la grappe c_j dans la file d'attente des travaux 5, et informe le contrôleur de travail 7_j du travail sélectionné.

Le contrôleur de travail 7_j extrait le travail notifié de la file d'attente des travaux 5, commence l'exécution du travail, et gère l'exécution du travail. Dès son achèvement, le contrôleur de travail 7_j informe le contrôleur de demande 3_j de la fin de l'exécution. Le contrôleur de demande 3_j, qui reçoit la notification, juge si l'utilisation des ressources de la grappe C_j qui accomplit l'exécution de travail est élevée, sur la base du résultat de la mesure stocké dans la mémoire de mesure 2_j. Si l'utilisation des ressources n'est pas élevée, le contrôleur de demande 3_j transmet une demande au sélecteur de travail 4_j afin

qu'un nouveau travail soit ordonnancé pour la grappe C_j. Le mécanisme de sélection du travail 4_j juge sur la base des circonstances à l'intérieur de la grappe C_j et sélectionne le travail dans la file d'attente des travaux 5.

La structure susmentionnée empêche qu'une grappe spécifiée reçoive des charges excessives pour la sélection du travail et permet que les grappes reçoivent des travaux adaptés d'une manière optimale à la grappe spécifiée. Ainsi, le système maintient sa performance en évitant de tels encombrements de sélection des travaux. Cependant, dans la structure susmentionnée, bien que les travaux attendant dans le système entier puissent être surveillés, il est difficile d'obtenir des informations concernant l'exécution individuelle des travaux des autres grappes. Ce problème est surmonté avec la structure de la figure 9.

En référence à la figure 9, un mode de réalisation dans lequel les sélecteurs de travail 4 sont situés uniquement dans la grappe C_i sera décrit. Dans cet exemple, les lignes de signalisation L_{11} à L_{nn} entre les processeurs et le commutateur X sont essentielles. Sont décrits ci-après la grappe spécifiée C_i ayant le sélecteur de travail 4 et une autre grappe C_j sans le sélecteur de travail 4.

Le travail entré est reçu par le preneur en charge de travail 6_k installé dans toutes les grappes ou dans une grappe spécifiée C, et enregistré au centre de la file d'attente des travaux 5 de la mémoire globale G, où il attend le début du traitement. Par ailleurs, le mécanisme de mesure 1_j est présent dans chaque grappe, et est activé à chaque période de temps déterminée (par exemple « 1 » seconde) pour mesurer l'utilisation des ressources de la grappe. Le mécanisme de mesure 1_j stocke les résultats des mesures dans la mémoire de

mesure 2 allouée dans la mémoire globale G, et informe le contrôleur de demande 3, installé dans la grappe C_i de l'achèvement de la mesure par l'intermédiaire de la ligne de signalisation 101.

Le contrôleur de demande 3 qui reçoit la notification juge si l'utilisation des ressources de la grappe C_j est faible, sur la base du résultat de la mesure stocké dans la mémoire de mesure 2. Si l'utilisation est faible, la demande de sélection du travail est transmise au sélecteur de travail 4 afin que la grappe C_j commence un nouveau travail.

Le sélecteur de travail 4 est initialisé par un signal envoyé par l'intermédiaire de la ligne de signalisation L même lorsqu'un nouveau travail est enregistré par un preneur en charge de travail 6_j d'une grappe arbitraire C_j. Le sélecteur de travail 4 qui reçoit la demande sélectionne le travail le plus approprié pour la grappe spécifiée Cj dans la file d'attente des travaux 5 (par exemple, sur la base des paramètres indiqués plus haut), et informe le contrôleur de travail 7_j du travail sélectionné par l'intermédiaire de la ligne de signalisation L.

Le contrôleur de travail 7_j extrait le travail notifié de la file d'attente des travaux 5, commence l'exécution du travail, et gère l'exécution. Dès l'achèvement, le contrôleur de travail 7_j informe le contrôleur de demande 3 de la grappe C_i de l'achèvement de l'exécution par l'intermédiaire d'un signal sur la ligne de signalisation L. Le contrôleur de demande 3, qui reçoit la notification, juge si l'utilisation des ressources de la grappe C_j qui achève l'exécution du travail est élevée, sur la base du résultat de la mesure stocké dans la mémoire de mesure 2. Si l'utilisation des ressources n'est pas élevée, le contrôleur de demande 3 transmet une demande au

sélecteur de travail 4 de la grappe C_i afin qu'un nouveau travail soit ordonnancé pour la grappe C_j.

Dans la sélection de travail centralisée susmentionnée, la mémoire de mesure 2 est allouée à la mémoire globale G, et le contrôleur de demande 3 et le sélecteur de travail 4 sont situés dans la même grappe C_i.

En tant que modification structurelle, bien que le sélecteur de travail 4 puisse être situé dans la grappe spécifiée C_i, la mémoire de mesure 2 et le contrôleur de demande 3 peuvent être situés dans toutes les grappes C_j qui sont soumises à la mesure et à la demande de sélections de travaux. Dans cette méthode, la notification entre le contrôleur de demande 3_j et le sélecteur de travail 4 est acheminée par l'intermédiaire de la ligne de signalisation L.

Dans les sélections de travaux centralisées susmentionnées, la combinaison des travaux en cours d'exécution pour toutes les grappes C_1 à C_n peut toujours être reconnue, et lorsqu'une demande de sélection de travail pour une grappe spécifiée C_j est reçue du contrôleur de demande 3 ou du contrôleur de travail 7, un travail optimal peut être sélectionné d'un point de vue non seulement de l'équilibre de la charge mais également des autres travaux en cours de traitement, ainsi que des travaux en attente dans le système entier.

Dans les modes de réalisations susmentionnés, la grappe C_j est une unité d'attribution des travaux. Cependant, pour un système dans lequel une grappe est logiquement divisée en une pluralité de sections, et où les sections divisées respectives sont contrôlées par un système d'exploitation indépendant, une gamme de sections divisées contrôlées par le système d'exploitation indépendant peut être désignée comme l'unité d'attribution des travaux. La mémoire globale G

peut être utilisée comme emplacement de stockage des informations concernant l'exécution (et/ou l'attente) des travaux dans les grappes respectives.

