



- (51) International Patent Classification:  
*G06F 9/48* (2006.01)
- (21) International Application Number:  
PCT/US2012/034191
- (22) International Filing Date:  
19 April 2012 (19.04.2012)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:  
13/101,626 5 May 2011 (05.05.2011) US
- (71) Applicant (for all designated States except US): **ALCATEL LUCENT** [FR/FR]; 3, avenue Octave Gréard, F-75007 Paris (FR).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **MUKHERJEE, Sarit** [US/US]; 1 Knob Hill Road, Morganville, NJ 07751 (US). **CHANG, Hyunseok** [KR/US]; 31 Fox Meadow

Lane, Holmdel, NJ 07733 (US). **LEE, Myung, Jin** [KR/US]; 3407 Cheswick Court Apartment S3, West Lafayette, IN 47906 (US). **KODIALAM, Muralidharan, Sampath** [US/US]; 5 Bella Vista Court, Marlboro, NJ 07746 (US). **LAKSHMAN, Tirunell, V.** [IN/US]; 115 Laredo Drive, Morganville, NJ 07751 (US).

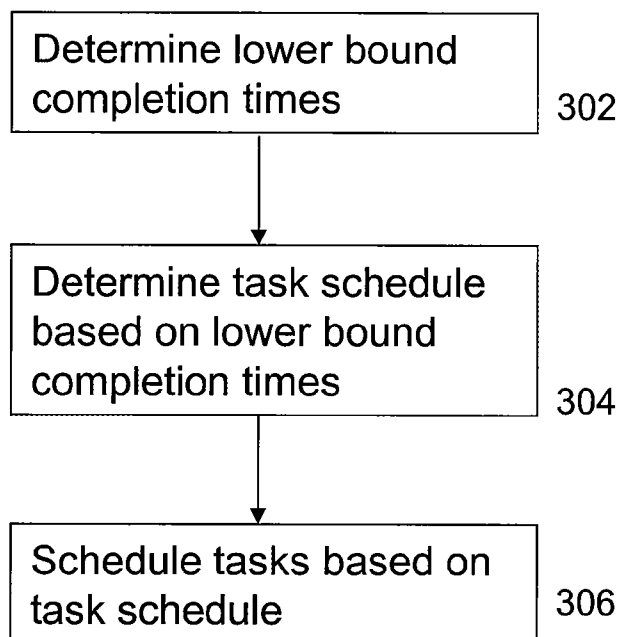
(74) Agent: **BILICSKA, Carl**; Alcatel-lucent Usa Inc., Attention: Docket Administrator - Room 3d-201, 600-700 Mountain Avenue, Murray Hill, NJ 07974-0636 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

[Continued on next page]

- (54) Title: SCHEDULING IN MAPREDUCE-LIKE SYSTEMS FOR FAST COMPLETION TIME

*Fig. 3*



(57) Abstract: A method and system for scheduling tasks is provided. A plurality of lower bound completion times is determined, using one or more computer processors and memory, for each of a plurality of jobs, each of the plurality of jobs including a respective subset plurality of tasks. A task schedule is determined for each of the plurality of processors based on the lower bound completion times.



(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published:**

- with international search report (*Art. 21(3)*)
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (*Rule 48.2(h)*)

## **SCHEDULING IN MAPREDUCE-LIKE SYSTEMS FOR FAST COMPLETION TIME**

### **FIELD OF THE INVENTION**

**[0001]** The present invention is generally directed to scheduling tasks in a computing environment and more particularly to optimized scheduling of related sets of tasks across multiple processors.

### **BACKGROUND**

**[0002]** Large-scale data processing needs are met using distributed and parallel computing data centers. MapReduce is a programming model used by these data centers for various applications such as indexing, mining, social networking, recommendation services, and advertising backends. MapReduce includes a map phase and a reduce phase. In the map phase, a dataset is partitioned into several smaller chunks that are assigned to individual nodes for partial computation of results. The results are computed at each node in the form of key-value pairs on the original data set. During the reduce phase, key-value pairs generated as a result of the map phase are aggregated based on the key-value pairs. Within a data center, a centralized master program orchestrates the assignment and scheduling of jobs, each of which includes several map and reduce tasks. The assignment and scheduling functions determine which tasks are assigned to a particular node. The master keeps track of the progress of individual tasks in order to determine where to assign and when to schedule individual tasks. Since several MapReduce jobs are often performed in parallel, scheduling problems often arise due to varying workloads and job size distributions.

### **SUMMARY**

**[0003]** In accordance with an embodiment of the invention, a method and system for scheduling tasks is provided. A plurality of lower bound completion times, using one or more computer processors and memory, is determined for each of a plurality of jobs, each of the plurality of jobs including a respective subset plurality of tasks. A task schedule is determined for each of a plurality of processors based on the lower bound completion times.

**[0004]** In an embodiment, tasks are scheduled at each processor based on the task schedule.

**[0005]** In an embodiment, determining the lower bound completion times includes solving a linear program to determine the lower bound completion times. The linear program is determined by the equation

$$Z^{LP} = \min \sum_{j \in J} w_j C_j$$

wherein  $Z^{LP}$  is a minimum weighted completion time,  $j$  is a job in a plurality of jobs  $J$ ,  $w_j$  represents a weight of each job, and  $C_j$  represents a calculated completion time of each job. Determining the lower bound completion times may be performed in response to the arrival of a new job or at a predetermined time in response to the arrival of new jobs during the predetermined time interval.

**[0006]** In an embodiment, determining the task schedule includes determining a ranked order of the plurality of jobs based on the lower bound completion time of each of the plurality of jobs.

**[0007]** In an embodiment, the lower bound completion times and task schedule are determined according to weighting factors assigned to each of the plurality of jobs.

**[0008]** These and other advantages will be apparent to those of ordinary skill in the art by reference to the following detailed description and the accompanying drawings.

## **BRIEF DESCRIPTION OF THE DRAWINGS**

**[0009]** Fig. 1 is an exemplary diagram of a computing environment in which tasks can be scheduled and assigned in accordance with an embodiment of the present invention;

**[0010]** Fig. 2 is an exemplary diagram of a task schedule in accordance with an embodiment of the present invention;

**[0011]** Fig. 3 is a flowchart of a process for scheduling tasks in accordance with an embodiment of the present invention;

**[0012]** Fig. 4 is a flowchart of an algorithm for scheduling tasks in accordance with an embodiment of the present invention; and

**[0013]** Fig. 5 is a high level diagram of a computer which may be used to implement certain embodiments of the present invention described herein.

