Optimal Power-of-Two Replenishment Strategies in Capacitated General Production / Distribution Networks

A. Federgruen • Yu-Sheng Zheng
Graduate School of Business, Columbia University, New York, New York 10027
Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104

In this paper we develop a model for a capacitated production / distribution network of general (but acyclic) topology with a general bill of materials, as considered in MRP (Material Requirement Planning) or DRP (Distribution Requirement Planning) systems. This model assumes stationary, deterministic demand rates and a standard stationary cost structure; it is a generalization of the uncapacitated model treated in the seminal papers of Maxwell and Muckstadt (1985) and Roundy (1986). The capacity constraints consist of bounds on the frequency with which individual items can or need to be replenished.

We derive a pair of simple and efficient algorithms capable of determining an optimal power-of-two policy. These algorithms consist of a limited number of maximum flow computations in networks closely related to the production / distribution network. The complexity of these algorithms, even when applied to the uncapacitated model, compares favorably with that of the existing alternative solution methods.

(*General Production / Distribution Networks; Capacity Constraints; Power-of-two Policies*)

For most production / distribution systems, the task of identifying cost effective replenishment strategies for raw materials, work in-progress and finished goods' inventories is complicated by the interdependencies that exist between different items and production / distribution stages. The replenishment strategy of a given item, at a given location or production stage, cannot be determined in isolation but needs to be closely coordinated with that of all preceding and subsequent production stages (or higher and lower level stocking points in a distribution system). Often, additional complications in identifying feasible and cost effective replenishment strategies arise due to *capacity* constraints which impose bounds on the *frequency* with which individual items can be replenished.

In this paper we develop a model for a capacitated production / distribution network of general (but acyclic) topology with a general bill of materials, as considered in MRP (Material Requirement Planning) or DRP (Distribution Requirement Planning) systems. This model assumes stationary, deterministic demand rates and a standard stationary cost structure; it is a generalization of the uncapacitated model treated in the seminal papers of Maxwell and Muckstadt (1985) and Roundy (1986) where it is shown that a *simple*, so-called *power-of-two* policy exists whose cost is guaranteed to come within a few percentage points of optimality. A power-of-two policy prescribes for each product i a replenishment interval T_i such that a replenishment for this product occurs at times $0, T_i, 2T_i, 3T_i, \cdots$ and its inventory level is *zero* at each such replenishment epoch. Moreover, every replenishment interval is chosen as a power-of-two multiple of a common base planning period. We derive a pair of simple and efficient algorithms capable of determining an optimal power-of-two policy. The complexity of these al-

0025-1909/93/3906/0710$01.25

gorithms, even when applied to the uncapacitated model, compares favorably with that of the existing alternative solution methods.

More specifically, consider a production/distribution network represented by a general directed acyclic network G, with node set $N(G)$ and arc set $A(G)$. Let $N = |N(G)|$ and $A = |A(G)|$. Each node is associated with a "product," where a "product" represents a specific in-process or finished item, at a given physical location and/or production stage. With this general "product" definition, a directed arc(i, j) between a pair of nodes $i, j \in N(G)$ indicates that product i is used to "produce" product j. The network is assumed to be acyclic to exclude circuits of products, each of which is consumed in producing its successor.

External demands may occur for any of the items, i.e., at any of the nodes in the network. These demands occur continuously at (item specific) constant rates. Components may be assembled in any given proportions. No backlogging is allowed. The cost structure consists of item-specific inventory carrying costs and variable and fixed production/distribution costs. Inventory carrying costs are incurred at constant rates per unit and per unit of time. Variable production costs are proportional to the production volumes. All production (distribution) orders are instantaneously delivered.

Upper and lower bound constraints may be imposed on the replenishment frequency of each individual item. We first describe a variety of settings in which *capacity* constraints of this type arise because of *externally imposed*, technical or physical limitations.

(1) *Constraints on Production and Distribution Resources*

Assume a product is manufactured in a dedicated facility with a capacity of C hours (e.g., per day). For each production run, a setup time S is incurred; variable processing times increase proportionally with the size of the production run at a rate of u hours per unit produced. If a policy replenishes the product every T days, the total capacity required for a single run cannot be larger than the available capacity in an interval of length T, i.e.,

$$u(Td) + S \le TC \quad \text{or} \quad T \ge b = S/(C - ud).$$

Similar capacity constraints arise, e.g., in multiechelon distribution systems in which a facility is supplied from a unique internal or external supplier, via a fleet of vehicles of given capacity C.

(2) *Limited Storage Space*

Assume a product with demand rate d is stored in a dedicated storage area with capacity Q. Thus if a zero-inventory ordering policy replenishes the product every T time units, the order quantity (dT) represents the maximum inventory level and cannot exceed the capacity Q. Thus, $dT \le Q$ or $T \le b = Q/d$.

(3) *Palletized Reorders*

Strongly advocated in Just-in-Time programs is the use of pallets capable of containing for item i a multiple, say b_i of its daily demand d_i. It is reasonable to consider situations where all pallet sizes $(b_i: i \in N)$ are chosen as power-of-two multiples of a common standard size T_L, i.e., for every item i, $b_i = 2^{n_i} T_L$ for some integer n_i.

As the product is progressively manufactured, it is also reasonable for lots to be split so that $n_i \ge n_j$ for $(i, j) \in A(G)$, perhaps because of the size of the now partially assembled product. As a result, setup costs fail to be independent of the batch size Q, but instead are proportional with the required number of pallets; thus,

$$K_i(Q) = K\lceil Q/(b_i d_i)\rceil \qquad (1)$$

where $\lceil x\rceil$ represents the smallest integer greater than or equal to x. It is easily verified (see Appendix 1) that an optimal nested power-of-two policy can be found for this setting, assuming a *simple* fixed setup cost K (instead of the step function (1)) prevails and imposing the constraint $T_i \le b_i$.

There are many other settings where the cost associated with a production run or distribution delivery is a *step function* of the production (distribution) volume, as in (1). Examples include production with containers or vessels or distribution with vehicles of given, standardized sizes. Order cost structures of this type have, for example, been considered in Lippman (1969, 1971), Swoveland (1975), Aucamp (1982) and Joneja (1989).

We refer to Zheng (1987) for a description of a branch-and-bound method which may be used when the b_i numbers fail to be nested power-of-two multiples. An instance of the model in this paper is used to evaluate each node of the branch-and-bound tree.

In addition, upper bounds for the replenishment cycles are often *self-imposed* in Just-in-Time programs because of the increasing realization that inventory re-

ductions which result from shorter cycles have benefits far beyond those of reduced direct inventory carrying charges. For example, warehouse and floor space needs are reduced significantly to the point where complete warehouses can be closed down and significant amounts of plant floor space become available for alternative uses. Material handling costs can be cut. Reduced inventories imply shorter leadtimes, which in turn allow for faster responses to changing customer needs and faster feedback with quality problems, thereby reducing scrap losses and rework costs. We refer to models *with* such capacity constraints as *capacitated* models, and those *without* such constraints as *uncapacitated* models.

As mentioned above, the uncapacitated model was introduced in a seminal paper by Maxwell and Muckstadt (1985). These authors identify an $O(N^4)$ (Divide-and-Conquer) algorithm for the determination of an optimal *nested power-of-two* policy. Under a *nested* policy, each node j places an order each time any one of its immediate predecessor nodes i does. (Node i is a predecessor of j, if $(i, j) \in A(G)$.)

Roundy (1986) shows that the cost of the best *nested* power-of-two policy may be arbitrarily bad compared to the optimum cost value, but that the cost of the best unrestricted (i.e., nested or unnested) power-of-two policy is guaranteed to come within 6% or 2% of the optimum cost value (depending upon whether the base planning period is fixed or may be varied respectively). He also shows that the Divide-and-Conquer algorithm can be used to determine an optimal power-of-two policy in $O(R^3N)$ time with R the number of routes in the network, i.e., the number of directed paths in G that start at an arbitrary node and terminate at a node with external demand. In this paper we consider a fixed base planning period.

As mentioned above, we derive a pair of algorithms capable of determining an optimal power-of-two policy for the general capacitated model. The fastest of these algorithms determines an optimal power-of-two policy in $O(RN^2)$ elementary operations. When used to determine the best *nested* power-of-two policy, its complexity is $O(N^3)$. The complexity of these algorithms, even when applied to the uncapacitated model, compares favorably with that of existing alternative methods, which, we recall, is $O(R^3N)$ to find the best unrestricted power-of-two policy and $O(N^4)$ to find the best nested power-of-two policy. Both of the proposed algorithms consist of a limited number of *maximum flow* computations thus enabling the use of *standard* software packages.

As shown in Maxwell and Muckstadt (1985) and Roundy (1986), the problem of determining an optimal power-of-two policy may be formulated as a nonlinear integer program with the replenishment intervals as decision variables. As is the case for the Maxwell and Muckstadt algorithm, the first of the two algorithms proposed in this paper is based on a *two-stage procedure*: the *first* stage constructs a solution of the *continuous relaxation* of the problem; in the *second* stage this solution is rounded to a power-of-two vector. When applied to uncapacitated models, the two-stage procedure may in fact be viewed as a variant of Maxwell and Muckstadt's Divide-and-Conquer algorithm. We develop in addition a *direct* algorithm which solves the integer problem directly. Its complexity is smaller than that of the two-stage algorithm by a factor N. The two stage procedure on the other hand, has the advantage of generating (as an intermediate result) a solution to the continuous relaxation, the cost of which provides a lower bound for the minimum cost value.

The two-stage algorithm is based on a characterization theorem describing necessary and sufficient conditions for an optimal solution of the relaxed program. This characterization theorem extends the corresponding theorem for uncapacitated models, as obtained in Jackson et al. (1988). Our proof is based on two simple duality results only: the duality theorem for convex programming and the max-flow min-cut theorem. The proof motivates our maximum flow based algorithm. A similar characterization theorem may be obtained for the original integer program and motivates the direct algorithm.

The remainder of this paper is organized as follows. In §1, we introduce the uncapacitated and capacitated model and discuss various settings in which bounds arise for the items' replenishment frequencies. Section 2 states the characterization theorem for the relaxed problem and the two-stage algorithm; in §3, we describe the direct algorithm and the underlying characterization algorithm for the original integer program. In §4 we describe how our algorithms specialize when applied to the uncapacitated model.

1. The Model

In this section we introduce notation and preliminaries required for the remainder of this paper. We first specify a mathematical programming formulation for the model without capacity constraints (subsection 1.1); next, capacity constraints are added to the model.

1.1. The Uncapacitated Model

Since the network G is acyclic, assume without loss of generality that the products are numbered such that $(i, j) \in A(G)$ only if $i < j$.

For each node $i \in N(G)$, let $P(i)$ $[S(i)]$ indicate the set of its immediate predecessors (successors) in the network, i.e.,

$$P(i) = \{ j \in N(G) | (j, i) \in A(G) \} \quad \text{and}$$
$$S(i) = \{ j \in N(G) | (i, j) \in A(G) \}.$$

For each arc $(i, j) \in A(G)$, λ_{ij} represents the number of units of product i required to produce one unit of product j. Let d_i' represent the rate at which *external* demands for product i arise; h_i' denotes the cost per unit of time for carrying one unit of product i. The incremental holding cost rate for product i is given by

$$h_i = h_i' - \sum_{j \in P(i)} \lambda_{ji} h_j'. \tag{1}$$

These incremental holding cost rates are assumed to be nonnegative, i.e., $h_i \geq 0$, $i \in N(G)$. Let N_0 be the set of products with no predecessors. Without loss of generality, we assume that $h_i > 0$ for all $i \in N_0$: if $h_i = 0$ for some $i \in N_0$, infinitely large quantities of this product may be ordered from external sources with zero resulting inventory costs. Such a product may clearly be eliminated from the model.

Let K_i denote the fixed cost of replenishing product i. Recall that the variable production / distribution costs are assumed to be proportional to the corresponding production / distribution volumes. The value of these cost components is thus constant under any reasonable replenishment strategy, i.e., any strategy under which the items' long run average production rates equal the corresponding demand rates, and may hence be ignored.