Comme mentionné précédemment, la mesure de l'utilisation des ressources qui forme la base de jugement de l'état d'une charge, peut être déterminée par le taux d'utilisation des processeurs, le taux d'utilisation des canaux, le taux d'utilisation de la zone de la mémoire principale, une fréquence de saisie des pages, une fréquence d'opérations d'entrée / de sortie, le taux d'utilisation d'une ressource de logiciel (par exemple, une table de commande) dans la grappe et similaire. De préférence, l'utilisation des ressources qui est la plus importante (et formerait sans doute un encombrement dans le système) est mesurée, et sur la base de l'utilisation mesurée, le travail doit être ordonnancé. L'utilisation des ressources pour la grappe entière peut être mesurée au niveau d'une seule synchronisation comme un ensemble, ou peut être effectuée à différentes synchronisations, individuellement.

Dans le mode de réalisation susmentionné, trois valeurs de seuils sont introduites. Les valeurs des seuils peuvent être modifiées, si cela est souhaitable et d'une manière sélective, pour la grappe entière, en fonction du type de ressources formant la base du jugement, un objet de la répartition de la charge et la structure du système, bien que ces valeurs soient généralement les mêmes pour les grappes respectives.

Pour la structure du matériel, toutes les grappes ont le même nombre de processeurs. Cependant, le nombre de processeurs de chaque grappe peut varier. La performance des processeurs individuels et la capacité de la mémoire principale peuvent ne pas être nécessairement identiques les unes aux autres. Si une

différence existe, il est souhaitable que la valeur du seuil soit modifiée selon les facteurs susmentionnés.

Comme décrit précédemment, selon la présente invention, une méthode d'ordonnancement dynamique du travail par lots est fournie dans laquelle la charge entre les grappes est toujours maintenue dans un état équilibré à court terme et à long terme, et l'état équilibré peut être maintenu sensiblement proche de ou à un taux cible d'utilisation des ressources.

Les résultats des simulations expérimentales ont démontré qu'avec l'application de la méthode et de la structure de l'invention employant le taux d'utilisation des processeurs en tant qu'utilisation des ressources, la dispersion du taux d'utilisation des processeurs peut être diminuée de 20 à 30 % comparé à la méthode de contrôle employant seulement le nombre d'exécutions de travaux simultanées. En conséquence, le système peut se permettre de traiter une charge en ligne ou une charge de haute priorité tout le temps, et la dispersion habituelle du temps de traitement d'un travail par lots peut être réduite (par exemple de 20 à 30 %).

La description qui précède des modes de réalisations préférés de l'invention a été présentée à des fins d'illustration et de description. Elle n'a pas pour objet d'être exhaustive ni de limiter l'invention à la forme précise sous laquelle elle est décrite, et des modifications et des variantes sont possibles à la lumière des enseignements ci-dessus ou peuvent être acquises à partir de la pratique de l'invention. Le mode de réalisation a été choisi et décrit afin d'expliquer les principes de l'invention et son application pratique afin de permettre aux hommes de l'art d'utiliser l'invention dans divers modes de réalisation et avec les diverses modifications qui sont appropriées pour l'usage particulier envisagé.

REVENDICATIONS

1. Système informatique ayant des grappes, chacune desdites grappes incluant au moins un processeur, ledit système informatique comprenant :

un mécanisme de mesure (2) pour mesurer l'utilisation de chaque grappe desdites grappes ;

un preneur en charge du travail (6) pour prendre en charge un travail à exécuter dans une grappe desdites grappes ;

un contrôleur de travail (7) pour contrôler le travail exécuté dans ladite grappe desdites grappes, et détecter l'achèvement du travail ;

un contrôleur de demande (3) pour demander la sélection d'un travail après l'achèvement du travail audit contrôleur de travail selon un résultat de mesure dudit mécanisme de mesure ; et

un sélecteur de travail (4) pour sélectionner un travail à exécuter dans une grappe desdites grappes après une demande de sélection de travail dudit contrôleur de travail (7) et la prise en charge d'un travail par ledit preneur en charge du travail (6).

2. Système informatique selon la revendication 1, dans lequel ledit contrôleur de demande (3) demande en outre la sélection d'un travail audit sélecteur de travail (4) suite à l'achèvement de la mesure par ledit mécanisme de mesure (2).

3. Système informatique selon la revendication 2, dans lequel ledit contrôleur de demande (3) juge si l'utilisation d'au moins une grappe desdites grappes a un état prédéterminé sur la base dudit résultat de la mesure, et demande audit sélecteur de travail (4) d'ordonnancer un travail pour au moins une grappe

desdites grappes lorsque l'utilisation d'au moins une grappe est jugée avoir un état prédéterminé.

4. Système informatique selon la revendication 3, dans lequel ledit contrôleur de demande (3) juge en outre, après l'achèvement d'un travail, si l'utilisation d'une grappe ayant achevé un travail dans lequel le travail vient d'être terminé a un second état prédéterminé selon ledit résultat de la mesure, et demande audit sélecteur de travail (4) d'ordonnancer un travail pour ladite grappe ayant achevé un travail lorsque l'utilisation de ladite grappe ayant achevé un travail est jugée ne pas avoir un second état prédéterminé.

5. Système informatique selon la revendication 2, dans lequel le contrôleur de demande (3) juge dès l'achèvement d'un travail si l'utilisation d'une grappe ayant achevé un travail dans lequel le travail vient d'être terminé a un second état prédéterminé selon ledit résultat de la mesure, et demande audit sélecteur de travail (4) d'ordonnancer un travail pour ladite grappe ayant achevé un travail lorsque l'utilisation de ladite grappe ayant achevé un travail est jugée ne pas avoir un second état prédéterminé.

6. Système informatique selon la revendication 5, dans lequel ledit mécanisme de mesure (2) mesure l'utilisation de ladite grappe à chaque synchronisation prédéterminée, et ledit contrôleur de demande (3) juge si l'utilisation de ladite grappe a un second état prédéterminé, et demande audit sélecteur de travail (4) d'ordonnancer un travail pour ladite grappe lorsque le résultat de la mesure est continuellement inférieur à la valeur du premier seuil multipliée par un nombre de fois prédéterminé.