#### **DETAILED DESCRIPTION**

**[0014]** Fig. 1 is an exemplary diagram of a computing environment in which tasks can be scheduled and assigned in accordance with an embodiment of the present invention. Server 102 may reside at a data center 106, which is responsible for receiving and assigning jobs within a distributed computing environment to clients 104-A, 104-B, 104-C, 104-D, and 104-E. Each job includes tasks, such as map and reduce tasks. Server 102 is also responsible for generating a task schedule representing an order of tasks to be performed by each of the clients.

**[0015]** Server 102 may receive groups of jobs in bulk or receive individual jobs one at a time. As jobs arrive at server 102, the jobs are processed such that all tasks associated with each job are assigned to clients 104. The tasks may be assigned based on several factors, such as a current workload of a particular client or based on the capabilities of a particular client. After all tasks are assigned to clients 104, server 102 determines an optimal scheduling of the tasks with each client in order to minimize total job completion times across all of the clients 104. The completion time of a job is defined as the time needed for all tasks belonging to a respective job to be performed by clients 104.

**[0016]** In order to determine the optimal scheduling of tasks for each client, server 102 determines a lower bound completion time for each job. The lower bound completion time for a job represents the shortest possible completion time for performing a particular job across all clients 104. Thus, each lower bound completion time represents an ideal completion time. The lower bound completion times are determined based on a linear program. A linear program is determined based on constraints such as the processing times of each job, release times of each job, and weights associated with each job specifying the importance of each job. The release

time of each job represents the arrival time of the job at server 102. More specific information regarding the linear program is provided in the paragraphs below.

**[0017]** The lower bound completion times may be determined each time server 102 receives a new job or alternatively at predetermined time intervals as long as new jobs are received by server 102 during the predetermined time interval.

**[0018]** After the lower bound completion time for each job has been determined, the lower bound completion times are used to determine a task schedule for any clients that have been assigned tasks associated with the jobs. The task schedule is determined by ranking all jobs based on their lower bound completion times. Typically, to ensure that all jobs are performed in the shortest time, the jobs are ranked in order of shortest lower bound completion time to longest lower bound completion time. The actual ranking of jobs is performed using an approximation algorithm that ensures that the actual summed weighted completion time of performing the jobs is less than or equal to a factor of the summed weighted lower bound completion time of jobs. More detail regarding the approximation algorithm is provided in the paragraphs below.

**[0019]** The task schedule is then used by server 102 to schedule tasks according to the order of jobs in the task schedule at each client 104A, 104-B, 104-C, 104-D, and 104-E. It should be further understood that while the embodiment described herein is with respect to a particular number of clients, any number of clients may be utilized. It should be understood that while described with respect to a central server 102 and multiple separate clients 104, the present invention can be adapted to a multi-processor computer or combination of a distributed environment with multi-processor computers such that server 102 may represent a master processor of a computer and each client 104 may represent a processor in a parallel computing environment.

**[0020]** Fig. 2 is an exemplary diagram of a task schedule in accordance with an embodiment of the present invention. An exemplary task schedule 202 is shown including a listing of jobs 204 and corresponding lower bound completion times 206, and a ranking 208. As shown by task schedule 202, the jobs are ranked based on shortest lower bound completion time to longest lower bound completion time. For example, job 25 has lower bound completion time of '5' and thus the highest task

schedule ranking of '1.' In contrast, job 1 has the longest corresponding lower bound completion time of '19' and has a task schedule ranking of '142.'

**[0021]** As mentioned above, for each client 104, server 102 will schedule tasks based on the task schedule. Task scheduling is based on task schedule 202. For example, if client 104-A is assigned certain tasks corresponding to jobs 25, 17, and 142, according to the task schedule, client 104-A will perform the tasks according to the job schedule ranking. This means that all tasks corresponding to job 25 are performed first, followed by the tasks corresponding to job 17, and then the tasks corresponding to job 142. Tasks will be performed in a similar order if, for example, client 104-B is assigned tasks corresponding to jobs 19, 2, and 8. Tasks corresponding to job 19 will be performed first, followed by tasks corresponding to job 2, and then tasks corresponding to job 8. It is understood that task schedule 202 is exemplary and may include more than or less than the number of jobs shown. It is also understood that each client 104 may be assigned tasks corresponding to any number of jobs.

**[0022]** Fig. 3 is a flowchart of a process for scheduling tasks in accordance with an embodiment of the present invention. Server 102 is configured to schedule a plurality of tasks, wherein each task is assigned to one of a plurality of clients 104. At step 302, a plurality of lower bound completion times for each of a plurality of jobs is determined. Each of the plurality of jobs includes a respective subset plurality of tasks. The determination of the plurality of lower bounds may be performed either in response to server 102 receiving at least one new job including tasks that need to be assigned to clients 104, or in response to receiving at least one new job during a predetermined time interval. Determining the plurality of lower bounds includes solving a linear program to determine the lower bound completion times.

**[0023]** At step 304, a task schedule for each of the plurality of clients is determined based on the lower bound completion times. Specifically, a ranked order of the plurality of jobs is determined using the lower bound completion times of each of the plurality of jobs. The ranked order is used as the task schedule. At step 306, tasks that have been assigned to each client are scheduled based on the task schedule.

**[0024]** In order to fully discuss the aforementioned linear program and approximation algorithm, the theoretical framework for scheduling jobs and tasks to clients or processors follows. If a server is responsible for  $m$  clients or a computer with  $m$  processors with  $n$  jobs assigned to the  $m$  processors,  $M$  may be used to represent the set of processors and  $J$  to represent the set of jobs that need to be processed.

Each job  $j \in J$  includes of  $n_j$  tasks each of which is assigned to a processor.

$M_j \subseteq M$  represents the set of processors required to process job  $j$ . It is assumed that each processor can process at most one task at a time. In practice, processors may run multiple tasks simultaneously and preempt each other.

**[0025]** Each job  $j$  is assumed to be available for processing at time  $r_j$ , the release time or arrival time of job  $j$  at server 102. A weight  $w_j$  may also be applied to each job, which specifies its importance.