The long-run average holding costs under a power-of-two policy are most easily assessed by charging the incremental holding cost rates to *echelon* inventories. The echelon inventory of product i is the number of units of that product in inventory at node i, or as "components" of units of inventory at any of product i's direct or indirect successor nodes.

Note first that if a power-of-two replenishment policy $\mathbf{T}$ is *nested*, the echelon inventory of any product i reaches zero at each of its replenishment epochs, since these epochs represent replenishment epochs for all successor nodes as well. This is a consequence of the zero-inventory ordering property discussed in more detail in Roundy (1986) and Zheng (1987). It is also easily verified that, under a *nested* power-of-two policy, the echelon inventory of any node i is increased by a constant amount at each of its replenishment epochs, and decreases in between at a constant so-called *induced* demand rate d_i. This rate can be computed recursively, starting with $i = N$, via:

$$d_i = d_i' + \sum_{j > i} \lambda_{ij} d_j, \quad i \in N(G).$$

Thus, each node's echelon inventory follows a saw-toothed pattern, and the long run average holding costs are thus given by

$$H[\mathbf{T}] = \sum_{i \in N(G)} H_i T_i, \quad \text{where} \quad H_i = \frac{1}{2} d_i h_i. \tag{2}$$

The computation of $H[\mathbf{T}]$ becomes more involved when $\mathbf{T}$ fails to be nested. In this case, replenishment quantities fail to be constant for at least some nodes. However, these quantities follow a periodic pattern which may be determined by solving a recursive linear system of equations. Theorem 2.4 in Zheng (1987) shows that the solution of this system requires at most

$$2(A + N)(\log_2(T_{\max}/T_{\min}) + 1) + \tfrac{1}{2} A(\log_2(T_{\max}/T_{\min}))^2$$

operations. Neither the *echelon* inventories nor the physical inventories need to follow simple saw-toothed patterns. However, it is shown in Roundy (1986) that echelon inventories may be decomposed into several components which do follow such patterns.

Let $R(G)$ denote the set of *routes* in the network G where a route $r = (i_1, \ldots, i_m)$ is a directed path in G starting at an arbitrary node and terminating at an end-product. Let $R = |R(G)|$. For each route $r = (i_1, \ldots, i_m) \in R(G)$, let $d_r = \lambda_{i_1 i_2} \cdots \lambda_{i_{m-1} i_m} d_{i_m}'$ and $H_r = \frac{1}{2} h_r d_r$

denote its demand and holding cost rate respectively. It is shown in Roundy (1986) (see also Lemma 2.7 in Zheng (1987)) that

$$H[\mathbf{T}] = \sum_{r \in R(G)} H_r(\max_{i \in r} T_i). \tag{3}$$

The problem of finding the best *nested* power-of-two policy may thus be formulated as (see Maxwell and Muckstadt 1985):

$$\text{(NP)} \quad \min_{\mathbf{T}>0} \sum_{i \in N(G)} [K_i/T_i + H_i T_i]$$

$$\text{s.t. } T_i \geq T_j; \quad (i,j) \in A(G),$$

$$T_i = 2^{m_i} T_L; \quad m_i \text{ integer}; \quad i \in N(G)$$

where T_L is a fixed-base planning period.

The problem of finding the best (unrestricted) power-of-two policy may be formulated as (see Roundy 1986):

$$\text{(P)} \quad \min_{\mathbf{T}>0} \sum_{i \in N(G)} K_i/T_i + \sum_{r \in R(G)} H_r T_r \tag{4}$$

$$\text{s.t. } T_r = \max_{i \in r} T_i, \quad r \in R(G), \tag{5}$$

$$T_i = 2^{m_i} T_L; \quad m_i \text{ integer}; \quad i \in N(G). \tag{6}$$

We observe that (5) may be replaced by

$$T_r \geq T_i; \quad i \in r; \quad i \in N(G), \quad r \in R(G). \tag{5'}$$

Since $H_r > 0$ for all $r \in R(G)$, if $\mathbf{T}$ minimizes (4) subject to (5') and (6), then (5) is satisfied as well.

We conclude that both (P) and (NP) represent special cases of the generic problem:

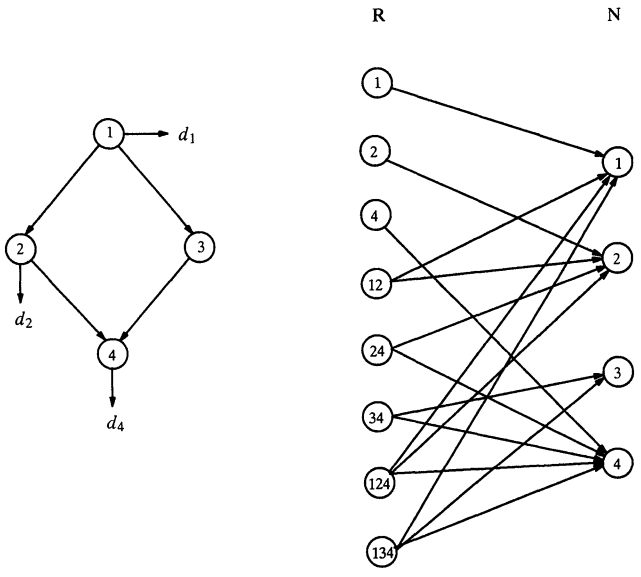
$$\text{(GP)} \quad \min_{\mathbf{T}>0} \sum_{i \in \mathbf{N}} [K_i/T_i + H_i T_i] \tag{7}$$

$$\text{s.t. } T_i \geq T_j; \quad (i,j) \in \mathbf{A}, \tag{8}$$

$$T_i = 2^{m_i} T_L; \quad m_i \text{ integer } (i \in \mathbf{N}). \tag{9}$$

For (P), $\mathbf{N} = R(G) \cup N(G)$ and $\mathbf{A} = \{(r, i): r \in R(G), i \in r\}$. We call $(\mathbf{N}, \mathbf{A})$ the route-product network. (We continue to use $\mathbf{N}$ and $\mathbf{A}$ both to denote the node and arc sets as well as their cardinalities.) Note that this network differs from the original network G, in that it is *bipartite* with arcs going from the route nodes to product nodes only, see Figure 1. It is also different from the "route network" used in Roundy (1986); our algorithms take essential advantage of its bipartite structure.

Figure 1 Route Product Network



The number of routes in a general product network may clearly grow exponentially with the number of products; in most practical production/distribution settings, however, the product network is extremely sparse, and R is of reasonable size. The route-product network can be generated in $0(R)$ time, see Zheng (1987). It is noteworthy that the complexity of the algorithms developed in this paper remains linear in R.

1.2. The Capacitated Model

The above model assumes that no restrictions apply to the products' replenishment frequencies. As explained in the introduction, there are many settings where such frequency constraints need to be enforced. When restricting oneself to the class of power-of-two policies, these frequency constraints can often be translated into simple upper bounds (or lower bounds) for the products' replenishment intervals. When added to (7)–(9), the formulation of (GP) becomes:

$$\text{(GCP)} \quad \min(7)$$

$$\text{s.t. } (8), (9);$$

$$T_i \leq b_i; \quad i \in \mathbf{N}. \tag{10}$$

If lower bounds $T_i \geq b_i$ are required rather than (10), the resulting model remains equivalent to (GCP) as is easily verified by applying the substitution of variables $T_i = Z_i^{-1}$ $(i \in \mathbf{N})$. The results of this paper can easily

be extended to include the case of both upper and lower bounds. We assume that all b_i are power-of-two multiples of T_L. Since T_L is fixed, this assumption is without loss of generality, as each b_i may be replaced by the largest power-of-two multiple of T_L, that is smaller than or equal to b_i ($i \in \mathbf{N}$) without changing the feasible region.

Problem (GCP) is a special case of a more general model in Jackson et al. (1988) which, for a given partition of the T-variables, imposes an upper bound on a weighted sum of the variables in each set of the partition. This more general problem is much harder to solve. The authors suggest a Lagrangean relaxation heuristic with (GP) instances or (GP) instances plus a single linear constraint as the relaxed problems.

1.3. A Preliminary Duality Result

The continuous relaxation of (GCP) is obtained by relaxing the power-of-two integrality constraints:

$$\text{(RCP)} \quad c^* = \min_{\mathbf{T}>0} c(\mathbf{T}) = \sum_{i\in\mathbf{N}} (K_i/T_i + H_iT_i) \tag{11}$$

$$\text{subject to } T_i \geq T_j, \quad (i,j) \in \mathbf{A}, \tag{12}$$

$$0 < T_i \leq b_i, \quad i \in \mathbf{N}. \tag{13}$$

A dual of the convex program (RCP) is:

$$\text{(CD)} \quad \max D(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v}) = \sum_{i\in\mathbf{N}} 2(K_i v_i)^{1/2} - \sum_{i\in\mathbf{N}} b_i\lambda_i \tag{14}$$

$$\text{subject to } \lambda_i + H_i + \sum_{l\in P_i} x_{li} - \sum_{j\in S_i} x_{ij} = v_i, \quad i \in \mathbf{N}, \tag{15}$$

$$\boldsymbol{\lambda} \geq 0, \quad \mathbf{x} \geq 0, \quad \mathbf{v} \geq 0. \tag{16}$$

LEMMA 1.

(a) (CD) *is the Lagrangian dual program of* (RCP). *For any* $\mathbf{T} > 0$ *and feasible solution* $(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v})$ *for* (CD) $c(\mathbf{T}) \geq D(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v})$;

(b) $\mathbf{T}^*$ *is optimal for* (RCP) *if and only if there exists a feasible solution* $(\boldsymbol{\lambda}^*, \mathbf{x}^*, \mathbf{v}^*)$ *for* (CD) *such that* $c(\mathbf{T}^*) = D(\boldsymbol{\lambda}^*, \mathbf{x}^*, \mathbf{v}^*)$;

(c) *the complementary slackness conditions*

$$x_{ij}^*(T_j^* - T_i^*) = 0, \quad (i,j) \in \mathbf{A},$$

$$\lambda_i^*(T_i^* - b_i) = 0, \quad i \in \mathbf{N},$$

hold for any optimal solution $\mathbf{T}^*$ *of* (RCP) *and optimal solution* $(\boldsymbol{\lambda}^*, \mathbf{x}^*, \mathbf{v}^*)$ *of* (CD).

PROOF. To prove part (a) let x_{ij} and λ_i be Lagrange multipliers associated with constraints (12) and (13) respectively. The Lagrangian of (RCP) is given by

$$L(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{T}) = \sum_{i\in\mathbf{N}} (K_i/T_i + H_iT_i) + \sum_{(i,j)\in\mathbf{A}} x_{ij}(T_j - T_i) + \sum_{i\in\mathbf{N}} \lambda_i(T_i - b_i).$$

The Lagrangian dual program is given by (see Geoffrion 1971)

$$\text{(CD)} \quad \max_{\mathbf{x},\boldsymbol{\lambda}\geq 0} \inf_{\mathbf{T}>0} L(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{T}). \tag{17}$$

Collecting the terms in T_i ($i \in \mathbf{N}$), we have

$$L = \sum_{i\in\mathbf{N}} \left(K_i/T_i + \left(\lambda_i + H_i + \sum_{l\in P_i} x_{li} - \sum_{j\in S_i} x_{ij}\right)T_i - b_i\lambda_i\right).$$

Let

$$v_i \overset{\text{def}}{=} \lambda_i + H_i + \sum_{l\in P_i} x_{li} - \sum_{j\in S_i} x_{ij} \quad (i \in N).$$

Since $L(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{T})$ is unbounded below when some $v_i < 0$, we may impose the constraint $\mathbf{v} \geq 0$ without affecting the value of the maximum in (14). Under these additional constraints,

$$\inf_{\mathbf{T}>0} L = \sum_{i\in\mathbf{N}} 2(K_i v_i)^{1/2} - \sum_{i\in\mathbf{N}} b_i\lambda_i.$$

Parts (b) and (c) follow from Theorem 3 in Geoffrion (1971) and the stability of (RCP). To verify the latter, it suffices to show that Slater's condition holds, i.e., there exists some $\mathbf{T} > 0$ which is an interior point to the feasible region described by (12) and (13); note that

$$\mathbf{T} = \{T_i = \epsilon(N - i + 0.1), i \in \mathbf{N}\}$$

is an interior point for ϵ sufficiently small. □

In the remainder we need the following definitions.