7. Système informatique selon la revendication 5, dans lequel ledit mécanisme de mesure (2) mesure l'utilisation de ladite grappe à chaque synchronisation

prédéterminée, ledit contrôleur de demande (3) estime une utilisation des ressources selon un résultat de mesure précédemment obtenu, et ledit contrôleur de demande (3) juge que l'utilisation de ladite grappe a un second état prédéterminé, et demande audit sélecteur de travail (4) d'ordonnancer un travail pour ladite grappe lorsque l'utilisation estimée devient inférieure à la valeur du premier seuil.

8. Système informatique selon la revendication 6, dans lequel ledit contrôleur de demande juge que l'utilisation d'au moins une première grappe a ledit état prédéterminé lorsqu'un résultat d'une mesure actuelle de ladite première grappe est supérieur à la valeur d'un second seuil qui est plus élevée que la valeur du premier seuil.

9. Système informatique selon la revendication 8, dans lequel le contrôleur de demande juge que l'utilisation d'au moins une seconde grappe a ledit second état prédéterminé lorsqu'un résultat d'une mesure actuelle de ladite seconde grappe est inférieur à la valeur d'un troisième seuil qui est plus faible que la valeur du premier seuil.

10. Système informatique selon la revendication 7, dans lequel le contrôleur de demande (3) juge que l'utilisation d'au moins une troisième grappe a ledit état prédéterminé lorsqu'un résultat d'une mesure actuelle de ladite troisième grappe est supérieur à la valeur d'un second seuil qui est plus élevée que la valeur du premier seuil.

11. Système informatique selon la revendication 1, dans lequel ledit sélecteur de travail (4) est réparti individuellement pour chacune desdites grappes une par une.

12. Système informatique selon la revendication 1, dans lequel seule une desdites grappes inclut ledit sélecteur de travail (4).

13. Méthode d'ordonnancement d'un travail dans un système informatique ayant des grappes, chacune desdites grappes incluant au moins un processeur, ladite méthode comprenant les étapes consistant à :

mesurer d'une utilisation de chaque grappe desdites grappes ;

prendre en charge d'un travail à exécuter dans une grappe desdites grappes ;

détecter l'achèvement d'un travail en cours d'exécution ;

demander la sélection d'un premier travail dès l'achèvement du travail selon un résultat de ladite étape de mesure ; et

sélectionner un travail à exécuter dans une grappe desdites grappe suite à une demande de sélection d'un travail dans ladite étape de sélection d'un premier travail et la prise en charge du travail dans ladite étape de prise en charge du travail.

14. Méthode d'ordonnancement d'un travail dans un système informatique selon la revendication 13, comprenant par ailleurs une étape consistant à :

demander la sélection d'un second travail dès l'achèvement d'une mesure sur la base d'un résultat de ladite étape de mesure.

15. Méthode d'ordonnancement d'un travail dans un système informatique selon la revendication 14, dans laquelle :

ladite étape de demande de sélection d'un premier travail inclut les étapes consistant à :

juger tout d'abord si l'utilisation d'une grappe ayant achevé un travail desdites grappes dans lequel le travail vient d'être achevé a un premier état prédéterminé selon le résultat obtenu dans ladite étape de mesure suite à l'achèvement du travail ; et

demander tout d'abord l'ordonnancement d'un travail pour ladite grappe ayant achevé un travail lorsque l'utilisation de ladite grappe ayant achevé un travail est jugée ne pas avoir un dit premier état prédéterminé dans ladite étape de premier jugement, et

ladite étape de demande de sélection d'un second travail inclut les étapes consistant à :

juger en second lieu si l'utilisation d'au moins une grappe a un second état prédéterminé selon le résultat obtenu dans ladite étape de mesure suite à l'achèvement d'une mesure ; et

demander en second lieu l'ordonnancement pour au moins une grappe desdites grappes lorsque l'utilisation de ladite grappe est jugée avoir un dit second état prédéterminé dans ladite étape de second jugement.

16. Méthode d'ordonnancement d'un travail dans un système informatique selon la revendication 14, dans laquelle :

ladite étape de demande de sélection d'un premier travail inclut les étapes consistant à :

juger si l'utilisation d'une grappe ayant achevé un travail desdites grappes dans lequel le travail vient d'être achevé a un état prédéterminé selon le résultat obtenu dans ladite étape de mesure suite à l'achèvement du travail ; et

demander l'ordonnancement d'un travail pour ladite grappe ayant achevé un travail lorsque l'utilisation de ladite grappe ayant achevé un travail est jugée ne pas avoir un dit état prédéterminé dans ladite étape de jugement.

17. Méthode d'ordonnancement d'un travail dans un système informatique selon la revendication 16, dans laquelle l'étape de mesure inclut une étape de mesure

de l'utilisation d'au moins une desdites grappes à chaque synchronisation prédéterminée, et

dans laquelle l'étape de demande de sélection d'un second travail inclut une étape permettant de juger si l'utilisation a un second état prédéterminé lorsque le résultat de la mesure est continuellement inférieur à la valeur d'un premier seuil multipliée par un nombre de fois prédéterminé.

18. Méthode d'ordonnancement d'un travail dans un système informatique selon la revendication 17, dans laquelle ladite étape de demande de sélection d'un premier travail inclut une étape permettant de juger si l'utilisation a un dit premier état prédéterminé lorsque le résultat de la mesure actuelle est supérieur à la valeur d'un second seuil qui est plus élevée que la valeur du premier seuil.

19. Méthode d'ordonnancement d'un travail dans un système informatique selon la revendication 18, danq laquelle ladite étape de demande de sélection d'un second travail inclut une étape permettant de juger si l'utilisation a un dit second état prédéterminé lorsque le résultat de la mesure actuelle d'au moins une grappe devient inférieur à la valeur d'un troisième seuil qui est plus faible que la valeur du premier seuil.

20. Méthode d'ordonnancement d'un travail dans un système informatique selon la revendication 16, dans laquelle ladite étape de demande de sélection d'un second travail inclut une étape d'estimation de l'utilisation des ressources selon un résultat de mesure obtenu précédemment et permettant de juger si l'utilisation a un second état prédéterminé lorsque l'utilisation estimée devient inférieure à la valeur d'un premier seuil.

