Implementing Supply-Routing Optimization in a Make-to-Order Manufacturing Network

John Foreman  
Booz Allen Hamilton, Atlanta, Georgia 30303, foreman_john@bah.com

Jérémie Gallien  
MIT Sloan School of Management, Cambridge, Massachusetts 02142, jgallien@london.edu

Julie Alspaugh, Fernando Lopez  
Dell Computers, Inc., Round Rock, Texas 78682 {julie_alspaugh@dell.com, fernando_m_lopez@dell.com}

Rohit Bhatnagar, Chee Chong Teo  
Nanyang Technological University, Singapore 639798 {arbhatnagar@ntu.edu.sg, teocc@ntu.edu.sg}

Charles Dubois  
HEC Paris, 78352 Jouy en Josas, France, charles.dubois@mailnec.net

Dell’s supply chain for desktops involves Asian vendors shipping components by sea to several U.S. plants. Although suppliers are responsible for shipping enough inventory to meet total needs across all production sites, Dell can reroute and expedite their shipments while in transit, and also transfer on-hand inventory in order to balance supply across sites. This paper describes the development, implementation, and impact of the process and optimization-based control system now used by Dell to address this supply-routing challenge for its U.S.-bound monitors. In a first phase, Dell created a new job definition focused solely on supply routing and implemented a supporting visualization tool. In a second phase, a decision support system relying on a mixed-integer programming formulation was implemented, overcoming two main challenges: (i) the estimation of shortages as a function of expected inventory, accounting for actual forecast quality; and (ii) the estimation of a meaningful shortage cost. This new methodology is estimated to have reduced Dell’s inventory-repositioning costs for monitors by about 60%.

Key words: inventory theory and control; supply chain management; empirical research; logistics and transportation

History: Received: August 11, 2008; accepted: November 8, 2009. Published online in Articles in Advance April 9, 2010.

1. Introduction

Dell’s growth over the last 10 years has coincided with a significant increase in the complexity of its operations. For its North American desktop division, this evolution has specifically taken the following forms: (i) increase in the number of assembly plants and warehouse facilities; (ii) replacement of most U.S.-based suppliers with Asian suppliers; and (iii) increasing variety of products offered to customers. Although these changes have directly impacted most of Dell’s operational functions, they have in particular substantially complicated the task of its procurement group. Indeed, this group has thus become responsible for maintaining the availability of more components in more locations, working with suppliers having longer transportation lead times.

To address this supply availability challenge, Dell has long relied on vendor-managed inventory (VMI) relationships. That is, its suppliers are expected to maintain sufficient inventory of components in each of Dell’s relevant locations, based on a demand forecast periodically communicated by Dell, e.g., 14 days of supply in inventory (DSI). As part of that relationship, component inventory continues to be owned by suppliers until only a couple of hours before that inventory is pulled by Dell’s assembly lines (or ware-
house pick process), and suppliers are mostly free to follow any schedule of shipments as long as it meets some minimum service level. To benefit from transportation volume discounts however, these shipments are typically sent by ocean and air carriers directly contracted by Dell. Also, Dell centralizes inventory and shipment information, in part because it often uses several suppliers for the same component. As a result, Dell has retained the function of managing both the routing of its pipeline inventory and the transshipments of its on-hand inventory between various facilities (supply routing), regardless of that inventory’s ownership (see Reyner 2006 and Kapuscinski et al. 2004 for more background and references on Dell’s business model, supply chain, and history).

The supply-routing function just defined is particularly important for components such as desktop chassis and monitors, which account for a substantial proportion of total supply transportation costs. These components are shipped by ocean from Asia to the United States in full containers of a single part type because of their large volume and weight. As a result, gaps between actual realized demand in each assembly or warehouse facility and the forecasts driving these shipments can become quite large over this transportation delay. This may cause large imbalances in the inventory positions of Dell’s various sites, and may in turn lead to customer delivery delays due to component shortages as well as additional inventory holding costs. To mitigate these problems, Dell can change, at some cost, the final destination of any container still in transit on the ocean (diversion) as well as its planned ground transportation mode (expediting) up until a couple of days before it is disembarked in Long Beach, CA. The available ground transportation modes include, with increasing cost and decreasing lead time, the default rail and truck mode; a single-driver truck-only mode; and a two-driver (team) truck-only mode. In addition, Dell can also perform transshipments (transfers) of on-hand inventory between its facilities. The available transportation modes for transfers include a set schedule of precontracted truck “milk runs” between Dell U.S. facilities that have low relative cost, but limited capacity, as well as special single or team trucks contracted on the spot. Following the terminology used by Dell and its carriers, we define milk runs, special single-driver trucks, and team trucks as different transportation modes, even though they all actually rely on the same type of transportation vehicle.