DEFINITION 1. G_l is a *subgraph* of $\mathbf{G}$ if (i) $N(G_l) \subset \mathbf{N}$, (ii) $(i, j) \in A(G_l)$ whenever $(i, j) \in \mathbf{A}$ and $i, j \in N(G_l)$.

DEFINITION 2. $(G_1, G_2, \ldots, G_M)$ is a *partition* of $\mathbf{G}$ if the node sets of subgraphs $G_1, G_2, \ldots, G_M$ form a partition of $\mathbf{N}$.

DEFINITION 3. A partition $(G_1, G_2, \ldots, G_M)$ is *directed* if there is no arc$(i, j) \in \mathbf{A}$ with $i \in N(G_a)$, $j \in N(G_b)$, and $a < b$.

DEFINITION 4. For any given subgraph G_l of G, we define the *extended* graph G_l' of G_l as the graph obtained by adding a source node s and a sink node t and connecting these with all of the nodes in $N(G_l)$ by arcs(s, i) and (i, t) $(i \in N(G_l))$.

In the uncapacitated model, the dual program (CD) has $\lambda = 0$ and the constraints (15) may thus be viewed as flow conservation constraints in the extended graph G_l', with x_{ij} as the flow on arc$(i, j) \in \mathbf{A}$, v_i as the flow on arc(i, t) and the flow on arc(s, i) is H_i. The objective is to maximize a specific separable concave function of $\mathbf{v}$, the vector of outflows to the sink. Network flow problems of this type have been studied by Veinott (1971), Fujishige (1980), Groenevelt (1985), and Federgruen and Groenevelt (1986, 1987). Efficient algorithms for this more general class of network flow problems have been proposed by Groenevelt (1985) and Federgruen and Groenevelt (1986, 1987).

In the case of uncapacitated *nonnested* models, $\mathbf{G}$ represents the route product network (see Figure 1) and problem (CD) takes the form

$$\text{(D1)} \quad \max \sum_{i \in N(G)} 2(K_i v_i)^{1/2}$$

$$\text{subject to} \sum_{i \in r} x_{ri} = H_r, \quad r \in R(G),$$

$$\sum_{r: i \in r} x_{ri} = v_i, \quad i \in N(G),$$

$$\mathbf{x}, \mathbf{v} \geq 0.$$

These above network flow problems are thus *bipartite*, a special structure which may be exploited to achieve considerable efficiency improvements.

2. A Two-stage Procedure Based on a Characterization Theorem for (RCP)

In this section we derive a two-stage procedure for the general capacitated problem (GCP). In the first stage an optimal solution of its continuous relaxation (RCP) is determined. As pointed out in the introduction, the optimal value of (RCP) is a lower bound, not only for the minimum cost among all power-of-two policies but also for the overall minimum cost among *all* (feasible) policies. For the uncapacitated model this result is shown in Roundy (1986). For the capacitated model this is proved in Appendix 2. From this lower bound result, it follows (see Theorem 3) using standard arguments that the average cost of an optimal power-of-two policy comes within 6% of this lower bound and hence of optimality.

We start with a theorem characterizing optimal solutions of the relaxed problem. The proof of this characterization theorem is based on a simple application of the duality theorem of convex programming and the max-flow min-cut theorem. Throughout this and the following section, we need to discuss maximum flows in the following associated networks $G_l'(\tau)$ and their "*relaxed*" associated networks $G_l^+(\tau)$.

DEFINITION 5. For any subgraph G_l of G and any $\tau > 0$, define the *associated graph* $G_l'(\tau)$ as the extended network G_l' with upper bounds H_i and K_i/τ^2 on the arcs(s, i) and (i, t) respectively, and with infinite capacities on all other arcs.

For any subgraph $G_l \subset \mathbf{G}$, let

$$b(G_l) = \min_{i \in N(G_l)} b_i, \quad N^0(G_l) = \{i \in N(G_l): b_i = b(G_l)\},$$

$$K(G_l) = \sum_{i \in N(G_l)} K_i \quad \text{and} \quad H(G_l) = \sum_{i \in N(G_l)} H_i.$$

DEFINITION 6. For a given subgraph $G_l \subset \mathbf{G}$ and $\tau > 0$, a relaxed associated network $G_l^+(\tau)$ is a network obtained by relaxing the upper bounds on $\{(s, i): i \in N^0(G_l)\}$ in the associated network $G_l'(\tau)$.

THEOREM 1. *Let* $\mathbf{T}^* \in R_+^{\mathbf{N}}$ *and* $(G_1, G_2, \ldots, G_M)$ *be a partition of* $\mathbf{G}$ *such that* $N(G_l) = \{i \in \mathbf{N}: T_i^* = T(l)\}$ *with* $T(1) < T(2) < \cdots < T(M)$. $\mathbf{T}^*$ *is an optimal solution of* (RCP) *if and only if the following three conditions hold*:

(i) $(G_1, \ldots, G_M)$ *is a directed partition*,

(ii) $T(l) = \min\{b(G_l), (K(G_l)/H(G_l))^{1/2}\}$,

(iii) *for each* $l = 1, \ldots, M$, *any directed partition* $(\underline{G}_l, \bar{G}_l)$ *of* G_l *satisfies*

(a) $\min((K(\bar{G}_l)/H(\bar{G}_l)^{1/2}, b(\bar{G}_l)) \leq T(l)$, *and*

(b) $(K(\underline{G}_l)/H(\underline{G}_l))^{1/2} \geq T(l)$.

PROOF. We show the sufficiency part first. Suppose $\mathbf{T}^*$ satisfies (i), (ii) and (iii). Feasibility of $\mathbf{T}^*$ follows

immediately from (i) and (ii). To prove optimality, it suffices in view of Lemma 1 to show that there exists a feasible dual solution $(\lambda^*, \mathbf{x}^*, \mathbf{v}^*)$ such that $D(\lambda^*, \mathbf{x}, \mathbf{v}^*) = c(\mathbf{T}^*)$. By (ii) and some algebra

$$c(\mathbf{T}^*) = \sum_{l \in L_1} 2(K(G_l)H(G_l))^{1/2} + \sum_{l \in L_2} (K(G_l)/b(G_l) + H(G_l)b(G_l)), \quad \text{where}$$

$$L_1 = \{l \in \{1, \ldots, M\}: T(l) = (K(G_l)/H(G_l))^{1/2}\};$$

$$L_2 = \{l \in \{1, \ldots, M\}: T(l) = b(G_l) < (K(G_l)/H(G_l))^{1/2}\}.$$

To construct a dual feasible solution achieving the objective function value $c(\mathbf{T}^*)$, let $x_{ij}^* = 0$ for $(i, j) \in \mathbf{A}$ with $i \in N(G_a)$, $j \in N(G_b)$, $a \neq b$. By doing so, we completely decompose (CD) into M separate subproblems of the same type as (CD) itself:

$$(\mathrm{CD}_l) \quad \max D_l(\lambda^l, \mathbf{x}^l, \mathbf{v}^l) = 2 \sum_{i \in N(G_l)} (K_i v_i)^{1/2} - \sum_{i \in N(G_l)} b_i \lambda_i \quad (18)$$

$$\text{s.t. } \lambda_i + H_i + \sum_{j \in P_i} x_{ji} - \sum_{j \in S_i} x_{ij} = v_i, \quad i \in N(G_l), \quad (19)$$

$$\lambda^l \geq 0, \quad \mathbf{x}^l \geq 0, \quad \mathbf{v}^l \geq 0, \quad (20)$$

where $\lambda^l = \{\lambda: i \in N(G_l)\}$, $\mathbf{x}^l = \{x_{ij}: (i, j) \in A(G_l)\}$ and $\mathbf{v}^l = \{\mathbf{v}_i: i \in N(G_l)\}$.
Also let

$$\lambda(G_l) = \sum_{i \in N(G_l)} \lambda_i, \quad v(G_l) = \sum_{i \in N(G_l)} v_i.$$

It thus suffices to show that for every $l \in L_1$ there exists a feasible solution of (CD_1) such that

$$D_l(\lambda^l, x^l, v^l) = 2(K(G_l)H(G_l))^{1/2}$$

and for every $l \in L_2$,

$$D_l(\lambda^l, x^l, v^l) = K(G_l)/b(G_l) + H(G_l)b(G_l).$$

For $l \in L_1$, it suffices to set $\lambda^l = 0$ (as suggested by the complementary slackness conditions in Lemma 1 part (c)) and $v_i^* = K_i/T(l)^2$, $i \in N(G_l)$:

$$D_l(\lambda^l, \mathbf{x}^l, \mathbf{v}^l) = 2\left(\sum_{i \in N(G_l)} K_i\right) \Big/ T(l) = 2(K(G_l)H(G_l))^{1/2},$$

by (ii) and $l \in L_1$.

It thus suffices to show that a vector $\mathbf{x}^l$ exists such that $(\lambda^l, \mathbf{x}^l, \mathbf{v}^l)$ satisfy (19) and (20), i.e., that a flow exists in the extended network G_l' with the flow on arc(s, i) equal to H_i and the flow on arc(i, t) equal to v_i $(i \in N(G_l))$. Alternatively, imposing H_i and v_i^* as upper bounds on the arcs(s, i) and (i, t) respectively $(i \in N(G_l))$ and infinite capacities on all arcs in $A(G_l)$, it suffices to show that a maximum flow exists in the thus capacitated extended network G_l' in which all arcs from the source and all arcs to the sink are saturated, i.e., in which the minimum cut capacity equals $\sum_{i \in N(G_l)} H_i = \sum_{i \in N(G_l)} v_i^*$. Note, however, that any finite capacity cut in this network must be of the type $(\underline{G}_l \cup \{s\}, \bar{G}_l \cup \{t\})$ with $(\underline{G}_l, \bar{G}_l)$ a directed partition. The capacity of such a cut is given by

$$\sum_{i \in N(\underline{G}_l)} v_i^* + \sum_{i \in N(\bar{G}_l)} H_i = K(\underline{G}_l)/T(l)^2 + H(G_l) - H(\underline{G}_l) \geq H(G_l)$$

since $K(\underline{G}_l)/H(\underline{G}_l) \geq T(l)^2$, by (iii) (b).

For $l \in L_2$, $T(l) = b(G_l)$. We construct a feasible solution $(\lambda^l, \mathbf{x}^l, \mathbf{v}^l)$ for (CD) such that

$$D(\lambda^l, \mathbf{x}^l, \mathbf{v}^l) = K(G_l)/b(G_l) + H(G_l)b(G_l). \quad (21)$$

Note that due to (iii) and

$$T(l) = b(G_l) < (K(G_l)/H(G_l))^{1/2},$$

we have for any directed partition $(\underline{G}_l, \bar{G}_l)$ of G_l

$$(K(\underline{G}_l)/H(\underline{G}_l))^{1/2} \geq b(G_l) \geq \min((K(\bar{G}_l)/H(\bar{G}_l))^{1/2}, b(\bar{G}_l)).$$

We construct $(\lambda^l, \mathbf{x}^l, \mathbf{v}^l)$ in the following two-step procedure:

Step 1. Find a maximum flow y^0 in $G_l'(b(G_l))$.

Step 2. Find a maximum flow y in $G_l^+(b(G_l))$ by using any shortest augmenting path algorithm starting with initial flow y^0. Then let $\lambda_i = y_{si} - y_{si}^0$, $v_i = y_{it}$, $i \in N(G_l)$; also let $x_{ij} = y_{ij}((i, j) \in A(G_l))$. We show that $(\lambda^l, \mathbf{x}^l, \mathbf{v}^l)$ is a feasible solution of (CD_l) and that (21) is satisfied. It is easily verified that (20) is satisfied:

$\mathbf{x}^l \geq 0$ and $\mathbf{v}^l \geq 0$ hold while $\lambda_i = y_{si} - y_{si}^0 \geq 0$ ($i \in N(G_l)$) because a shortest augmenting path algorithm never reduces the flow on any arc(s, i). (In other words, arc(s, i) never serves as a backward arc in a shortest augmenting path.)