21. Système informatique ayant des grappes, chacune desdites grappes incluant au moins un processeur, ledit système informatique comprenant :

un mécanisme de mesure (2) pour mesurer l'utilisation de chaque grappe desdites grappes ;

un preneur en charge du travail (6) pour prendre en charge un travail à exécuter dans chaque grappe desdites grappes ;

un contrôleur de travail (7) pour contrôler le travail exécuté dans chaque grappe desdites grappes, et détecter l'achèvement du travail ;

un contrôleur de demande (3) pour demander la sélection d'un travail après l'achèvement du travail audit contrôleur de travail (7) selon un résultat de mesure dudit mécanisme de mesure (2) ; et

un sélecteur de travail (4) pour sélectionner un travail à exécuter dans une grappe desdites grappes après une demande de sélection de travail dudit contrôleur de travail (7) et la prise en charge d'un travail par ledit preneur en charge du travail (6).

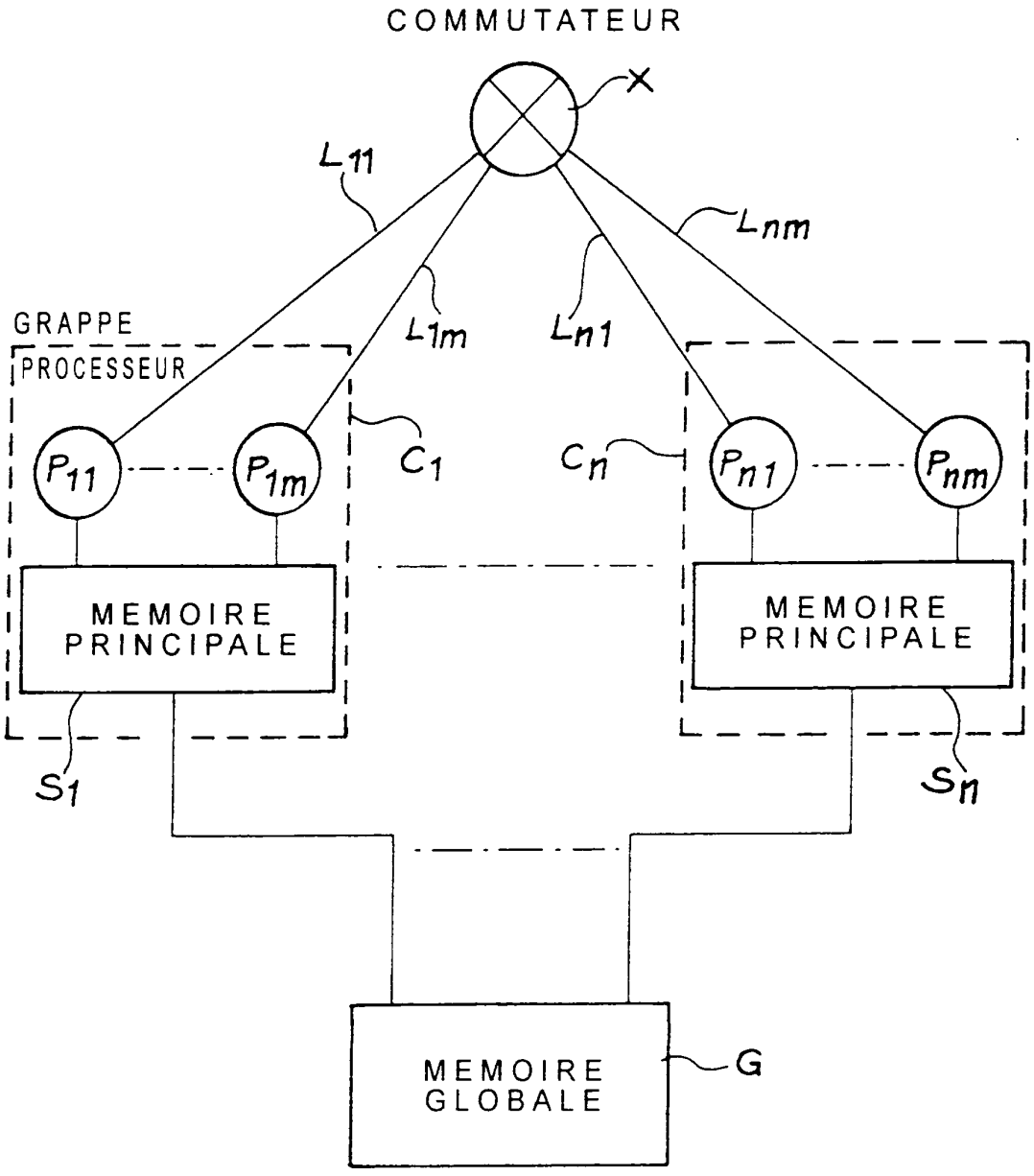

COMMUTATEUR
X
L11
Lnm
L1m
Ln1
GRAPPE
PROCESSEUR
P11
P1m
C1
Cn
Pn1
Pnm
MEMOIRE
PRINCIPALE
MEMOIRE
PRINCIPALE
S1
Sn
MEMOIRE
GLOBALE
G