**[0026]** The finish time of job  $j$  on a client or processor  $p \in M_j$ , is denoted by  $f_{jp}$ , the time at which the task of job  $j$  is completed on processor  $p$ . The completion time of a job,  $C_j$  is defined as the time at which all tasks belonging to a job  $j$  are finished. Therefore the completion time of job  $j$ , is

$$C_j = \max_{p \in M_j} f_{jp},$$

where  $J_p$  denotes the set of jobs that have a task assigned to processor  $p$ . Therefore

$$j \in J_p \text{ if and only if } p \in M_j.$$

**[0027]** The objective of scheduling is to determine the ordering of tasks on each processor in order to minimize the weighted sum of the job completion times

$\sum_{j \in J} w_j C_j$ . Assuming that scheduling takes place offline, all job arrival times,  $r_j$  are known in advance. Scheduling problems, however, arise online when jobs are released over time and the job release times are not known in advance. Also, while it is assumed that processing times are known processing times are generally estimated and thus estimation errors must be taken into account.

**[0028]** An algorithmic approach is ideal for solving the problem of scheduling tasks. As discussed a linear program is derived to determine a lower bound completion time for each job. However since the lower bound completion times represent the

absolute minimum possible completion time, it is not always feasible that this completion time be achieved. Thus, the lower bound completion times are used in order to determine the resulting task schedule where the weighted completion time of completing the jobs according to the task schedule is within a factor of the weighted sum of the lower bound completion times.

**[0029]** A linear program for determining the lower bound completion times for jobs will be described. Assuming  $t_{jp}$  denotes the processing time of job  $j$  on processor  $p$ , and  $J_p$  denotes the set of jobs on processor  $p$ , the linear program (LP\_OPT\_SCHED) discussed below provides the lower bound on a minimum weighted completion time schedule.

**[0030]** The linear program is represented by:

$$Z^{LP} = \min \sum_{j \in J} w_j C_j$$

$$\sum_{j \in S} t_{jp} C_j \geq f(S, p) \quad \forall S \subseteq J_p \quad \forall p \quad (1)$$

$$C_j \geq r_j + h_j \quad \forall j \quad (2)$$

where

$$f(S, p) = \frac{1}{2} \left[ \sum_{j \in S} t_{jp}^2 + \left( \sum_{j \in S} t_{jp} \right)^2 \right] \text{ and } h_j = \max_{p \in M_j} t_{jp}.$$

**[0031]**  $Z^{LP}$  represents an optimal solution to linear program (LP\_OPT\_SCHED) representing the lower bound completion time and  $f(S, p)$  represents the finish time of all jobs  $S$  on a particular processor  $p$ .

**[0032]** A proof for the aforementioned linear program is described for completeness. Fix a processor  $p$  and consider the set of jobs  $J_p$  that are assigned to processor  $p$ . Consider a subset  $S \subset J_p$  and let  $S = \{1, 2, \dots, k\}$ . Let  $f_{jp}$  denote the finish time of job  $j$  on processor  $p \in M_j$ . Assume that the jobs in  $S$  are processed before any job in  $J_p$  in a subset  $S$ . Assume that the jobs are processed in the order 1, 2, ...  $k$ . Since a lower bound is determined,  $r_j$  is ignored, and in this case

$$f_{jp} = \sum_{i=1}^j t_{ip} \text{ and } \sum_{j \in S} t_{jp} f_{jp} = \sum_{j=1}^k t_{jp} \left( \sum_{i=1}^j t_{ip} \right) = f(S, p)$$

**[0033]** Note that  $f(S, p)$  is symmetric in  $t_{jp}$  and therefore its value is independent of the order in which the jobs in set  $S$  are processed. Moreover,  $\sum_{j \in S} t_{jp} f_{jp}$  only increases if some job in  $J_p$  in a subset  $S$  is processed along with jobs in  $S$ . This holds for any subset  $S$  of  $J_p$ . Since

$$C_j \geq f_{jp}, \sum_{j \in S} t_{jp} C_j \geq \sum_{j \in S} t_{jp} f_{jp}$$

thus giving the constraints in (1). The constraints in (2) states that the completion time of job  $j$  on cannot be less than the sum of the arrival time of the job and the maximum processing time of the job on any processor.

**[0034]** Note that in Equation (1) of the linear program, there are an exponential number of constraints for each processor. The polyhedron for each processor is a polymatroid. In this case, the job completion times lie in the intersection of  $m$  polymatroids, one belonging to each processor. The constraints in the linear program only represent necessary conditions, and therefore,  $Z^{LP}$ , the optimal solution to the linear program, only represents a lower bound of the optimal solution, which is the lower bound completion time. We illustrate this by the following example: Consider a two job, two processor problem. The processing times are  $t_{11} = 1$ ;  $t_{12} = 2$ ;  $t_{21} = 2$ ;  $t_{22} = 1$ . Let  $w_1 = w_2 = 1$  and  $r_1 = r_2 = 0$ . In this case the linear programming problem is

$$\begin{aligned} \min & C_1 + C_2 \\ & C_1 \geq 1 & 2C_1 & \geq 4 \\ & 2C_2 \geq 4 & C_2 & \geq 1 \\ & C_1 + 2C_2 \geq 7 & 2C_1 + C_2 & \geq 7 \end{aligned}$$

The three constraints on the left are for processor 1 and the three constraints on the right are for processor 2. The optimal solution to the linear program is  $C_1 = C_2 = 7/3$  giving  $Z^{LP} = 14/3 = 4.67$ . An optimal schedule is to schedule job 1 on both machines first, followed by job 2. The completion time of job 1 is 2 and the completion time of job 2 is 3 giving a total completion time of 5. Since the linear program does not give an

actual schedule, an algorithm is necessary to produce a feasible ordering of jobs on different processors, which is discussed further in the paragraphs below.

**[0035]** One approach to solving the linear program involves using an ellipsoid algorithm with a separation oracle. Though this will result in a polynomial time algorithm, it is not a practical approach due to the high level of complexity associated with polynomial time algorithms. Thus, a primal dual algorithm is better suited for solving the linear program.