To see that (19) is satisfied, we need to show

$$y_{si} = H_i + \lambda_i \quad \text{or equivalently} \quad y_{si}^0 = H_i \quad i \in N(G). \tag{22}$$

By the max-flow min-cut theorem, it is equivalent to show that $\{\varnothing, G_l\}$ is a minimum cut of $G'(b(G_l))$, i.e., for any directed partition $(\underline{G}_l, \bar{G}_l)$ of G_l, $H(G_l) \leq H(\bar{G}_l) + K(\underline{G}_l)/b^2(G_l)$, and the latter follows from (iii) (b). Thus $(\lambda^l, \mathbf{x}^l, \mathbf{v}^l)$ is feasible, and it remains to be shown that (21) is satisfied. For the latter, it suffices to show that

$$v_i = K_i/b^2(G_l), \quad i \in N(G_l) \tag{23}$$

because (22) and (23) imply that

$$\lambda(G_l) = \sum_{i \in N(G_l)} y_{si} - H(G_l) = v(G_l) - H(G_l)$$

$$= K(G_l)/b^2(G_l) - H(G_l).$$

$b_i \neq b(G_l)$ implies that $i \notin N^0(G_l)$, and hence that $\lambda_i = 0$. Therefore,

$$b(G_l)\lambda(G_l) = \sum_{i \in N(G_l)} b_i \lambda_i.$$

Thus,

$$D_l(\lambda^l, \mathbf{x}^l, \mathbf{v}^l)$$

$$= 2K(G_l)/b(G_l) - b(G_l)\lambda(G_l)$$

$$= 2K(G_l)/b(G_l) - b(G_l)(K(G_l)/b^2(G_l) - H(G_l))$$

$$= K(G_l)/b(G_l) + H(G_l)b(G_l).$$

To verify (23), it suffices to show that the arcs $\{(i, t): i \in N(G_l)\}$ are saturated by $\mathbf{y}^l$ or equivalently that $\{(i, t): i \in N(G_l)\}$ is a minimum cut of $G^+(b(G_l))$. Note that the capacity of any directed partition $(\underline{G}_l, \bar{G}_l)$ of G_l, with $N(\bar{G}_l) \cap N^0(G_l) = \varnothing$, is

$$H(\bar{G}_l) + K(\underline{G}_l)/b^2(G_l) \geq K(G_l)/b^2(G_l)$$

since $K(\bar{G}_l)/H(\bar{G}_l) \leq b^2(G_l)$ in view of (iii) (a). All other cuts have infinite capacity.

We prove the necessity part via the following simple perturbation argument. Let $\mathbf{T}^*$ be an optimal solution of (RCP). (i) is necessary for feasibility. Suppose to the contrary that (ii) does not hold for some G_l. If $T(l) > (K(G_l)/H(G_l))^{1/2}$ take ϵ sufficiently small such that $T(l) - \epsilon > T(l-1)$ and let $T_i' = T_i^* - \epsilon$ if $i \in N(G_l)$ and $T_i' = T_i^*$ otherwise. (We refer to $\mathbf{T}'$ as a negative ϵ-perturbation of $\mathbf{T}^*$.) $\mathbf{T}'$ is clearly feasible for (RCP) and $c(\mathbf{T}') < c(\mathbf{T}^*)$, a contradiction. Similarly, if

$$T(l) < \min(b(G_l), (K(G_l)/H(G_l))^{1/2})$$

then a positive ϵ-perturbation of $\mathbf{T}^*$ on G_l would improve $c(\mathbf{T}^*)$. This proves (ii).

For (iii), assume to the contrary that there exists a directed partition $(\underline{G}_l, \bar{G}_l)$ of G_l (for some l) such that (a) or (b) is violated. If (a) does not hold, a positive ϵ-perturbation of $\mathbf{T}^*$ on $\bar{G}_l$ would improve $c(\mathbf{T}^*)$; otherwise, a negative ϵ-perturbation of $\mathbf{T}^*$ on G_l would improve $c(\mathbf{T}^*)$. □

Theorem 1 suggests that solutions to (RCP) may be characterized by partitions of the node set $\mathbf{N}$. For a given partition $(G_1, G_2, \ldots, G_M)$ let the *associated solution* $\mathbf{T}$ be defined by:

$$T_i = \min[b(G_l), (K(G_l)/H(G_l))^{1/2}],$$

$$i \in N(G_l), \quad l = 1, \ldots, M.$$

A partition will be referred to as *optimal*, if the associated $\mathbf{T}$-vector is an optimal solution of (RCP). In view of Theorem 1, a partition $(G_1, \ldots, G_M)$ is optimal if and only if the partition is (i) *directed*; (ii) *monotone*:

$$\min[b(G_l), (K(G_l)/H(G_l))^{1/2}]$$

$$\leq \min[b(G_{l+1}), (K(G_{l+1})/H(G_{l+1}))^{1/2}],$$

$$l = 1, \ldots, M-1;$$

(iii) *maximally fine*: for each $l = 1, \ldots, M$ and any directed partition $(\underline{G}_l, \bar{G}_l)$ of G_l,

$$\min[b(\bar{G}_l), (K(\bar{G}_l)/H(\bar{G}_l))^{1/2}]$$

$$\leq \min[b(\underline{G}_l), (K(\underline{G}_l)/H(\underline{G}_l))^{1/2}].$$

It is easy to find a partition which is directed and monotone: the singleton $\{\mathbf{G}\}$, for example, represents such a partition. (The associated solution $\mathbf{T}$ has all components equal to $\min[b(\mathbf{G}), (K(\mathbf{G})/H(\mathbf{G}))^{1/2}]$.) As shown in Lemma 2 and Lemma 3 below, maximum

flow (or minimum cut) calculations may be used to *check* whether each set in a given partition is maximally fine; otherwise a refined partition may be generated from the computed minimum cut, which remains directed and monotone. This suggests an iterative algorithm which, starting with the singleton $\{\mathbf{G}\}$, results in an optimal partition after a limited number of maximum flow computations and as many partition refinements. Indeed, algorithm (RCP) below generates a finite sequence of progressively refined, directed and monotone partitions.

Algorithm RCP

Step 0. $M := 1$; $l := 1$; $G_1 := G$.

Step 1. Let $\tau = \min(b(G_l), (K(G_l)/H(G_l))^{1/2})$; find a maximum flow in $G'(\tau)$. If $(\varnothing, G_l)$ is a minimum cut go to Step 2. Otherwise we have a nontrivial minimum cut $(\underline{G}_l, \bar{G}_l)$ of G_l. Renumber $(\underline{G}_l, \bar{G}_l, G_{l+1}, \ldots, G_M)$ as $(G_l, G_{l+1}, G_{l+2}, \ldots, G_{M+1})$; $M := M + 1$. Repeat Step 1.

Step 2. If $\tau = (K(G_l)/H(G_l))^{1/2}$ go to Step 3; otherwise find a maximum flow in $G_l^+(\tau)$. If $(G_l, \varnothing)$ is a minimum cut, go to Step 3; otherwise we have a nontrivial cut $(\underline{G}_l, \bar{G}_l)$ of G_l. Renumber $(\underline{G}_l, \bar{G}_l, G_{l+1}, \ldots, G_M)$ as $(G_l, G_{l+1}, G_{l+2}, \ldots, G_{M+1})$; $M := M + 1$.

Step 3. $T(l) := \tau$, $T_i^* = T(l)$ $(i \in N(G_l))$. If $l = M$, stop. Otherwise $l := l + 1$, go back to Step 1.

In Theorem 2 we show that the above algorithm solves (RCP). We first need the following lemmas which characterize minimum cuts in the associated networks $G_l'(\tau)$ and $G_l^+(\tau)$ respectively. These lemmas are proven in Appendix 2.

LEMMA 2. *Let G_l be a subgraph of $\mathbf{G}$ and $\tau > 0$; let (G_1, G_2) be a minimum cut of $G_l'(\tau)$. Then,*

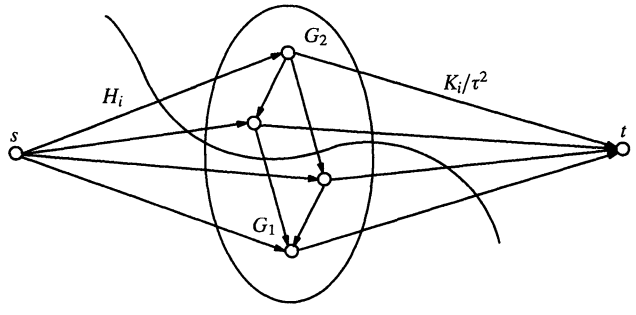
(i) *(G_1, G_2) is a directed partition of G_l.*

(ii) *$K(G_1)/H(G_1) \le \tau^2 \le K(G_2)/H(G_2)$. If the cuts $(\varnothing, G_l)$ or $(G_l, \varnothing)$ are not minimal, these inequalities are strict.*

(iii) *Let $\tau^2 = K(G_l)/H(G_l)$. G_l is maximally fine if the cuts $(\varnothing, G_1)$ and $(G_l, \varnothing)$ are minimal in $G_l'(\tau)$.*

(iv) *If $(\underline{G}_1, \bar{G}_1)$ is a directed partition of G_1, then $K(\bar{G}_1)/H(\bar{G}_1) \le \tau^2$. Similarly, if $(\underline{G}_2, \bar{G}_2)$ is a directed partition of G_2, then $\tau^2 \le K(\underline{G}_2)/H(\underline{G}_2)$.*

Figure 2 **A Minimum Cut in the Associated Network $G_l'(\tau)$**



LEMMA 3. *Let G_l be a subgraph of $\mathbf{G}$, and $\tau > 0$. Let (G_1, G_2) be a minimum cut of $G_l^+(\tau)$.*

(i) *(G_1, G_2) is a directed partition of G_l.*

(ii) *$b(G_1) = b(G_l)$; if $G_2 \neq \varnothing$, $b(G_2) > b(G_l)$, i.e., $N(G_2) \cap N^0(G_l) = \varnothing$.*

(iii) *$K(G_2)/H(G_2) \ge \tau^2$.*

(iv) (a) *If $(\underline{G}_1, \bar{G}_1)$ is a directed partition of G_1, $K(\bar{G}_1)/H(\bar{G}_1) \le \tau^2$ or $b(\bar{G}_1) = b(G_l)$.*

(b) *If $(\underline{G}_2, \bar{G}_2)$ is a directed partition of G_2, $K(\underline{G}_2)/H(\underline{G}_2) \ge \tau^2$.*

THEOREM 2. *The vector $\mathbf{T}^*$ generated by Algorithm RCP is an optimal solution for* (RCP).

PROOF. It suffices to show that $T(1) \le T(2) \le \cdots \le T(M)$ and that (i), (ii), (iii) of Theorem 1 are satisfied. (i) is easily verified by induction using Lemma 2(i) and Lemma 3(i). (ii) holds by the specification of $\mathbf{T}$ and $T(l)$ $(l = 1, \ldots, M)$ in the algorithm. In order to show (iii), for $l = 1, \ldots, M$ let $(\underline{G}_l, \bar{G}_l)$ be any directed partition of G_l. Note $(\varnothing, G_l)$ is a minimum cut of $G_l'(T(l))$. In view of Lemma 2(iv) we have $K(\underline{G}_l)/H(\underline{G}_l) \ge T^2(l)$, i.e., (iii)(b) holds. To prove (iii)(a) and in view of the specification in Step 2, we only need to distinguish among the following three cases:

(1) $T(l) = (K(G_l)/H(G_l))^{1/2}$. Since $K(G_l)/T^2(l) = H(G_l)$, $(G_l, \varnothing)$ is also a minimum cut of $G_l'(T(l))$. By Lemma 2(iii), we have $K(\bar{G}_l)/H(\bar{G}_l) \le T^2(l)$, so that (iii)(a) follows.