FIG. 1

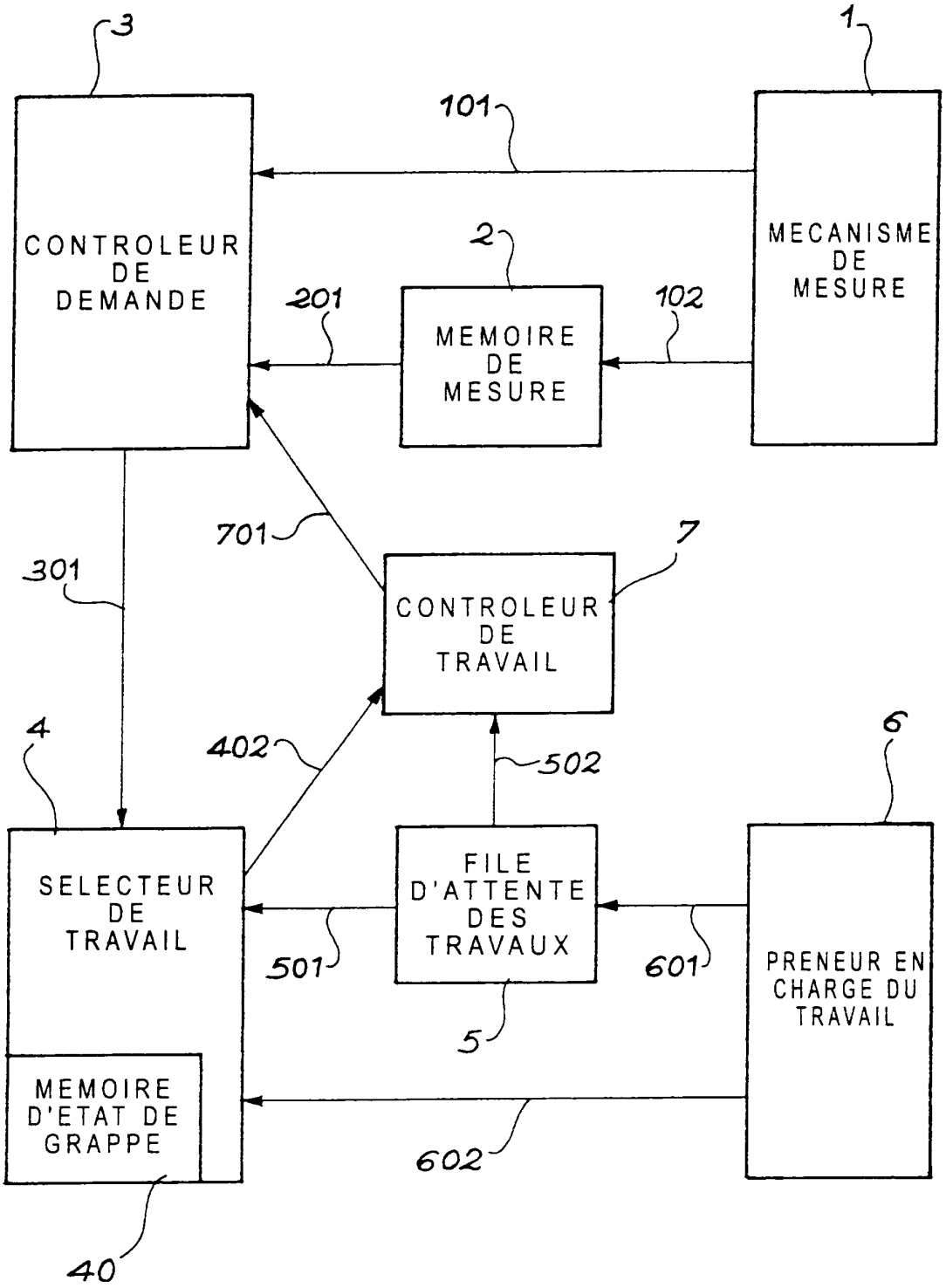

3
1
101
CONTROLEUR DE DEMANDE
MECANISME DE MESURE
2
201
102
MEMOIRE DE MESURE
701
7
301
CONTROLEUR DE TRAVAIL
4
402
502
6
SELECTEUR DE TRAVAIL
FILE D'ATTENTE DES TRAVAUX
PRENEUR EN CHARGE DU TRAVAIL
501
601
5
MEMOIRE D'ETAT DE GRAPPE
602
40