Figure 1  Dell’s Supply Chain Structure and Supply-Routing Decisions for U.S.-Bound Desktop Chassis and Monitors

---

1 Supplier DSI targets in individual locations constitute widely followed but noncontractual operational guidelines. Contractual supplier obligations only relate to the sum of all inventory provided across all of Dell’s facilities over time. Failure to meet these obligations results in negotiation of liability for any additional costs incurred by Dell (see Tsay 1999 for a relevant discussion of manufacturing supply contracts).
structure and the associated supply-routing decisions just defined, and also shows the four main locations in the United States where chassis and monitors are shipped for assembly and/or inclusion into customer orders as well as typical transportation lead times.

The volume of material continuously going through the supply chain just described is very significant: a rough estimate from Dell’s public 10K filing for fiscal year 2008 reveals that tens of thousands of units of each component type must have been shipped every week, on average, to Dell U.S. facilities over that period. The challenge of making all the associated diversion, expediting, and transfer decisions in a timely and cost-effective manner thus constitutes a supply chain control problem that is both difficult and important: Although we are not at liberty to provide any specific financial information here, it is reported, for example, in the public thesis of Reyner (2006) that “A presentation was given to Dell’s management in the second quarter of 2006 that quoted expedite costs in the tens of millions per quarter […]” (p. 27). The present paper summarizes the collaboration between Dell and university researchers over several years to develop the optimization-based control system now used by Dell to address this supply-routing challenge for its U.S.-bound monitors. It contributes to the operations management literature by providing a detailed description of a real-world control challenge that has not been discussed extensively so far, even though it is critical to the operation of an important supply chain. It also describes a model for addressing this challenge, along with an implementation process, which have both been tested and validated by practice.

The remainder is organized as follows. After a discussion of the related literature in §2, we present the two main successive phases of that collaboration in §§3 and 4. As described in §3.1, the first phase consisted of creating a new position referred to as supply-routing analyst, who was solely dedicated to the inventory-routing decisions described above. It also included the development of a spreadsheet-based visualization tool in support of that role. The associated implementation issues are discussed in §3.2, and the resulting impact in §3.3. The second phase involved the development of a more sophisticated decision support system relying on a mixed-integer program (MIP) solved independently for every monitor type over a rolling horizon, as described in §4.1. The main modeling challenge encountered at that stage consisted of embedding into this optimization problem formulation a function representing the expected shortage costs resulting from the routing decisions considered. This function thus captures Dell’s actual demand uncertainty in a model that is otherwise deterministic. Next, §4.2 discusses the main challenges we overcame to implement the model, in particular, the determination of a sensible value for the unit shortage cost. The financial impact of this work is then discussed in §4.3, which also explores the qualitative differences between the supply-routing decisions determined by the analyst and those recommended by the optimization model. Finally, §5 contains concluding remarks pertaining to the limitations of our work, ongoing related developments, possible future research, and key learnings from this collaboration. An important notational convention used throughout this paper consists of using symbols in bold for random variables, and the same symbols with no highlight for their mathematical expectations, e.g., $\bar{d} \triangleq \mathbb{E}[d]$. Also, notations with an upper bar denote cumulative quantities, e.g., $\bar{d}_t = \sum_{k=1}^t d_k$. We use throughout the following cost terminology: Supply transportation costs refer to the sum of all costs that Dell incurs directly or indirectly to transport its components from the supplier location to the plant where they will be assembled into (or, in the case of monitors, packed with) a computer. They comprise embedded transportation costs and repositioning transportation costs, from which we omit the word “transportation” when no ambiguity arises. Embedded costs correspond to the default transportation mode for a given component type (for monitors, transit by ocean from Asia and then by rail to the U.S. destination originally intended) and are included in the “on dock” price per part charged by suppliers. Repositioning costs, which are paid directly by Dell to the carriers, comprise all other ground transportation costs incurred as a result of the supply-routing decisions defined above (i.e. diversions, expediting and transfers). Finally, some of the numerical data included in this paper are disguised in order to protect the confidentiality of Dell’s sensitive information.
2. Literature Review

The reader may have noted from §1 that the high-level structure of Dell’s supply chain for large desktop components in North America closely resembles the one captured by the inventory distribution model of Eppen and Schrage (1981). Among common features, Dell’s supply-routing problem also involves the centralized allocation of incoming inventory among several facilities where it is stored and consumed, and its cross-docking disembarkment operation in Long Beach, California exactly matches the definition of a “stockless depot” considered in that paper. Consequently, many of the results and insights described in the body of literature on multiechelon inventory allocation that started with that seminal paper (see Axson et al. 2002 for a recent survey) are conceptually relevant to the problem considered.