**[0036]** To solve the primal dual algorithm, the dual variable  $\pi_p^S$  is associated with equation (1) of the linear program. The dual variable is then used in the equation shown:

$$\begin{aligned} \max \sum_m \sum_{S \subseteq J_p} f(S, p) \pi_p^S \\ \sum_m t_{jp} \sum_{S \subseteq J_p: S \ni j} \pi_p^S \leq w_j \quad \forall j \end{aligned}$$

$\pi_p^S$  represents dual variables and  $C_j$  represents primal variables.

**[0037]** An outline of algorithm PRIMAL\_DUAL is presented below:

- 1) Initialize  $C_j = \frac{\delta}{w_j}$  for all jobs  $j$ .
- 2) Find the processor  $\bar{p}$  and subset of jobs  $\bar{S} \subseteq J_p$  that minimizes Equation (4).
- 3) Find the job  $\bar{j} \in \bar{S}$  that minimizes  $\frac{w_j}{p_{jm}}$ .
- 4) Set  $\pi_{\bar{p}}^{\bar{S}} \leftarrow \pi_{\bar{p}}^{\bar{S}} + \frac{w_{\bar{j}}}{p_{\bar{j}\bar{m}}}$ .
- 5) Set  $C_j \leftarrow C_j + \epsilon \left( 1 + \frac{w_j/p_{jm}}{w_{\bar{j}}/p_{\bar{j}\bar{m}}} \right)$ .

**[0038]** Fig. 4 is a flowchart of an algorithm for scheduling tasks in accordance with an embodiment of the present invention. More specifically, Fig. 4 is a flowchart of the PRIMAL\_DUAL algorithm. Given any  $\epsilon > 0$ , a combinatorial approach may be used to solve the maximization linear program to within a factor of  $(1 - \epsilon)^2$  of the optimal solution.

**[0039]** The primal dual-algorithm initializes completion time  $C_j = \delta/w_j$  for all jobs  $j$  at step 402, where  $\delta$  is a parameter based on  $\epsilon$ .

**[0040]** At step 404, the processor subset combination  $(\bar{p}, \bar{S})$  that minimizes

$$\frac{\sum_{j \in S} t_{jp} C_j}{f(S, p)} \quad (4)$$

is identified for current values of  $C_j$ .

**[0041]** This processor-subset combination represents a column in the linear program. This is equivalent to picking the entering variable in a standard simplex algorithm.

**[0042]** At step 406, the blocking row, a job  $j$  from set  $S$  is found. The blocking

row is the job that minimizes  $\frac{w_j}{p_{jp}}$ .

**[0043]** At step 408,

$$\pi_{\bar{p}} \leftarrow \pi_{\bar{p}} + \frac{w_j}{p_{jm}}$$

representing the dual variables, is set.

**[0044]** At step 410, the  $C_j$  values are updated and the algorithm PRIMAL\_DUAL outlined above may be repeated. The following paragraphs provide a theorem that serves as a proof for the primal dual algorithm described.

**[0045]** Given  $\epsilon > 0$  and

$$\delta = (1 + \epsilon) \left( (1 + \epsilon)n \right)^{\frac{1}{\epsilon}}$$

the algorithm PRIMAL\_DUAL computes an optimal solution to the linear program in at most

$$n^{\frac{1}{\epsilon}} \log_{1+\epsilon} n$$

iterations, where  $n$  is the number of jobs in the system. Each iteration represents a lower bound completion time for a particular job.

**[0046]** One aspect of solving the linear program is to determine a processor-subset combination that minimizes Equation (4). Although completion time values  $C_j$  are known, there may be an exponential number of subsets to be checked. Thus, a straightforward brute force approach is not appropriate. Instead, the structure of the scheduling polyhedron can be exploited to develop a simple polynomial time algorithm to solve the problem. Thus, on any processor, at most  $n$  sets need to be checked.

**[0047]** Since jobs  $J_p$  on processor  $p$  are configured such that

$$C_1 \leq C_2 \leq \dots C_k \text{ where } |J_p| = k$$

Equation (4) is minimized by a set  $S$  of  $\{1, 2, \dots, m\}$  for  $m=2, 3, \dots, k$ . Assuming that  $S$  minimizes Equation (4) for processor  $p$ , a job  $k \in S$  is found in order to determine what happens to Equation (4) when evaluating for  $S \setminus k$ . Letting

$$g(S, p) = \sum_{i \in S} t_{ip} C_i$$

results in:

$$\frac{g(S \setminus k, p)}{f(S \setminus k, p)} = \frac{g(S, p) - t_{kp} C_k}{f(S, p) - \left( \sum_{j \in S} t_{jp} \right) t_{kp}}$$

Since  $S$  minimizes Equation (4),

$$C_k \leq \frac{g(S, p)p(S, p)}{f(S, p)}$$

Where  $P(S, p) = \sum_{j \in S} t_{jp}$ . Considering that some  $k \in J_p \setminus S$ , then

$$\frac{g(S \cup k, p)}{f(S \cup k, p)} = \frac{g(S, p) + t_{kp} C_k}{f(S, p) + t_{kp}^2 + \left( \sum_{j \in S} t_{jp} \right) t_{kp}}$$

Since  $S$  minimizes Equation (4), it can be shown that

$$C_k \geq \frac{g(S, p)p(S, p)}{f(S, p)} + \frac{t_{kp}g(S, p)}{f(S, p)}$$

**[0048]** Note that  $k \in S$  if any only if

$$C_k \leq \frac{g(S, p)p(S, p)}{f(S, p)}$$

This implies that if two jobs  $j, k \in J_p$  and  $C_j \leq C_k$ , then  $j \in S$  if  $k \in S$ .

**[0049]** Since tasks of each job can potentially be assigned to all available processors, the running time of the algorithm is  $O(nm)$ . The primal-dual algorithm for solving LP\_OPT\_SCHED is similar to the primal-dual algorithm PRIMAL\_DUAL outlined above. The only change is that there are  $n$  additional dual variables corresponding to the release time constraint for each job. These  $n$  additional variables are updated separately.

[0050] The solution of LP\_OPT\_SCHED can be used to design a 3-approximation algorithm. This algorithm is referred to as OFFA. The superscript LP is used to denote the linear program solution and OFFA to denote the approximation

algorithm. Letting  $C_j^{LP}$  denote job completion times corresponding to the optimal linear program solution, algorithm OFFA can be used to determine a task schedule representing the optimal ordering of jobs on processors.