FIG. 2

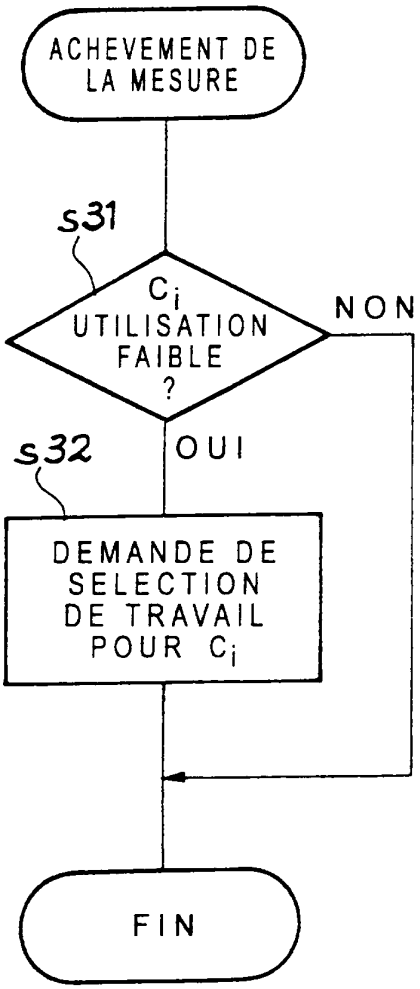

ACHEVEMENT DE LA MESURE
s31
C_i UTILISATION FAIBLE ?
NON
OUI
s32
DEMANDE DE SELECTION DE TRAVAIL POUR C_i
FIN

FIG. 3

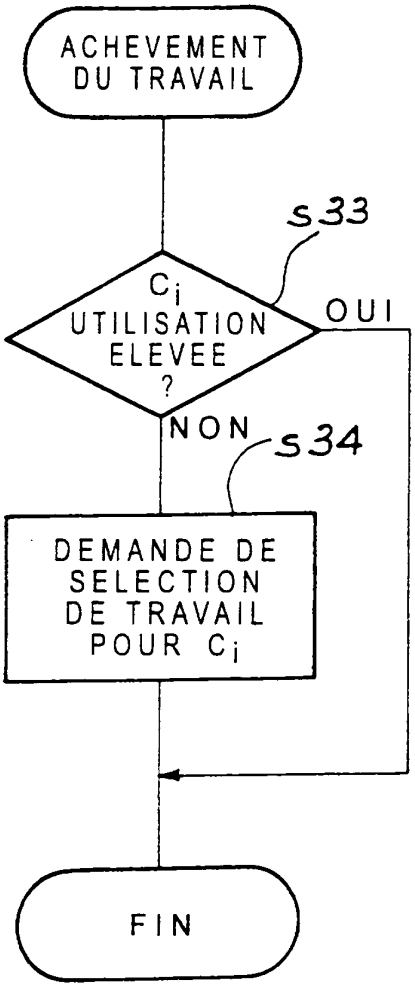

ACHEVEMENT DU TRAVAIL
s33
C_i UTILISATION ELEVEE ?
OUI
NON
s34
DEMANDE DE SELECTION DE TRAVAIL POUR C_i
FIN

FIG. 4

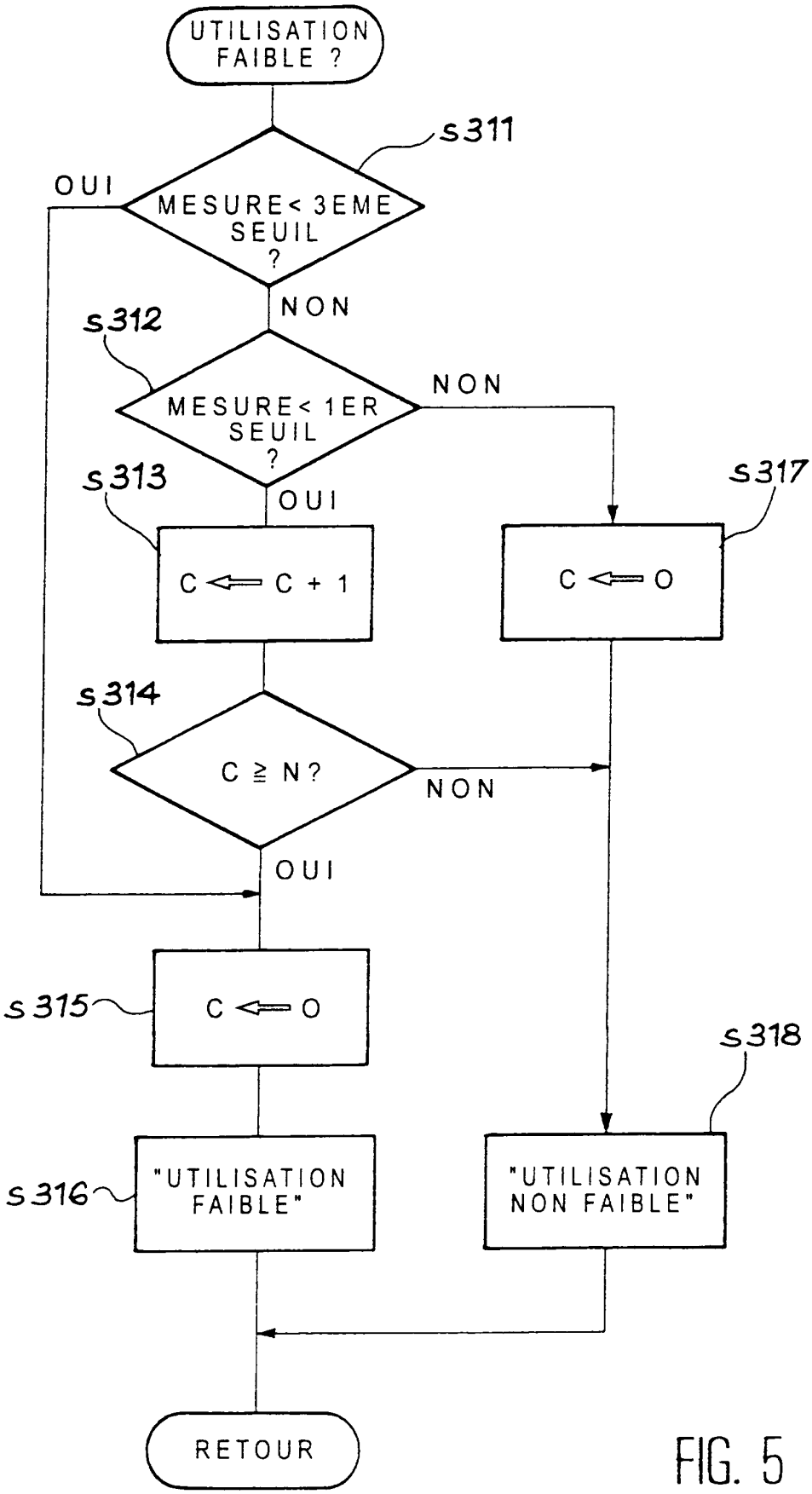
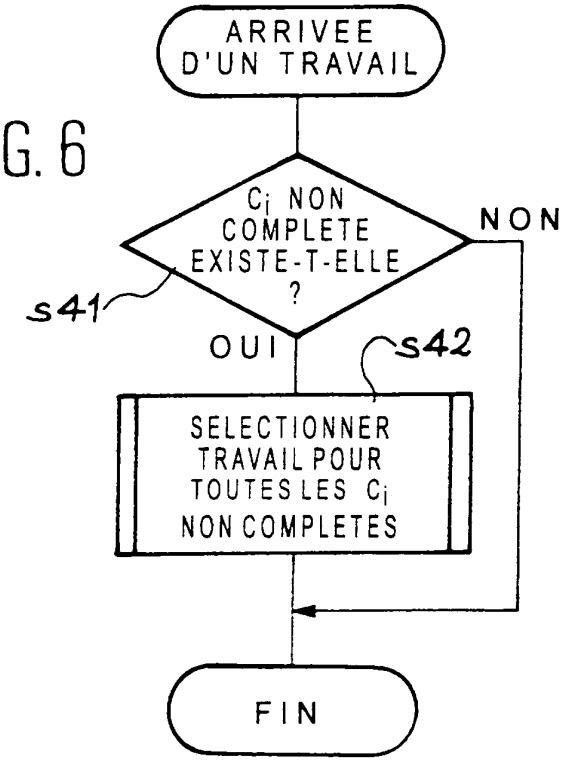

UTILISATION FAIBLE ?
s311
MESURE < 3EME SEUIL ?
OUI
NON
s312
MESURE < 1ER SEUIL ?
NON
OUI
s313
C ⇐ C + 1
s317
C ⇐ 0
s314
C ≧ N ?
NON
OUI
s315
C ⇐ 0
s318
s316
"UTILISATION FAIBLE"
"UTILISATION NON FAIBLE"
RETOUR