In spite of all these papers’ relevance, however, both our goal and methodology differ substantially from theirs. Specifically, our objective is to develop and implement an operational system for a large existing supply chain, as opposed to deriving theoretical insights from a stylized model. Consequently, our approach sacrifices tractability for realism and operational applicability, and the model we formulate is a mixed-integer program solved over a rolling horizon using numerical (branch-and-bound) algorithms, as opposed to, say, a dynamic program—see Chand et al. (2002) for a more general review of rolling horizon methods.

Some insights on the supply chain motivating our work may be gained from Kapuscinski et al. (2004), which describes the development and implementation by Dell of replenishment models for its component inventory. That paper is thus an important complement to ours, in that it focuses on how the inventory-ordering decisions, which we assume to be exogenous here, should be generated by Dell’s suppliers as part of their VMI relationship (see §1).

The paper most related to ours, however, is Caggiano et al. (2006), which considers operational models for inventory and capacity allocation decisions in a multi-item reparable service part system with a central warehouse and field stocking locations. In particular, their Extended Stock Allocation Model (ESAM), which leaves the repair decisions aside, is similar in many respects to the one we describe in §4: It is a mathematical program meant to be solved on a rolling horizon basis, its decision variables comprise inventory allocation and expedited shipment decisions, its objective function includes a transportation component and a newsboylike backorder component, and it assumes deterministic lead times and an exogenous supply pipeline. However, the ESAM is still simpler than our model, in that it does not capture transshipments, considers a single expedited transportation mode, assumes a linear transportation cost structure, and ignores transportation-scheduling and capacity constraints. These differences are material, because the solution approach ultimately followed in Caggiano et al. (2006) consists of developing heuristic solutions by exploiting the structural properties that can be established in their setting, whereas we compute instead solutions to an approximate (linearized) version of our model (see §4.1). Most importantly, however, our paper describes an actual implementation of the optimization model, presented along with an assessment of its impact, and thus offers a grounded perspective on the many important practical issues involved. This practical focus is reflected in the structure of this paper, whereby we now describe in turn the two successive phases followed by Dell as part of that implementation.

3. First Phase: Process Design

3.1. Development

The first phase of this project, which is described more extensively in Reyner (2006), started in the spring of 2005. Its goal was to correct an observed increase in expediting and transfer costs for large components by coordinating the associated decision process across the relevant groups within Dell. The associated solution designed and implemented later that year focused on monitors because they are very large contributors to Dell’s supply transportation costs. It comprised two main components: The first was organizational, and involved the creation (and staffing) of a specific job definition named supply-routing analyst, with the responsibility of gathering and analyzing all

\[\text{allocation function (splitting incoming quantities among final destinations)}\]

\[\text{ordering function (determining incoming quantities)}\]
Figure 2 Visualization of Dynamic Routing Decisions with the Balance Tool

<table>
<thead>
<tr>
<th></th>
<th>Demand forecast</th>
<th>Starting inventory</th>
<th>Planned deliveries</th>
<th>Delivery adjustment</th>
<th>Ending inventory</th>
<th>DSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austin</td>
<td>600</td>
<td>1,400</td>
<td>6,000</td>
<td></td>
<td>800</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>800</td>
<td>200</td>
<td></td>
<td>5,600</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>200</td>
<td></td>
<td></td>
<td>5,000</td>
<td>9.3</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td></td>
<td></td>
<td></td>
<td>5,000</td>
<td>8.3</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td></td>
<td></td>
<td></td>
<td>2,100</td>
<td>5.3</td>
</tr>
</tbody>
</table>


relevant information in order to make and implement all supply-routing decisions. The primary objective specified for this new position consisted of reducing the repositioning transportation costs incurred by Dell, subject to acceptable levels of inventory availability at the various sites (Austin, TX; Nashville, TN; Reno, NV; Winston-Salem, NC).