[0051] Specifically, algorithm OFFA is described below.

Algorithm OFFA

- 1) Solve LP\_OPT\_SCHED and let  $C_j^{LP}$  denote the completion time of job  $j$ . On each processor  $p$ , the jobs in  $J_p$  are processed in the order output by the LP.
- 2) Given a job  $j \in J_p$ , let

$$P(j, p) = \{k : k \in J_p, C_k^{LP} \leq C_j^{LP}\}$$

denote the set of jobs on  $p$  whose LP completion time is lower than  $j$ . These are jobs that will be scheduled before job  $j$  on processor  $p$ .

- 3) A job is started as soon as the previous job is complete or the job becomes available whichever is later. The finish time of job  $j$  on machine  $p$  denoted by  $f_{jp}^{OFFA}$  is

$$f_{jp}^{OFFA} = \max \left\{ r_j, \max_{k \in P(j, p)} f_{kp}^{OFFA} \right\} + t_{jp}.$$

- 4) The completion time of job  $j$

$$C_j^{OFFA} = \max_{p \in M_j} f_{jp}^{OFFA}.$$

[0052]  $Z^{LP}$  and  $Z^{OFFA}$  denote the objective function value of LP\_OPT\_SCHED and the sum weighted completion time from algorithm OFFA, respectively. Thus,

$$Z^{OFFA} \leq 3 Z^{LP}.$$

[0053]  $f_{jp}^{OFFA}$  denotes the finish time of job  $j$  on processor  $p$  and

$$P(j, p) = \{k : k \in J_p, C_k^{LP} < C_j^{LP}\} \cup \{j\}.$$

$f_{jp}^{OFFA} \leq \max_{k \in P(j,p)} r_k + \sum_{k \in P(j,p)} t_{kp}$ . From Equation (2) of the linear program, note that  $r_k \leq C_k^{LP}$  and  $C_k^{LP} \leq C_j^{LP}$  from the definition of  $P(j,p)$ . Therefore,

$$f_{jp}^{OFFA} \leq C_j^{LP} + \sum_{k \in P(j,p)} t_{kp}. \quad (5)$$

**[0054]** Applying Equation (1) to the set  $S=P(j,p)$  on processor  $p$ , then

$$\sum_{k \in P(j,p)} t_{kp} C_k^{LP} \geq f(P(j,p), p) \geq \frac{1}{2} \left( \sum_{k \in P(j,p)} t_{kp} \right)^2$$

Based on the fact that  $C_j^{LP} \geq C_k^{LP}$  for all  $k \in P(j,p)$ , then

$$C_j^{LP} \sum_{k \in P(j,p)} t_{kp} \geq \frac{1}{2} \left( \sum_{k \in P(j,p)} t_{kp} \right)^2$$

Therefore,

$$\sum_{k \in P(j,p)} t_{kp} \leq 2C_j^{LP}$$

and applying these to Equation (5) results in

$$f_{jp}^{OFFA} \leq 3 C_j^{LP}$$

This result holds for all  $p \in M_j$ . Since

$$C_k^{OFFA} = \max_{p \in M_j} f_{jp}^{OFFA}$$

then

$$C_k^{OFFA} \leq 3 C_k^{LP} \quad \forall k$$

implying that the solution given by algorithm OFFA is less than three times the optimal solution.

**[0055]** When all jobs are available initially, a specific linear program LP\_BULK\_ARRIVAL is used. LP\_BULK\_ARRIVAL is represented by

$$Z_B^{LP} = \min \sum_j w_j C_j$$

$$\sum_{j \in S} t_{jp} C_j \geq f(S, p) \quad \forall S \subseteq J_p \quad \forall p$$

This linear program provides lower bound completion times and is used to generate an approximation algorithm when all jobs are available at time zero. This approximation algorithm is referred to as OFFB. OFFB operates similarly to OFFA described above. Jobs are processed on processor  $p$  in the order of their LP\_BLK\_ARRIVAL completion times. Since job arrival times are not used,

$$f_{jp}^{OFFB} = t_{jp} + \sum_{k \in P(j,p)} t_{kp}$$

**[0056]** Letting  $Z_B^{LP}$  and  $Z^{OFFB}$  denote the objective function value of LP\_BULK\_ARRIVAL and the sum weighted completion time from algorithm OFFB respectively. Then

$$Z^{OFFB} \leq 2 Z_B^{LP}$$

**[0057]** Algorithm OFFB processes jobs in the order of linear program completion times and thus

$$f_{jp}^{OFFB} = \sum_{k \in P(j,p)} t_{kp}. \quad (6)$$

**[0058]** Applying Equation (1) to set  $S=P(j,p)$  on processor  $p$ , then  $\sum_{k \in P(j,p)} t_{kp} \leq 2 C_j^{LI}$ . Applying these to Equation (6) results in  $f_{jp}^{OFFB} \leq 2 C_j^{LP}$ . This result holds for all  $p \in M_j$ . Since

$$\tilde{C}_k^{OFFB} = \max_{p \in M_j} f_{jp}^{OFFB}, \text{ then } C_k^{OFFB} \leq 2 C_k^{LP}$$

for all  $k$  implying that the solution given by algorithm OFFB is less than two times the optimal solution.

**[0059]** In practice, release times or arrival times of jobs are not always known in advance. Thus, when a new job arrives, all jobs currently in the system as well as the new job may be taken and used for LP\_BULK\_ARRIVAL. An optimal task schedule can then be determined based on the lower bound completion times. This task schedule may be used until the arrival of the next job. Alternatively, groups of arrival jobs may be gathered and LP\_BULK\_ARRIVAL run periodically.

**[0060]** The above-described methods and apparatus for scheduling tasks can be implemented on a computer using well-known computer processors, memory units,

storage devices, computer software, and other components. A high-level block diagram of such a computer is illustrated in Fig. 5. Computer 500 contains a plurality of processors 510 which controls the overall operation of computer 500 by executing computer program instructions which define such operations. The computer program instructions may be stored in a storage device 520, or other computer readable medium (e.g., magnetic disk, CD ROM, etc.), and loaded into memory 530 when execution of the computer program instructions is desired.