FIG. 5

FIG. 6


ARRIVEE D'UN TRAVAIL
Ci NON COMPLETE EXISTE-T-ELLE ?
NON
s41
OUI
s42
SELECTIONNER TRAVAIL POUR TOUTES LES Ci NON COMPLETES
FIN

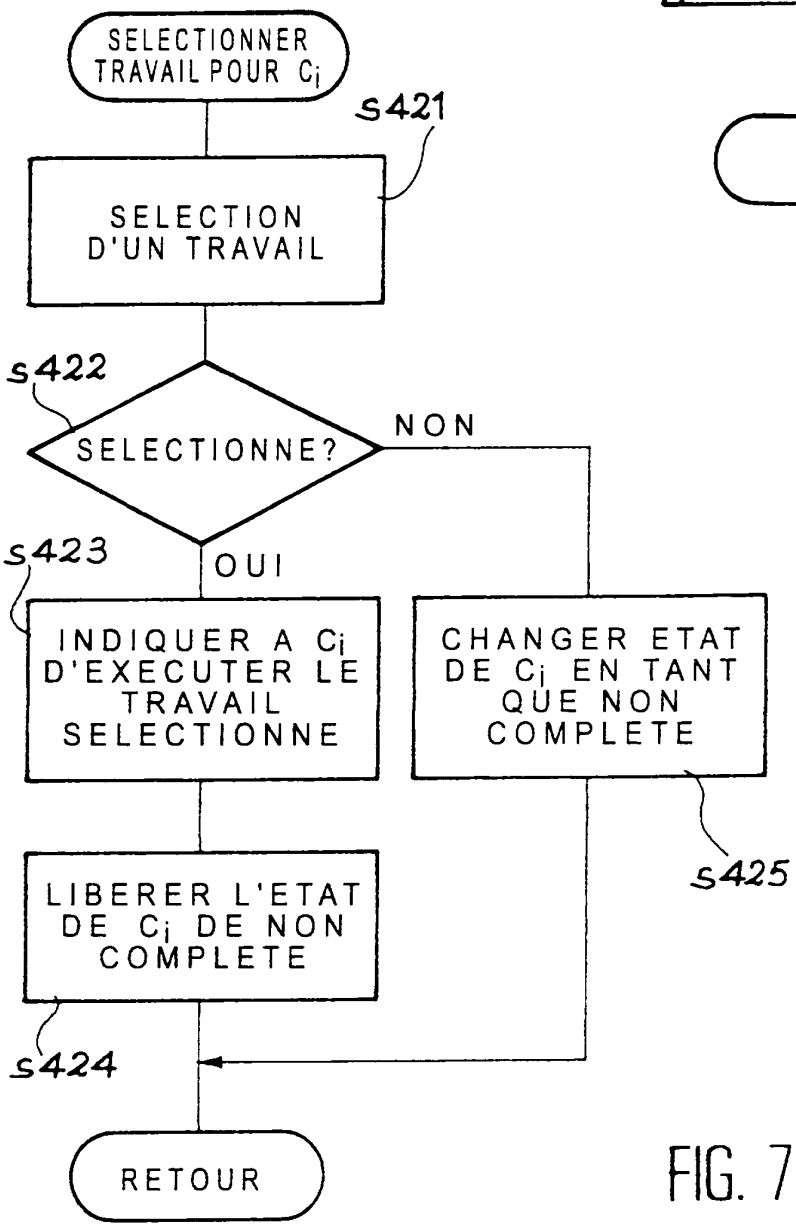

SELECTIONNER TRAVAIL POUR Ci
s421
SELECTION D'UN TRAVAIL
s422
SELECTIONNE?
NON
s423
OUI
INDIQUER A Ci D'EXECUTER LE TRAVAIL SELECTIONNE
CHANGER ETAT DE Ci EN TANT QUE NON COMPLETE
s425
LIBERER L'ETAT DE Ci DE NON COMPLETE
s424
RETOUR