(2) $T(l) = b(G_l) < (K(G_l)/H(G_l))^{1/2}$ *and* $(G_l, \varnothing)$ *is a minimum cut of* $G_l^+(T(l))$. Applying Lemma 3(iv)(a) with $G_1 = G_l$, $G_2 = \varnothing$, $\bar{G}_1 = \bar{G}_l$, we have $K(\bar{G}_l)/H(\bar{G}_l) \le T^2(l)$ or $b(\bar{G}_l) = b(G_l)$ which is (iii)(b).

(3) $T(l) = b(G_l) < (K(G_l)/H(G_l))^{1/2}$ *and* (G_l, G_{l+1}) *is a minimum cut of* $(G_l \cup G_{l+1})^+(T(l))$. Applying Lemma 3(iv)(a) with $G_1 = G_l$, $G_2 = G_{l+1}$, $\bar{G}_1 = \bar{G}_l$, we have (iii)(b) again.

It remains to be proven that $T(1) \le T(2) \le \cdots \le T(M)$. Consider a pair (G_l, G_{l+1}) $(l = 1, \ldots, M-1)$.

At some execution of Step 1 or Step 2 some subgraph of $\mathbf{G}$, e.g., G_s, containing G_l and G_{l+1} is partitioned into $(\underline{G}_s, \bar{G}_s)$ with $G_l \subset \underline{G}_s$, $G_{l+1} \subset \bar{G}_s$. Note that $(\underline{G}_s \backslash G_l, G_l)$ and $(G_{l+1}, \bar{G}_s \backslash G_{l+1})$ are directed partitions of $\underline{G}_s$ and $\bar{G}_s$ respectively. Consider the following two cases:

(a) $(\underline{G}_s, \bar{G}_s)$ *is generated in some execution of step* 1. In view of Lemma 2(ii),

$$K(G_l)/H(G_l) \le \min(K(G_s)/H(G_s), b^2(G_s)) \le K(G_{l+1})/H(G_{l+1}).$$

Since $b(G_{l+1}) \ge b(G_s)$ we have

$$\min(K(G_l)/H(G_l), b^2(G_l)) \le \min(K(G_l)/H(G_l), b^2(G_s)) \le \min(K(G_{l+1})/H(G_{l+1}), b^2(G_{l+1}))$$

or $T(l) \le T(l+1)$.

(b) $(\underline{G}_s, \bar{G}_s)$ *is generated in some execution of step* 2. In this case $(\varnothing, G_s)$ is a minimum cut of $G'_s(T(l))$ with $T(l) = b(G_s)$, and $(\underline{G}_s, \bar{G}_s)$ is a minimum cut of $G^+_s(b(G_s))$ with $\underline{G}_s = G_l$. The former implies by Lemma 2(iii) and Lemma 3(ii), that

$$K(G_l)/H(G_l) \ge b^2(G_s) = b^2(G_l);$$

the latter implies by Lemma 3(iv) that

$$K(G_{l+1})/H(G_{l+1}) \ge b^2(G_s) = b^2(G_l).$$

Therefore by Lemma 3(ii)

$$\min((K(G_l/H(G_l))^{1/2}, b(G_l)) = b(G_l) \le \min((K(G_{l+1})/H(G_{l+1}))^{1/2}, b(G_{l+1})),$$

i.e., $T(l) \le T(l+1)$. □

The Second Stage: Rounding Up Procedure

Let $(G_1, \ldots, G_M)$ be the partition obtained by the first-stage procedure (Algorithm RCP) and let $\mathbf{T}$ be the associated solution vector. The following well-known rounding procedure transforms $\mathbf{T}$ into a power-of-two vector $\mathbf{T}^*$.

Rounding procedure. For all $l = 1, \ldots, M$, find the unique integer m_l such that

$$2^{m_l-(1/2)}T_L \le T(l) < 2^{m_l+(1/2)}T_L,$$

and set the common reorder interval for $N(G_l)$ as $T^*(l) = 2^{m_l}T_L$.

Note that the power-of-two vector $\mathbf{T}^*$ satisfies the capacity constraints since every component T_i, when increased, is rounded up to the smallest power-of-two value $\ge T_i$; hence $T^*_i \le b_i$ for all i. Also, when $T_i = b_i$, $T^*_i = T_i = b_i$. It is also easily verified that every nested vector $\mathbf{T}$ is transformed into a nested vector $\mathbf{T}^*$, thus maintaining feasibility.

The complexity of the two-stage algorithm is clearly determined by that of (RCP). To characterize the latter, note that each execution of Step 1 follows an increase of either M or l by one unit and Step 2 is executed M times. Since M cannot exceed $\mathbf{N}$, at most $(3\mathbf{N} - 1)$ maximum flow computations need to be performed. The complexity of the algorithm is thus given by $O(\mathbf{N}M_f)$ where M_f is the effort required for a single maximum flow computation. If the network $\mathbf{G}$ were dense, $M_f = O(\mathbf{N}^3)$, see, e.g., Malhotra et al. (1978). Below we argue, however, that the network $\mathbf{G}$ is *sparse* for both the nested and nonnested model considered.

Complexity of Algorithm RP for Nested Models. In the original production / distribution network the in-degrees and / or out-degrees of the nodes may be assumed to be uniformly bounded (by a small number usually), i.e., $A = O(N)$. Thus, the algorithm by Sleator and Tarjan (1983) or Goldberg and Tarjan (1986) for sparse networks is to be preferred, with

$$M_f = O(NA \log N) \quad \text{and} \quad M_f = O(NA \log(N/A))$$

respectively, i.e., $M_f = O(N^2 \log N)$. (See Ahuja et al. (1989) for a recent survey of efficient max-flow algorithms.)

Complexity of Algorithm RP for Nonnested Models. Since G is *bipartite* in this case, the same is true for all associated networks G'_l, see Definition 5. Some care is required when constructing the relaxed associated networks G^+_l (see Definition 6), in particular to ensure that these are bipartite as well. Since no direct arc(s, i) exists for any product $i \in N^0(G_l)$, one needs to relax the upperbound on the arc(s, r) for $r = \{i\}$. (If $d'_i = 0$, so that $r = \{i\} \notin R$, a node r and an uncapacitated arc(s, r) need to be added to the network.) We conclude that all associated and relaxed associated networks are bipartite so that special maximum flow algorithms by Gusfield et al. (1985) or Ahuja et al. (1988) may be invoked, with $M_f = O(RN^2)$ and $M_f = O(N\mathbf{A} + N^3)$ respectively. If the number of nodes per route is uni-

formly bounded, $\mathbf{A} = O(R)$ and the Ahuja et al. algorithm has $M_f = O(NR + N^3)$.

Note in addition that each of the subgraphs G_l ($l = 1, \ldots, M$) generated by Algorithm RCP must contain at least one product node, i.e., $M \le N$. To verify this, note first that when $(\underline{G}_l, \bar{G}_l)$ is a nontrivial cut of some associated network G_l' or relaxed associated network G_l^+, $\underline{G}_l$ must contain at least one product node, since the cut would otherwise be of infinite capacity. We conclude that in the final partition $(G_1, \ldots, G_M)$ generated by RCP, G_1 contains at least one product node. Thus, by Theorem 1,

$$0 < \min[b(G_1), (K(G_1)/H(G_1))^{1/2}]$$
$$\le \min[b(G_l), (K(G_l)/H(G_l))^{1/2}]$$

for all $l = 1, \ldots, M$ so that *each* of the disjoint sets G_l contains at least one product node. The overall complexity of the algorithm is thus $O(RN^3)$ or $O(N^2R + N^4)$ if $\mathbf{A} = O(R)$ and the Ahuja et al. (1988) max-flow algorithm is used. In other words, even when $R \gg N$, the complexity of the algorithm remains "linear" in R. (Recall that the generation of the route product network $\mathbf{G}$ requires $O(R)$ operations itself.) We have in fact:

THEOREM 3. *The two-stage algorithm, i.e.,* (RCP) *followed by the rounding procedure generates an optimal solution of* (CP).

The proof of Theorem 3 is similar to that given in Zheng (1987, Theorem 3.6) for uncapacitated models.

3. An Integrated Algorithm Based on a Characterization Theorem for the Model's Integer Program

In this section we derive a characterization theorem for optimal solutions of the integer program (GCP), based on which a direct solution procedure is derived.

THEOREM 4. *If the components of* $\mathbf{T}^*$ *take on M distinct power-of-two values* $T(1) < T(2) < \cdots < T(M)$, *and if* $(G_1, G_2, \ldots, G_M)$ *is a partition of* $\mathbf{G}$ such that $N(G_l) = \{i \in \mathbf{N}: T_i^* = T(l)\}$, *then* $\mathbf{T}^*$ *is an optimal solution of* (CP) *if and only if the following three conditions hold*:

(i) $(G_1, G_2, \ldots, G_M)$ *is a directed partition of* $\mathbf{G}$;

(ii) $T(l) = \min(b(G_l), T^*(l))$ *where* $T^*(l) = 2^{m_l} T_L$ (m_l *integer*); *and*

$$(1/\sqrt{2})(K(G_l)/H(G_l))^{1/2}$$
$$\le T^*(l) \le \sqrt{2}(K(G_l)/H(G_l))^{1/2}.$$

(iii) *For any directed partition* $(\underline{G}_l, \bar{G}_l)$ *of* G_l,

(a) $T(l) \le \sqrt{2}(K(\underline{G}_l)/H(\underline{G}_l))^{1/2}$;

(b) $\min(b(\bar{G}_l), (1/\sqrt{2})(K(\bar{G}_l)/H(\bar{G}_l))^{1/2}) \le T(l)$.

PROOF. We first prove the necessity part. Let $\mathbf{T}^*$ be an optimal solution of (GCP). (i) follows from the feasibility of $\mathbf{T}^*$. To prove (ii), assume to the contrary that for some $l = 1, \ldots, M$, $T(l) \ne \min(b(G_l), T^*(l))$. The feasibility of $\mathbf{T}^*$ implies $T(l) \le b(G_l)$. It suffices to consider the following two cases:

(1) $T(l) < \min(b(G_l), (1/\sqrt{2})(K(G_l)/H(G_l))^{1/2})$;

(2) $T(l) > \sqrt{2}(K(G_l)/H(G_l))^{1/2}$.

Let $C_l(t) \overset{\text{def}}{=} K(G_l)/t + H(G_l)t$.

It is easily verified that in case (1)

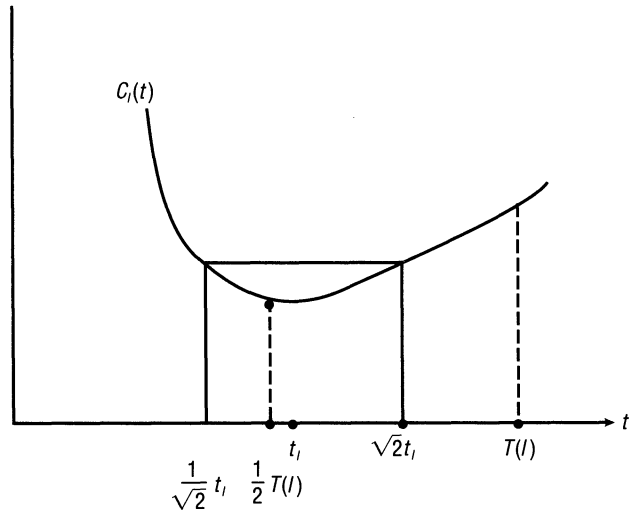
$$C_l(2T(l)) < C_l(T(l))$$

while in case (2)

$$C_l(\tfrac{1}{2}T(l)) < C_l(T(l))$$

(see Figure 3). In case (1), define $\mathbf{T}'$ by $T_i' = 2T_i^*$ ($i \in N(G_l)$), $T_i' = T_i^*$ ($i \notin N(G_l)$). $\mathbf{T}'$ is feasible

Figure 3 **Function** $C_l(t)$



(if $T(l) < b(G_l)$, $2T(l) \le b(G_l)$) with $C(\mathbf{T}^*) > C(\mathbf{T}')$, contradicting the optimality of $\mathbf{T}^*$.