**[0061]** Thus, the method steps of Figs. 3 and 4 can be defined by the computer program instructions stored in the memory 530 and/or storage 540 and controlled by processors 410 executing the computer program instructions. For example, the computer program instructions can be implemented as computer executable code programmed by one skilled in the art to perform an algorithm defined by the method steps of Figs. 3 and 4. Accordingly, by executing the computer program instructions, the processors 510 execute an algorithm defined by the method steps of Figs. 3 and 4.

**[0062]** While computer 500 has been described as being used for compensating for frequency dependent imbalances in a signal in accordance with the method steps shown in Figs. 3 and 4, computer 500 may also perform functionalities related to those described above in connection with the other Figures. Computer 500 also includes one or more network interfaces 540 for communicating with other devices via a network. Computer 500 further includes input/output devices 550 that enable user interaction with the computer 500 (e.g., display, keyboard, mouse, speakers, buttons, etc.) One skilled in the art will recognize that an implementation of an actual computer could contain other components as well, and that Fig. 5 is a high level representation of some of the components of such a computer for illustrative purposes.

**[0063]** The foregoing Detailed Description is to be understood as being in every respect illustrative and exemplary, but not restrictive, and the scope of the embodiments disclosed herein is not to be determined from the Detailed Description, but rather from the claims as interpreted according to the full breadth permitted by the patent laws. It is to be understood that the embodiments shown and described herein are only illustrative of the principles of the present embodiments and that various modifications may be

implemented by those skilled in the art without departing from the scope and spirit of the embodiments described herein. Those skilled in the art could implement various other feature combinations without departing from the scope and spirit of the embodiments of the present disclosure.

## I Claim:

1. A computer-implemented method for scheduling a plurality of tasks, wherein each task is assigned to one of a plurality of processors, the method comprising:

determining, using one or more computer processors and memory, a lower bound completion time for each of a plurality of jobs, each of the plurality of jobs comprising a respective plurality of tasks; and

determining, using one or more computer processors and memory, a task schedule for each of the plurality of processors based on the lower bound completion times.

2. The method of claim 1, wherein determining the lower bound completion time comprises:

solving a linear program to determine the lower bound completion time, wherein

$$Z^{LP} = \min \sum_{j \in J} w_j C_j$$

the linear program is determined by the equation  $Z^{LP}$ , wherein  $Z^{LP}$  is a minimum weighted completion time,  $j$  is a job in a plurality of jobs  $J$ ,  $w_j$  represents the weight of each job, and  $C_j$  represents a calculated completion time of each job.

3. The method of claim 1, wherein determining the task schedule comprises:

determining a ranked order of the plurality of jobs based on the lower bound completion time of each of the plurality of jobs.

4. An apparatus for scheduling a plurality of tasks, wherein each task is assigned to one of a plurality of processors, the apparatus comprising one or more computer processors and memory, the one or more computer processors and memory configured to:

determine a lower bound completion time for each of a plurality of jobs, each of the plurality of jobs comprising a respective plurality of tasks; and

determine a task schedule for each of the plurality of processors based on the lower bound completion times.

5. The apparatus of claim 4, wherein the one or more computer processors and memory further configured to:

schedule, at each processor, tasks based on the task schedule.

6. The apparatus of claim 4, wherein the one or more computer processors and memory are further configured to:

solve a linear program to determine the lower bound completion time, wherein

$$Z^{LP} = \min \sum_{j \in J} w_j C_j$$

the linear program is determined by the equation  $Z^{LP}$ , wherein  $Z^{LP}$  is a minimum weighted completion time,  $j$  is a job in a plurality of jobs  $J$ ,  $w_j$  represents the weight of each job, and  $C_j$  represents a calculated completion time of each job.

7. The apparatus of claim 4, wherein the one or more computer processors and memory are further configured to:

determine the lower bound completion time of each of the plurality of jobs in response to the arrival of a new job.

8. An article of manufacture including a tangible non-transitory computer-readable medium having instructions stored thereon, that in response to execution by a computing device cause the computing device to perform operations comprising:

determining a lower bound completion time for each of a plurality of jobs, each of the plurality of jobs comprising a respective plurality of tasks; and

determining a task schedule for each of the plurality of processors based on the lower bound completion times.

9. The article of manufacture of claim 8, wherein the operation of determining the lower bound completion time for each of the plurality of jobs is performed at a predetermined time interval in response to the arrival of new jobs during the predetermined time interval.

10. The article of manufacture of claim 8, wherein the lower bound completion time for each of the plurality of jobs and task schedule are determined according to a weighting factor assigned to each of the plurality of jobs.