FIG. 7

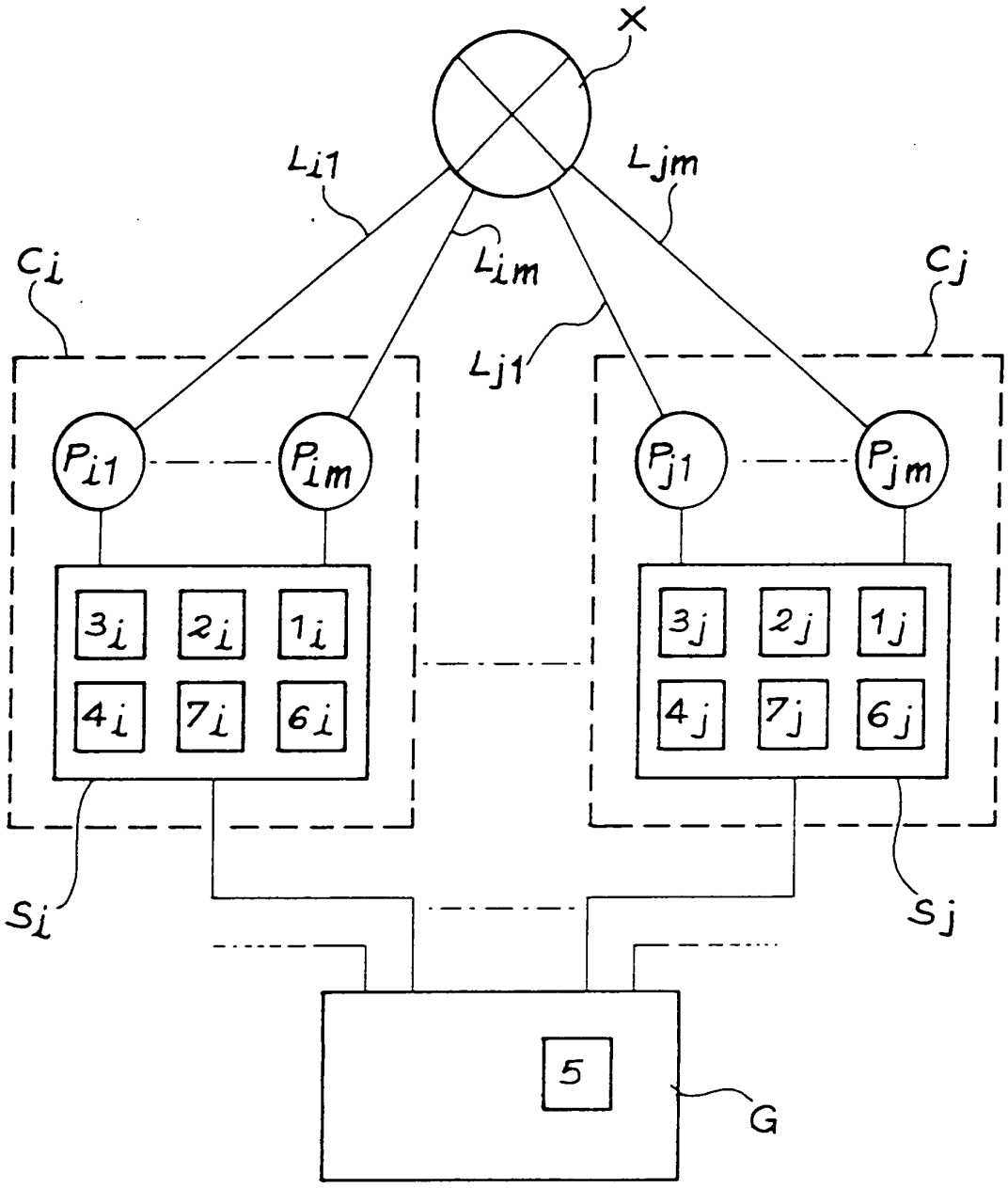

X
Li1
Ljm
Lim
Lj1
Ci
Cj
Pi1
Pim
Pj1
Pjm
3i
2i
1i
4i
7i
6i
3j
2j
1j
4j
7j
6j
Si
Sj
5
G

FIG. 8

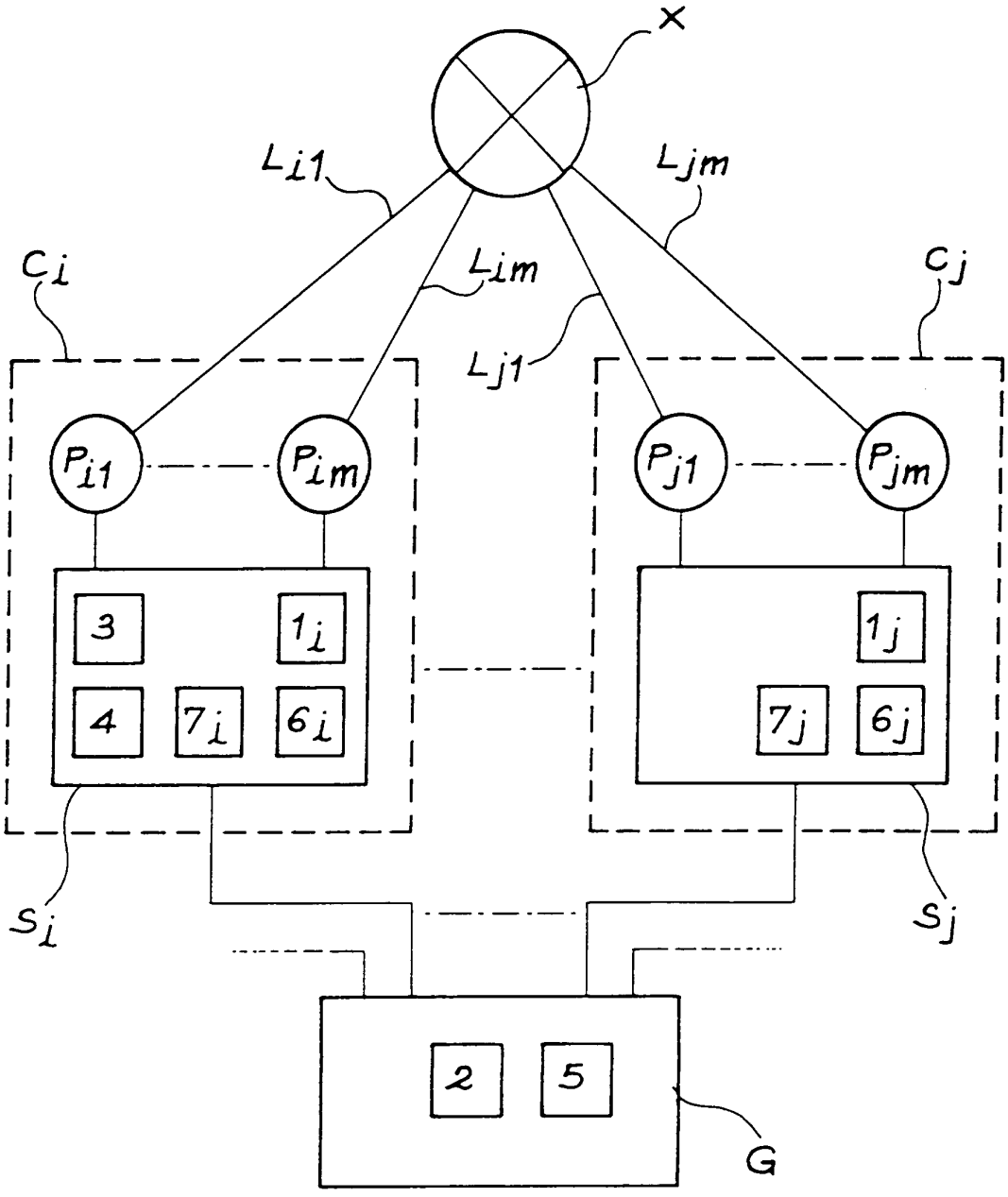

X
Li1
Ljm
Lim
Lj1
Ci
Cj
Pi1
Pim
Pj1
Pjm
3
1i
4
7i
6i
1j
7j
6j
Si
Sj
2
5
G

FIG. 9