In case (2) let $\mathbf{T}'$ be defined as $T_i' = \frac{1}{2}T_i^*$ $(i \in N(G_l))$, $T_i' = T_i^*$ $(i \notin N(G_l))$. $\mathbf{T}'$ is feasible with $C(\mathbf{T}^*) > C(\mathbf{T}')$, again contradicting the optimality of $\mathbf{T}^*$. A similar argument, doubling T_i for $i \in N(\bar{G}_l)$, if (b) fails and multiplying T_i by $1/2$ for $i \in N(\underline{G}_l)$, if (a) fails, establishes (iii).

We now prove the sufficiency part. Suppose that a vector $\mathbf{T}^*$ satisfies (i), (ii), (iii). Feasibility follows from (i) and (ii). To show optimality, define subproblems (CP_l) $(l = 1, \ldots, M)$:

$$(\mathrm{CP}_l) \quad C_l^* = \min_{\mathbf{T}>0} \sum_{i \in N(G_l)} (K_i/T_i + H_i T_i)$$

$$\text{subject to } T_i \ge T_j, \quad (i,j) \in A(G_l),$$

$$T_i = 2^{m_i} T_L, \quad i \in N(G_l),$$

$$T_i \le b_i, \quad i \in N(G_l).$$

Clearly $\sum_{l=1}^{M} C_l^*$ is a lower bound for the optimal value of (GCP). Thus it suffices to show that $\{T_i = T(l): i \in N(G_l)\}$ is an optimal solution for (CP_l), $l = 1, \ldots, M$. Assume there is an optimal solution $\mathbf{T}^{**}$ for (CP_l) such that its components take on distinct power-of-two values $T(l_1) < \cdots < T(l_a)$. Let $(G_{l_1}, \ldots, G_{l_a})$ be a partition of G_l such that

$$N(G_{l_r}) = \{i \in N(G_l): T_i^{**} = T(l_r)\}, \quad r = 1, \ldots, a.$$

We show that $T(l_1) \ge \frac{1}{2}T(l)$. By the proved necessity part of the present theorem,

$$T(l_1) = \min(b(G_{l_1}), T^*(l_1)),$$

which implies

$$T(l_1) \ge \min(b(G_{l_1}), (1/\sqrt{2})(K(G_{l_1})/H(G_{l_1}))^{1/2}). \tag{24}$$

Since $(G_{l_1}, G_l \backslash G_{l_1})$ is a directed partition of G_l, we have in view of (iii)

$$(K(G_{l_1})/H(G_{l_1}))^{1/2} \ge (1/\sqrt{2})T(l). \tag{25}$$

Note

$$b(G_{l_1}) \ge b(G_l) \ge T(l) > \tfrac{1}{2}T(l). \tag{26}$$

We have

$$T(l_1) \ge \min(b(G_{l_1}), \tfrac{1}{2}T(l)) = \tfrac{1}{2}T(l).$$

If $T(l_1) = \frac{1}{2}T(l)$, it follows from (26) that $b(G_{l_1}) > T(l_1)$ so that by (24) and (25),

$$T(l_1) \ge (1/\sqrt{2})(K(G_{l_1})/H(G_{l_1}))^{1/2}$$

$$\ge \tfrac{1}{2}T(l) = T(l_1), \quad \text{i.e.,}$$

$$\sqrt{2}T(l_1) = (K(G_{l_1})/H(G_{l_1}))^{1/2} = (1/\sqrt{2})T(l).$$

In this case, if we replace the common reorder interval $T(l_1)$ of $N(G_{l_1})$ by $T(l)$, the new solution $\mathbf{T}'$ is easily verified to remain feasible and the objective value remains the same, i.e., $\mathbf{T}'$ is optimal. We thus conclude that an optimal solution $\mathbf{T}'$ for (CP_l) exists with $(G_{l_1}, \ldots, G_{l_a})$ as the associated partition, and $T(l) \le T(l_1)$. We now show that an optimal solution $\mathbf{T}''$ of (CP_l) exists with

$$T(l) \le T(l_1) \le \cdots \le T(l_a) \le T(l).$$

From the necessity part of this very theorem,

$$T(l_a) \le \min(b(G_l), \sqrt{2}[K(G_{l_a})/H(G_{l_a})]^{1/2}),$$

see (ii). In view of (iii),

$$\min(b(G_{l_a}), (1/\sqrt{2})[K(G_{l_a})/H(G_{l_a})]^{1/2}) \le T(l)$$

since $(G_l \backslash G_{l_a}, G_{l_a})$ is a directed partition of G_l.

Thus

$$(1/\sqrt{2})T(l_a) \le \min(b(G_{l_a}), (K(G_{l_a})/H(G_{l_a}))^{1/2})$$

$$\le \sqrt{2}T(l) \tag{27}$$

or $T(l_a) \le 2T(l)$. $T(l_a) = 2T(l)$ holds only if the inequalities in (27) hold as equalities. In this case,

$$b(G_{l_a}) \ge (K(G_{l_a})/H(G_{l_a}))^{1/2}.$$

(Otherwise

$$\min(b(G_{l_a}), [K(G_{l_a})/H(G_{l_a})])^{1/2} = b(G_{l_a}) = \sqrt{2}T(l)$$

but the equality $b(G_{l_a}) = \sqrt{2}T(l)$ cannot hold since both $b(G_{l_a})$ and $T(l)$ are power-of-two multiples of T_L.) Therefore

$$(1/\sqrt{2})T(l_a) = (K(G_{l_a})/H(G_{l_a}))^{1/2} = \sqrt{2}T(l).$$

One now verifies as above that an alternative optimal solution exists with $T(l) \le T(l_1) \le T(l_a) \le T(l)$. □

The above characterization theorem suggests the following integrated algorithm: Let $\tau_0 = (1/\sqrt{2})\,b(G)$.

Algorithm DIRECT (C)

Step 0. Let $\tau := \tau_0$. Find a min cut (G_1, G_2) of $G'(\tau)$. $M := 2$.

Step 1. If $G_1 = \varnothing$, then *begin* $\tau := 2\tau_0$, go to Step 2 *end*; $\tau := \tau/2$; find a min cut $(\underline{G}_1, \bar{G}_1)$ of $G'(\tau)$. Rename $(\underline{G}_1, \bar{G}_1, G_2, \ldots, G_M)$ as $(G_1, G_2, \ldots, G_{M+1})$. $M := M + 1$. $T_i^* := \sqrt{2}\tau$, $i \in N(G_2)$. Go back to Step 1.

Step 2. If $G_M = \varnothing$, then stop; otherwise, if $(1/\sqrt{2})\tau < b(G_M)$ then $G'_M := G'_M(\tau)$ else $G'_M := G^+_M(\tau)$. Find a min cut $(\underline{G}_M, \bar{G}_M)$ of G'_M. Rename $(G_1, \ldots, \underline{G}_M, \bar{G}_M)$ as $(G_1, \ldots, G_M, G_{M+1})$. Let $M := M + 1$. $T_i^* := (1/\sqrt{2})\tau$, $i \in N(G_{M-1})$. Go back to Step 2.

THEOREM 5. *The vector* $\mathbf{T}^*$ *generated by algorithm* DIRECT(C) *is an optimal solution of* (GP).

PROOF. It suffices to show that conditions (i), (ii) and (iii) of Theorem 4 are satisfied. The fact that $(G_1, \ldots, G_M)$ is a directed partition follows by complete induction using Lemma 2(i) and Lemma 3(i). We show (ii) and (iii) for those sets G_m that were generated in Step 1 first. For fixed m, consider the iteration in which G_m is generated. In this iteration a min cut $(\underline{G}_1, \bar{G}_1)$ of $G'_1(\tau)$ is found where $\bar{G}_1$ becomes G_m at the end of the algorithm and $\tau = (1/\sqrt{2})T(m)$. Note that $\tau \le (1/\sqrt{2})b(\mathbf{G}) \le (1/\sqrt{2})b(G_m)$, which implies

$$T(m) \le b(G_m). \tag{28}$$

Moreover, it follows from Lemma 2(ii) that

$$K(G_m)/H(G_m) \ge \tau^2 = T^2(m)/2 = 2^{2m-1}T_L. \tag{29}$$

Observe that the set G_1 is generated either in the previous execution of Step 1, or in the one time execution of Step 0, i.e., there exists a set $\bar{G}$, such that $(G_1, \bar{G})$ is a minimal cut of $(G_1 \cup \bar{G})'(2\tau)$. Applying Lemma 2(iii) we conclude that

$$K(G_m)/H(G_m) \le (2\tau)^2 = 2T^2(m) = 2^{2m+1}T_L. \tag{30}$$

Formula (28), (29) and (30) imply (ii). For any directed partition $(\underline{G}_m, \bar{G}_m)$, (28), (29) and (30) imply (ii). For any directed partition $(\underline{G}_m, \bar{G}_m)$ of G_m, we have from Lemma 2(iii)

$$(K(\underline{G}_m)/H(\underline{G}_m))^{1/2} \ge (1/\sqrt{2})T(m), \tag{31}$$

$$(K(\bar{G}_m)/H(\bar{G}_m))^{1/2} \le \sqrt{2}T(m). \tag{32}$$

(31) and (32) imply (iii).

Next we show (ii) for those G_k that were generated in step 2. Fix k and consider the iteration in which G_k was generated. G_k is, in that iteration, $\underline{G}_M$ in a min cut $(\underline{G}_M, \bar{G}_M)$ of G'_M, where $G'_M = G'_M(\tau)$ or $G'_M = G^+_M(\tau)$ and $\tau = \sqrt{2}T(k)$. On the other hand the subgraph G_M itself is generated either in the previous execution of Step 2 or in the one time execution of Step 0. In either case, applying Lemma 2(iii) or Lemma 3(iv) we have

$$(K(G_k)/H(G_k))^{1/2} \ge \tfrac{1}{2}\tau = (1/\sqrt{2})T(k). \tag{33}$$

We distinguish between the following two cases:

Case (1). $G'_M = G'_M(\tau)$. In this case we have

$$b(G_k) \ge b(G_M) > (1/\sqrt{2})\tau = T(k) \tag{34}$$

and in view of Lemma 2(ii)

$$(K(G_k)/H(G_k))^{1/2} \le \tau. \tag{35}$$

(33) and (35) imply

$$(1/\sqrt{2})(K(G_k)/H(G_k))^{1/2} \le T(k) \le \sqrt{2}(K(G_k)/H(G_k))^{1/2}. \tag{36}$$

In view of (34), (ii) is verified.

Case (2). $G'_M = G^+_M(\tau)$. In this case we have in view of Lemma 3(ii)

$$b(G_k) = b(G_M) \le (1/\sqrt{2})\tau. \tag{37}$$

To prove (ii), it suffices in view of (33) to show that $T(k) = b(G_k)$. Due to (37), we only need to show that $b(G_k) > \frac{\tau}{2}$. Since $b(G_k) = b(G_M)$, it suffices to show that $b(G_M) > \frac{\tau}{2}$. We prove this by complete induction. For $M = 2$, $b(G_M) \ge b(G) = \sqrt{2}\tau_0 > \tau_0$. Assume that $b(\underline{G} \cup G_M) \ge \frac{\tau}{4}$ holds. We prove $b(G_M) > \frac{\tau}{2}$. If $(\underline{G}, G_M)$ is a minimum cut of $(\underline{G} \cup G_M)'(\frac{\tau}{2})$, we have

$$b(G_M) \ge b(\underline{G} \cup G_M) > (1/(2\sqrt{2}))\tau,$$

i.e., $b(G_M) \ge \frac{\tau}{2}$. If $(\underline{G}, G_M)$ is a minimum cut of $(\underline{G} \cup G_M)^+(\frac{\tau}{2})$, we have, in view of Lemma 3(ii), $b(G_M) > b(\underline{G})$. Combining this inequality with $b(G_{k-1}) \ge \frac{\tau}{4}$, we have $b(G_M) > \frac{\tau}{2}$.