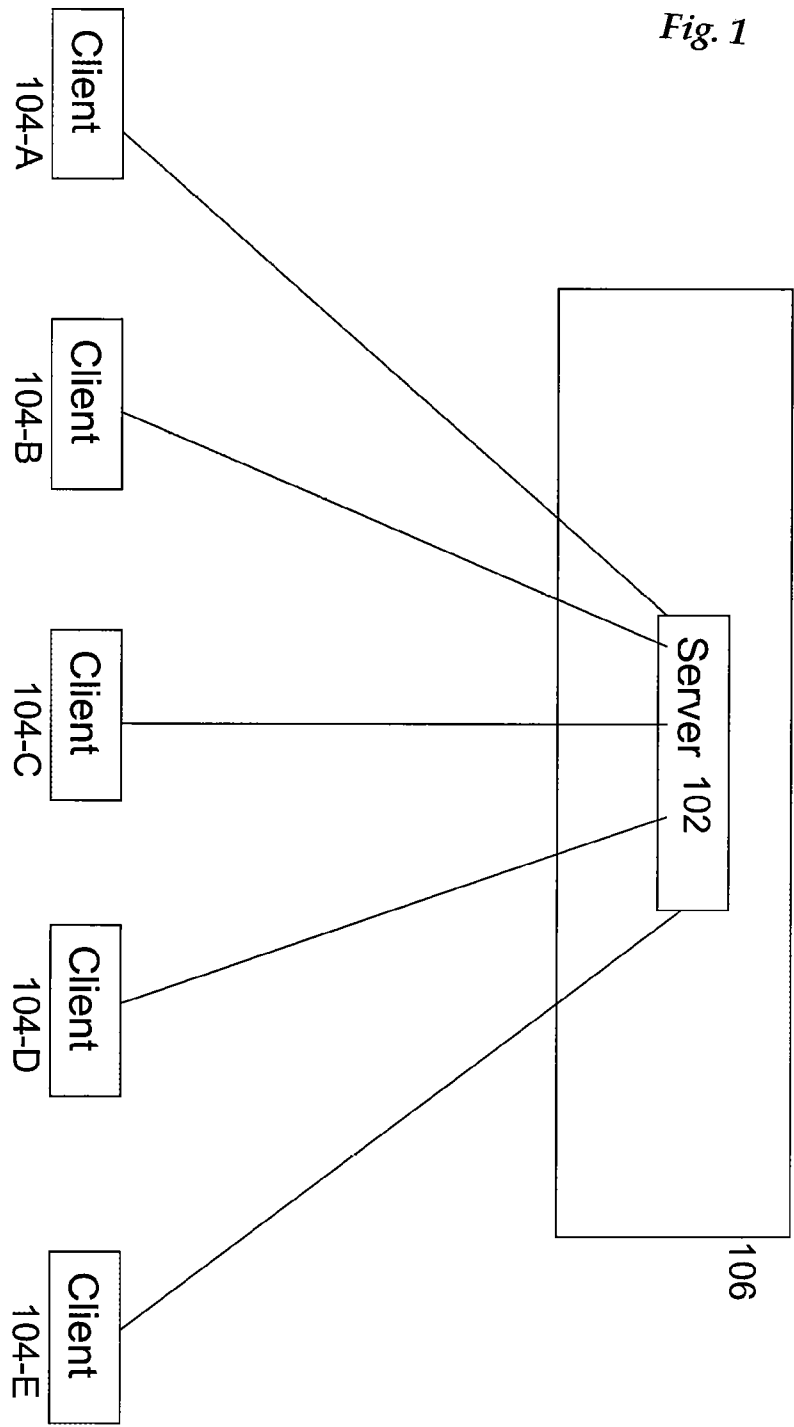
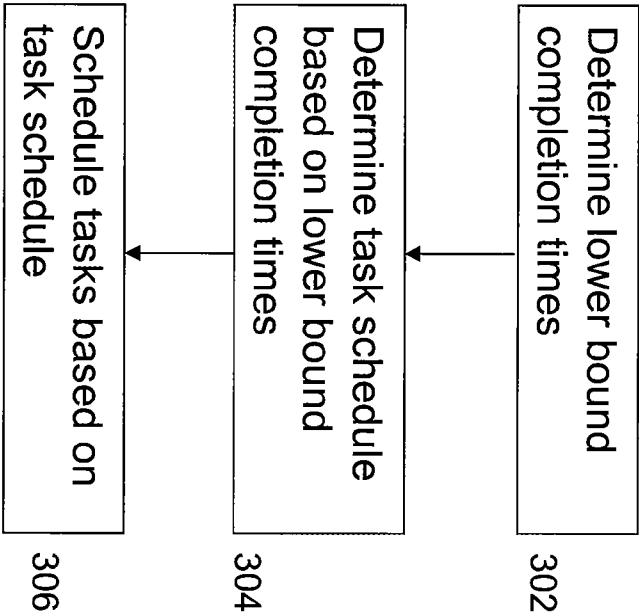


Fig. 1

Fig. 2

| 204 |                              | 206     |  | 208 |  |
|-----|------------------------------|---------|--|-----|--|
| Job | Lower Bound Completion Times | Ranking |  |     |  |
| 25  | 5                            | 1       |  |     |  |
| 19  | 11                           | 2       |  |     |  |
| 17  | 12                           | 3       |  |     |  |
| .   | .                            | .       |  |     |  |
| .   | .                            | .       |  |     |  |
| .   | .                            | .       |  |     |  |
| 2   | 13                           | 139     |  |     |  |
| 8   | 14                           | 140     |  |     |  |
| 10  | 17                           | 141     |  |     |  |
| 1   | 19                           | 142     |  |     |  |