Finally we show (iii) for G_k. Let $(\underline{G}_k, \bar{G}_k)$ be any directed partition of G_k. (iii)(a) is true in view of Lemma 3(iv) and Lemma 2(iii). In case $G'_M = G'_M(\tau)$, we have in view of Lemma 2(iii)

$$(K(\bar{G}_k)/H(\bar{G}_k))^{1/2} \le \tau. \tag{38}$$

In case $G'_M = G^+_M(\tau)$, we have in view of Lemma 3(iv)

$$(K(\bar{G}_k)/H(\bar{G}_k))^{1/2} \leq \tau \quad \text{or} \quad b(\bar{G}_k) = b(G_k) = T(k). \tag{39}$$

(38) and (39) imply (iii)(b). □

Complexity of DIRECT (C)

The complexity of the DIRECT (C) algorithm is $O(M_f L)$ with L the number of possible distinct reorder intervals in the optimal power-of-two vector. We refer to §2 for a discussion of the magnitude of M_f in various special cases (in particular in nested versus nonnested models). In practice, L is a small number; $L \leq 10$ (say). Thus it is reasonable to assume that $L = O(1)$. Under this assumption the DIRECT algorithm achieves an order of magnitude efficiency improvement over the two-stage procedure.

4. Algorithms for the Uncapacitated Model

In this section we describe how the algorithms (RCP) and DIRECT (C) simplify when applied to the uncapacitated model. We refer to the simplified version as RP and DIRECT respectively.

Algorithm RP

Step 0. $M := 1$, $l := 1$, $G_1 := \mathbf{G}$.

Step 1. $T = (K(G_l)/H(G_l))^{1/2}$. Find a maximum flow in $G'_l(T)$; if $(G_l, \varnothing)$ is a minimum cut, go to Step 2. Otherwise, we have a nontrivial minimum cut of G_l, $(\underline{G}_l, \bar{G}_l)$. Renumber $(\underline{G}_l, \bar{G}_l, G_{l+1}, \ldots, G_M)$ as $(G_l, G_{l+1}, \ldots, G_{M+1})$; $M := M + 1$ and repeat Step 1.

Step 2. If $l = M$, stop; $(G_1, G_2, \ldots, G_M)$ is the desired partition; otherwise $l := l + 1$ and go back to Step 1.

Algorithm DIRECT

Let $\tau_0 = (1/\sqrt{2})(K(G)/H(G))^{1/2}$.

Step 0. Let $\tau := \tau_0$. Find a min cut (G_1, G_2) of $G'(\tau)$; $M := 2$.

Step 1. Same as *Step* 1 in Algorithm DIRECT (C).

Step 2. If $G_M = \varnothing$ then stop; otherwise, find a min cut $(\underline{G}_M, \bar{G}_M)$ of $G'_M(\tau)$. Rename $(G_1, \ldots, \underline{G}_M, \bar{G}_M)$ as $(G_1, \ldots, G_M, G_{M+1})$; Let $M := M + 1$. $T^*_i := (1/\sqrt{2})\tau$, $i \in N(G_{M-1})$. Go back to Step 2.

The worst case complexity of the algorithms does not reduce below the bounds identified in §§2 and 3. As

Figure 4 **Optimal Partition for (RP)**

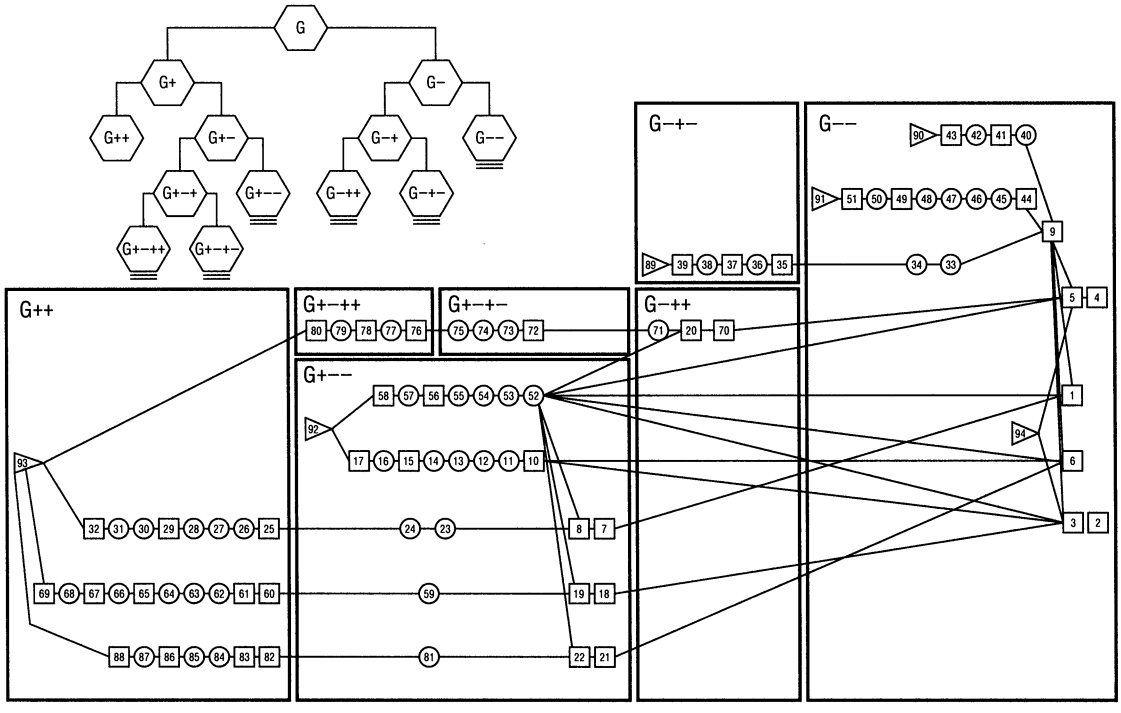
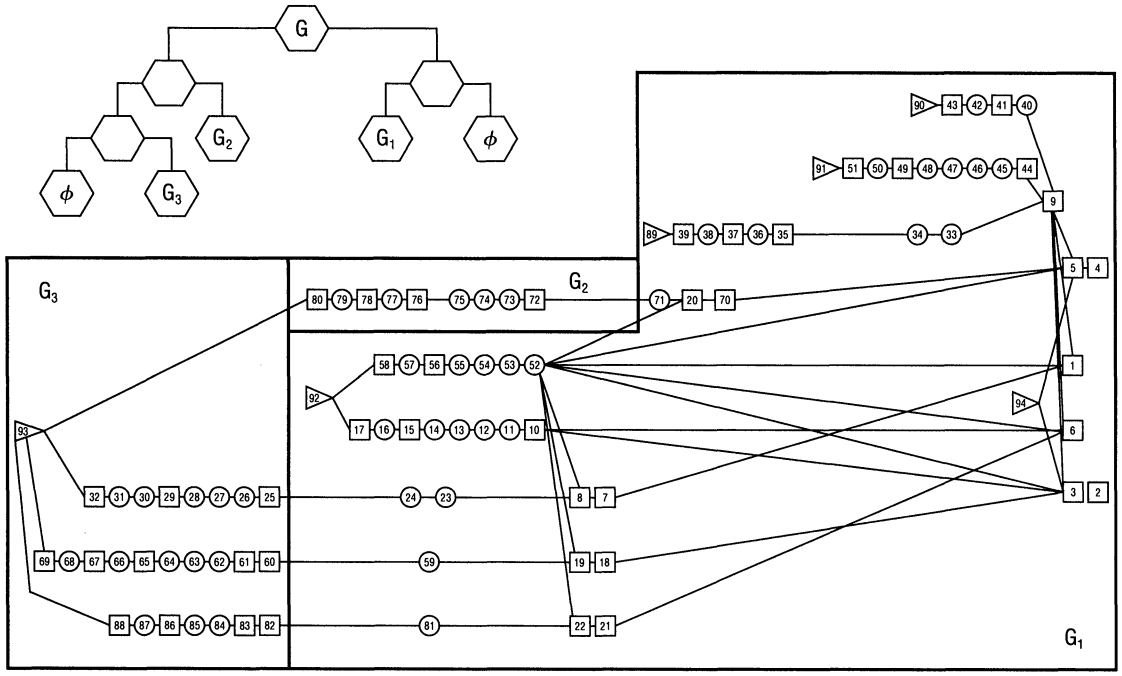


Figure 5 Optimal Solution Generated by DIRECT



discussed in the introduction, these complexity bounds compare favorably with the best existing alternatives.

We conclude this paper with an application of the DIRECT algorithm to the example model in Maxwell and Muckstadt (1985). This model represents a production system of a major U.S. automobile manufacturer with 94 distinct operations, i.e., $N = 94$. As pointed out above, Maxwell and Muckstadt restrict themselves to nested policies and report that their algorithm requires 13 iterations. In each a network flow problem is solved, the complexity of which is equivalent to that of a maximum flow computation. The partition corresponding to the obtained optimal solution of (RP) is depicted in Figure 4; it consists of seven node sets. An optimal power-of-two policy, however, uses only 3 distinct reorder intervals, 128 days, 64 days and 32 days. The DIRECT algorithm finds this solution in 4 iterations only. The solution is shown in Figure 5.[1]

[1] We wish to thank two anonymous referees for their most helpful comments.

Appendix 1. Lemma A.1

LEMMA A.1. *Consider the setup cost structure described by* (1) *and assume* $b_i = 2^{n_i} T_L$ (n_i *integer*) *and* $b_i \geq b_j$ *for all* $(i, j) \in A(G)$ *where* $b_i = B_i / d_i$. *There exists an optimal nested power-of-two policy* $\mathbf{T}$ *with* $T_i \leq b_i$ *for all* $i \in N(G)$.

PROOF. Note that the long-run average costs of an arbitrary power-of-two policy $\mathbf{T}$ is given by

$$\sum_{i \in N(G)} K_i(d_i T_i)/T_i + \sum_{i \in N(G)} H_i T_i.$$

Let the vector $[T_N^*, T_{N-1}^*, \ldots, T_1^*]$ be the lexicographically smallest optimal nested power-of-two vector. Assume to the contrary that $T_i^* > b_i$ for some $i \in N(G)$ and let n be the largest indexed component of $\mathbf{T}^*$ for which this is the case. Let $\mathbf{T}'$ be defined by $T_n' = b_n$ and $T_i' = T_i^*$ otherwise. Clearly $T_i' \geq T_n'$ for all $(i, n) \in A(G)$ while for $(n, j) \in A(G)$, $T_n' = b_n \geq b_j \geq T_j^* = T_j'$. Thus, $\mathbf{T}'$ is feasible. In view of (A1), its average setup costs are identical to those of $\mathbf{T}'$ while its average holding costs are no larger than those of $\mathbf{T}'$. We conclude that $\mathbf{T}'$ is optimal and lexicographically smaller than $\mathbf{T}^*$. □

Appendix 2. A Lower Bound Theorem for Capacitated Models

In this appendix, we show that in complete similarity to uncapacitated problems, the minimum value of the continuous relaxation (RCP)

constitutes a lower bound for the minimum average cost achievable by any feasible policy, i.e., any policy which orders product i with a frequency less than or equal to b_i^{-1} $(i \in N)$.

Even though in nonnested models the general formulation of (CP) allows for specific upper bounds on the auxiliary variables $\{T_r: r \in R\}$, in addition to upper bounds on the products' replenishment intervals $\{T_i: i \in N(G)\}$, we assume in Theorem A2 below that only the latter upper bounds prevail. (This is the case in all of the examples mentioned in §1.)

THEOREM A2. LOWER BOUND THEOREM. *Consider the general nonnested model, and assume the system starts with zero inventory. The minimum value c^* of* (RCP) *is a lower bound for the average cost of any feasible policy over any finite horizon $[0, \tau)$, i.e., any policy which replenishes product i's inventory at least $b_i^{-1}\tau$ times $(i \in N(G), \tau > 0)$.*

PROOF. Fix $\tau > 0$ and a feasible policy with c as its average cost in $[0, \tau)$. For the general nonnested model (CD) can be written as

$$(\mathrm{CD}_1) \quad \max D(\lambda, \mathbf{x}, \mathbf{v}) = \sum_{i \in N(G)} (2(K_i v_i)^{1/2} - b_i \lambda_i)$$

$$\text{subject to } \sum_{i \in r} x_{ri} = H_r, \quad r \in R,$$

$$\sum_{r: i \in r} x_{ri} = v_i - \lambda_i, \quad i \in N(G),$$

$$\lambda \geq 0, \quad \mathbf{x} \geq 0, \quad \mathbf{v} \geq 0.$$

Let $(\lambda^*, \mathbf{x}^*, \mathbf{v}^*)$ be an optimal solution of (CD_1). Consider a given policy, time $t > 0$, and route $r = (i_1, \ldots, i_m)$. Define, as in Roundy (1986), route r's echelon inventory E_r^t as the total number of units of product i_1 which are held in stock somewhere along the route r at time t (perhaps as components of more advanced products) and which have been specified to follow route r, measured in multiples of $\frac{1}{2}d_r$, i.e., as the number of time units of demand for route r's (unique) end item which this inventory is capable of supporting. For each $i \in N(G)$ and $t > 0$ define

$$n_i^t = \min\{E_r^t: r \in R_i\}.$$

We have

$$\sum_{r \in R} H_r E_r^t = \sum_{r \in R} \sum_{i \in r} x_{ri}^* E_r^t = \sum_{i \in N(G)} \sum_{r: i \in r} x_{ri}^* E_r^t$$

$$\geq \sum_{i \in N(G)} \sum_{r: i \in r} x_{ri}^* n_i^t = \sum_{i \in N(G)} (v_i^* - k_i^*) n_i^t.$$

The inequality follows since for any route $r = (i_1, \ldots, i, \ldots, i_m) \in i$, $n_i^t \leq E_{r'}^t \leq E_r^t$ where $r' = (i, \ldots, i_m)$. Let J_i denote the number of times product i is ordered in $[0, \tau]$. Thus,

$$\tau c \geq \sum_{i \in N(G)} (K_i J_i + \int_0^\tau (v_i^* - \lambda_i^*) n_i^t \, dt$$

$$\geq \sum_{i \in N(G)} (K_i J_i + (v_i^* - \lambda_i^*)\tau^2 / J_i)$$

$$\geq \sum_{i \in N(G)} \min\{K_i z_i + (v_i^* - \lambda_i^*)\tau^2 / z_i: z_i \geq b_i^{-1}\tau\}$$

$$= \sum_{i \in N(G)} \tau \min\{K_i / t_i + (v_i^* - \lambda_i^*) t_i: t_i \leq b_i\}$$

$$\geq \sum_{i \in N(G)} (2(K_i v_i^*)^{1/2} - \lambda_i^* b_i) = \tau c^*$$

where the first inequality holds in view of $v_i^* - \lambda_i^* \geq 0$ $(i \in N(G))$. We show the last inequality as follows: Let $\mathbf{T}^*$ be an optimal solution to (RCP) and let $(G_1, \ldots, G_M)$ be a corresponding partition of the node set $N(G)$, i.e., $N(G_l)$ is the set of nodes which share the l-the smallest replenishment interval. As in the proof of Theorem 1, let $L_1 = \{l: 1 \leq l \leq M \text{ and } K(G_l)/H(G_l) \leq b^2(G_l)\}$ and $L_2 = \{1, \ldots, M\} \setminus L_1$. For $i \in N(G_l)$ with $l \in L_1$, $\lambda_i^* = 0$ (see (23)), hence,

$$K_i / t_i + (v_i^* - \lambda_i^*) t_i = K_i / t_i + v_i^* t_i \geq 2(K_i v_i^*)^{1/2} - \lambda_i^* b_i.$$

For $i \in N(G_l)$ with $l \in L_2$, we have $v_i^* = K_i / b_i^2$, see (31). Note that $\lambda_i^* \geq 0$ implies that $K_i/(v_i^* - \lambda_i^*) = K_i/(K_i/b_i^2 - \lambda_i^*) \geq b_i^2$.

Therefore, since the function $K_i/t_i + (v_i^* - \lambda_i^*)t_i$ is nondecreasing for $t_i \leq (K_i/(v_i^* - \lambda_i^*))^{1/2}$, we have for $t_i \leq b_i$:

$$K_i / t_i + (K_i / b_i^2 - \lambda_i^*) b_i \geq 2K_i / b_i - \lambda_i^* b_i = 2(K_i v_i^*)^2 - \lambda_i^* b_i. \quad \square$$

Appendix 3. Proof of Lemmas 2 and 3

PROOF OF LEMMA 2.

(i) Immediate from the fact that in $G'(\tau)$ all arcs in $A(G)$ have infinite capacities.

(ii) The capacity of the minimum cut (G_1, G_2) is no larger than that of the cut $(\varnothing, G_l)$. Thus $H(G_2) + K(G_1)/\tau^2 \leq H(G_l)$ or $K(G_1)/H(G_1) \leq \tau^2$. Similarly, since $(G_l, \varnothing)$ is also a cut, $H(G_2) + K(G_1)/\tau^2 \leq K(G_2)/\tau^2$ or $\tau^2 \leq K(G_2)/H(G_2)$.

(iii) Let $\tau^2 = K(G_l)/H(G_l)$. $K(G_2)/H(G_2) \leq \tau^2 \leq K(G_1)/H(G_1)$ for *every directed* partition (G_1, G_2) if the cut $(\varnothing, G_l)$ or $(G_l, \varnothing)$ is minimal in $G_l'(\tau)$.

(iv) The capacity of the minimum cut (G_1, G_2) is no larger than that of the cut $(\underline{G}_1, G_2 \cup \bar{G}_1)$. Thus

$$H(G_2) + K(G_1)/\tau^2 \leq H(G_2 \cup \bar{G}_1) + K(\underline{G}_1)/\tau^2,$$

or $K(\bar{G}_1)/\tau^2 \leq H(\bar{G}_1)$ or $K(\bar{G}_1)/H(\bar{G}_1) \leq \tau^2$. The second part of (iii) follows similarly from the observation that the capacity of (G_1, G_2) is no larger than that of $(G_1 \cup \underline{G}_2, \bar{G}_2)$. $\square$

PROOF OF LEMMA 3.

(i) Note $(G_l, \varnothing)$ is a cut with finite capacity, while any undirected partition of G_s has infinite capacity.

(ii) If $N(G_2) \cap N^0(G_l) \neq \varnothing$, the capacity of (G_1, G_2) is infinite.

(iii) The capacity of $(G, \varnothing)$ is no smaller than that of (G_1, G_2), i.e., $K(G_l)/\tau^2 \geq H(G_2) + K(G_1)/\tau^2$ or $K(G_2)/H(G_2) \geq \tau^2$.

(iv) Assume $\underline{G}_1 \cap N^0(G_l) = \varnothing$, i.e., $b(\bar{G}_1) \neq b(G_l)$. The capacity of $(\underline{G}_1, \bar{G}_1 \cup G_2)$ is not smaller than that of (G_1, G_2), i.e.,

$$H(\bar{G}_1 \cup G_2) + K(\underline{G}_1)/\tau^2 \geq H(G_2) + K(G_1)/\tau^2,$$

i.e., $K(\bar{G}_1)/H(\bar{G}_1) \leq \tau^2$. Similarly the capacity of $(G_1 \cup \underline{G}_2, \bar{G}_2)$ is no smaller than that of (G_1, G_2). Since $\bar{G}_2 \subset G_2$, we conclude in view of (ii) that

$$K(G_1 \cup \underline{G}_2)/\tau^2 + H(\bar{G}_2) \geq K(G_1)/\tau^2 + H(G_2)$$

or $K(\underline{G}_2)/H(\underline{G}_2) \geq \tau^2$. □

References

Ahuja, R., T. Magnanti and J. Orlin, "Network Flows," In *Handbooks in OR & MS, Vol. I*, G. L. Nemhauser et al. (Eds.), 1989.

——, J. Orlin, C. Stein and R. Tarjan, "Improved Algorithms for Bipartite Network Flow Problems" (to appear).

Aucamp, D., "Nonlinear Freight Costs in the EOQ Problem," *European J. of Operations Res.*, 9 (1982), 61–63.

Federgruen, A. and H. Groenevelt, "Optimal Flow in Networks with Multiple Sources and Sinks, with Applications to Oil and Lease Investment Programs," *Operations Res.*, 34 (1986), 218–225.

—— and ——, "Polymatroidal Flow Network Models with Multiple Sinks," *Networks*, 18 (1987), 285–302.

Fujishige, S., "Lexicographically Optimal Base of a Polymatroid with Respect to a Weight Vector," *Mathematics of Operations Res.*, 5 (1980), 186–196.

Geoffrion, A., "Duality in Nonlinear Programming: A Simplified Applications Oriented Development," *SIAM Review*, 13, 1 (1971), 1–37.

Goldberg, A. and R. Tarjan, "A New Approach to the Maximum Flow Problem," *Proceedings of the 18th ACM Symposium on The Theory of Computing*, 1986, 136–146.

Groenevelt, H., "Two Algorithms for Maximizing a Separable Concave Function over a Polymatroid Feasible Region," *European J. Operational Res.*, 54, 2 (1991) 227–236.

Gusfield, D., C. Martel and D. Fernandez-Baca, "Fast Algorithms for Bipartite Network Flow," Working Paper, Department of Computer Science, Yale University, 1985.

Jackson, P. L., W. L. Maxwell and J. A. Muckstadt, "Determining Optimal Intervals in Capacitated Production-Distribution Systems," *Management Sci.*, 35 (1988), 938–958.

Joneja, D., "Planning for Joint Replenishment and Assembly Systems with Deterministic Non-Stationary Demands," Ph.D. Dissertation, School of ORSIE, Cornell University, Ithaca, NY, 1989.

Lippman, S., "Optimal Inventory Policy with Multiple Setup Costs," *Management Sci.*, 16 (1969), 118–138.

-, "Economic Order Quantity and Multiple Setups," *Management Sci.*, 18 (1971), 39–47.

Malhotra, V. M., M. Pramodh Kumar and S. N. Maheshwari, "An $O(v3)$ Algorithm for Finding the Maximum Flow in a Network," *Inform. Process. Lett.*, 7 (1978), 277–278.

Maxwell, W. L. and J. A. Muckstadt, "Establishing Consistent and Realistic Reorder Intervals in Production-Distribution Systems," *Operations Res.*, 33 (1985), 1316–1341.

Muckstadt, J. A., "Planning Component Delivery Intervals in Constrained Assembly Systems," In *Multi-Stage Production Planning and Inventory Control*, S. Axsäter, et al. (Eds.), Springer-Verlag, Berlin, 1985.

Roundy, R., "A 98% Effective Lot-Sizing Rule for a Multi-Product, Multi-Stage Production Inventory Systems," *Mathematics of Operations Res.*, 11 (1986), 699–727.

Sleator, D. and R. Tarjan, "A Data Structure for Dynamic Trees," *J. Comput. Systems Sci.*, 24 (1983), 362–391.

Swoveland, C., "A Deterministic Multi-period Production Planning Model with Piecewise Concave Production and Holding Costs," *Management Sci.*, 21 (1975), 1007–1013.

Veinott, A. F., Jr., "Least d-Majorized Network Flows with Inventory and Statistical Applications," *Management Sci.*, 17, 9 (1971), 547.

Zheng, Y. S., "Replenishment Strategies for Production / Distribution Networks with General Joint Setup Costs," Ph.D. Dissertation, Columbia University, New York, 1987.

Accepted by Stephen C. Graves; received November 20, 1991. This paper has been with the authors 2 months for 2 revisions.