Fig. 3



4/5

Fig. 4

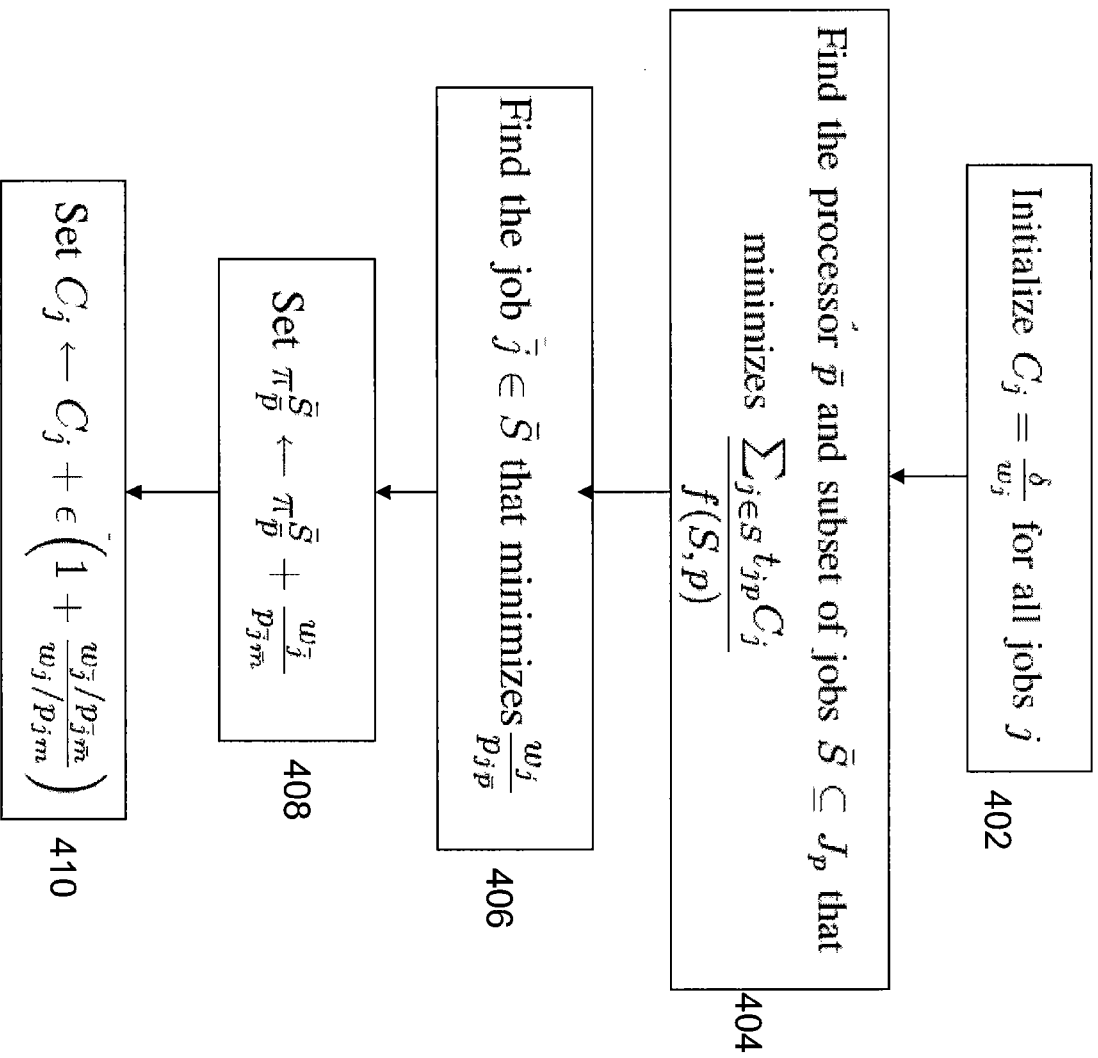
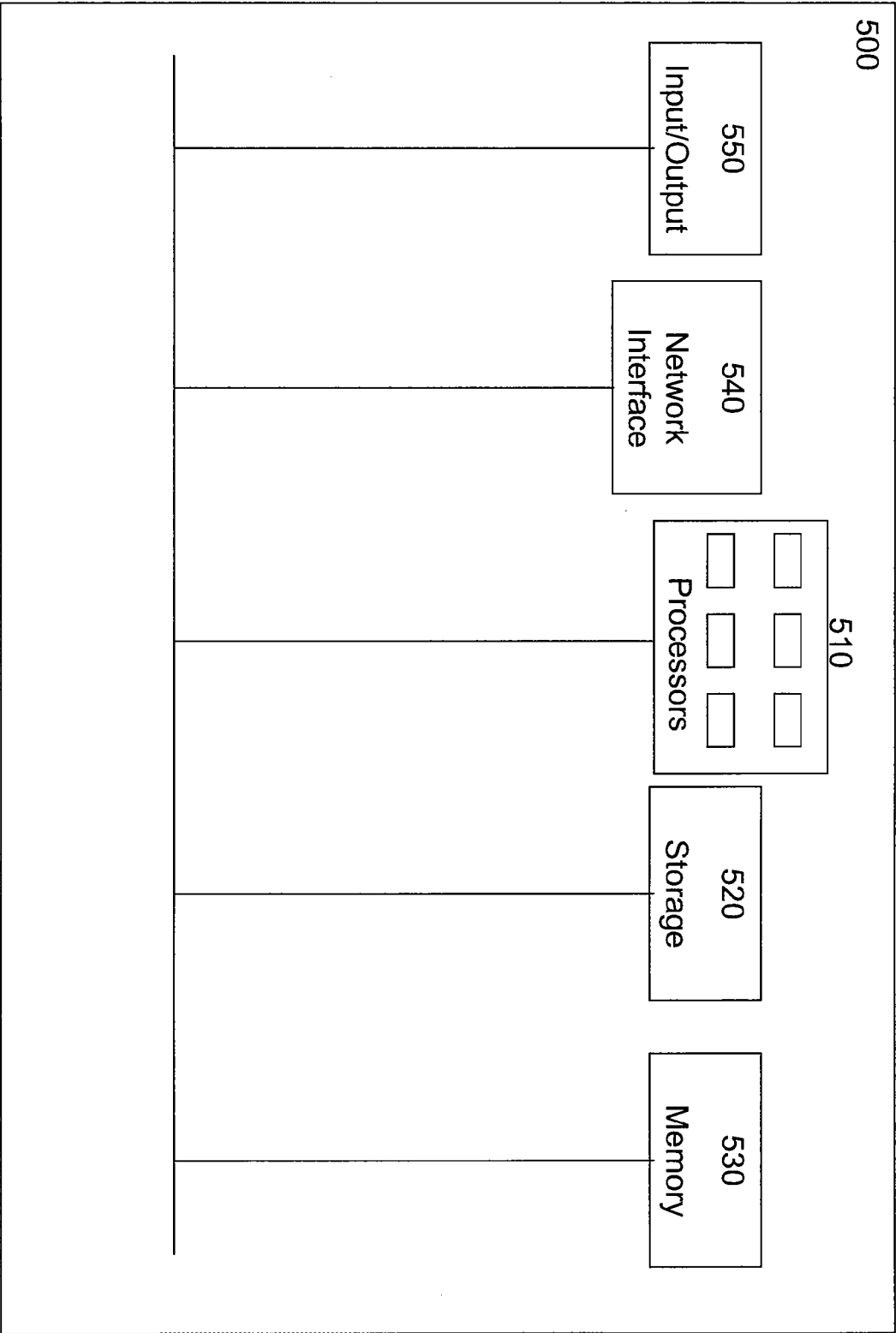


Fig. 5



# INTERNATIONAL SEARCH REPORT

International application No  
PCT/US2012/034191

A. CLASSIFICATION OF SUBJECT MATTER  
INV. G06F9/48  
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)  
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, INSPEC, WPI Data

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                            | Relevant to claim No. |
|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| X         | HYUNSEOK CHANG ET AL: "Scheduling in mapreduce-like systems for fast completion time",<br>INFOCOM, 2011 PROCEEDINGS IEEE, IEEE,<br>10 April 2011 (2011-04-10), pages<br>3074-3082, XP031953538,<br>DOI: 10.1109/INFCOM.2011.5935152<br>ISBN: 978-1-4244-9919-9<br>the whole document<br>----- | 1-10                  |

☐

Further documents are listed in the continuation of Box C.

☐

See patent family annex.

\* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

28 August 2012

Date of mailing of the international search report

05/09/2012

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040,  
Fax: (+31-70) 340-3016

Authorized officer

Noll, Joachim