The second component was the development of a supporting spreadsheet-based information acquisition and visualization tool, that became known as the Balance Tool. As seen from the information subset displayed in Figure 2, this tool simultaneously displays all available demand forecasts and scheduled supply deliveries for each monitor type in each of the relevant factories and warehouses over a rolling horizon of several weeks, with a planning period of one day. The corresponding information sources include Dell’s carriers, who update the scheduled supply deliveries on a daily basis, and Dell’s own forecasting group, which provides a weekly update of all demand forecasts (for each monitor type in each location) for every week over a forecasting horizon of several months; these weekly forecasts are then divided equally between all working days of the corresponding week in order to obtain daily forecasts. Combining that information with the current inventory on hand and backlog in the various sites allows Dell to compute projected net inventory equivalent DSI (days of supply in inventory) levels in all the relevant locations over this horizon, and highlight any anticipated shortages. Specifically, the balance tool uses a color code to show different categories of DSI levels on each day of the horizon in each facility; the color codes are red (critical situation), yellow (should be monitored), and green (sufficient inventory). Based on daily updates of the information displayed for each component and using special entry cells, the supply-routing analysts can then manually explore the implications of all possible routing decisions. For example, a container scheduled to arrive in Austin in the later part of the horizon could be diverted to Nashville and expedited by team truck, which the balance tool would reflect by removing that container from Austin’s supply line on its original scheduled arrival date and adding it to that of Nashville on a closer date (determined by the difference between the
transportation times from Long Beach to Austin by rail and to Nashville by team truck, respectively). The resulting new inventory and DSI levels in both sites resulting from such a move would then be instantly displayed, showing, for example, the extent to which this action would help correct a projected shortage situation in Nashville in the short term when Austin is projected to have excess inventory later in the horizon. Finally, the length of the planning horizon was chosen so that it would always include any containers located before the diversion cutoff point of two–three days before port, assuming the longest-possible ground transportation lead time and then adding an additional time buffer.

3.2. Implementation
The creation of the supply-routing analyst position was welcomed by the various groups previously involved in making these decisions, in part because many of those involved regarded supply routing as a nonexplicit yet time consuming part of their work assignment. The balance tool was implemented with the spreadsheet program Microsoft Excel, augmented with Visual Basic macros that automated certain functions such as retrieval of external data as well as creation and deletion of parts.

An important implementation decision in this phase was to organize a live pilot of the newly designed decision process for a selected subset of components (monitors) relatively early on (September 2005). This pilot uncovered many improvement opportunities for the balance tool, forced a grounded reflection on how supply-routing decisions should be made in specific situations, and helped quantify the impact of the new process, as we discuss next.

3.3. Impact
The financial impact of this first phase was estimated using a fairly coarse methodology. Specifically, managers reviewed all the decisions made over a limited period of time during the live pilot described above, along with the associated input data. In each case, they determined which alternative decisions would likely have been made under the previous process, along with their associated repositioning transportation costs. Because the part shortages were generally perceived to have decreased during the pilot, the repositioning cost savings calculated in this way were considered meaningful. Although this methodology involves many subjective and arguably biased inputs, its results were still deemed valid by Dell’s managers and led to their conclusion that the new process reduced repositioning transportation costs by about 40% (Reyner 2006). We suspect that this quantitative impact estimation was easily accepted because it had a clear qualitative explanation. Specifically, the new process generated comparatively more rail diversions (which involve only a small bill of lading splitting fee) and fewer transfers and less expediting (which cost considerably more). Indeed, rail diversions had been a neglected lever because they require more information than transfers and their organizational ownership had previously been unclear. Shortly after this live pilot was completed in January 2006, Dell started using the new process described above continuously for all its monitors (about two dozen different part types).

4. Second Phase: Optimization
In spite of its positive impact, the first project phase also revealed the following improvement opportunities:
—Relying exclusively on the analysts’ judgement proved problematic from a time efficiency standpoint because of the high number of parts to manage, the very high number of potential decisions involved for each part, the large amount of relevant information and the high decision-making frequency: Although forecasts can change daily, for example, whenever a large customer order is received, the analysts were only able to review and change the status of any particular part once a week on average;
—From a resiliency standpoint, it also seemed problematic for Dell to depend entirely on a handful of individuals for such frequent and critical control decisions;
—Finally, the balance tool only characterized expected shortages very coarsely through the net inventory levels displayed and the color code described above, and did not provide an estimate of the repositioning transportation costs associated with the decisions considered. It was thus suspected that even an experienced analyst could easily make sub-optimal decisions in this setting.
These observations motivated the second phase of this project, which started in September 2006. Its objective was to develop and implement an optimization-based decision support system to assist the supply-routing analysts. It was decided up-front that the structure of this optimization model would support the decision process established in the previous phase. Specifically, the model envisioned would need to generate recommendations on a rolling horizon basis and for each monitor considered independently, consistent with the supply-routing analysts’ practice when using the balance tool. We note here that managing each monitor type independently of all the other components, which greatly simplifies the problem, is only made possible by the shipments of monitors in full containers of a single part type. Because this does constitute a limitation of our approach, we return to this issue in §5. Another important design consideration stemmed from the envisioned execution of the model on a rolling horizon basis. Specifically, because the model was to consider a time horizon of several weeks and generate in each run a set of recommended supply-routing decisions for each part over that period, some of these decisions could possibly only be enacted on some distant day in the future. In this context, we defined the concept of time sensitivity for each individual decision as the number of days before the opportunity to enact that decision would disappear. For example, a recommended diversion decision affecting a container on a vessel five days away from Long Beach would have a time sensitivity of three days if the diversion cut-off point for this part was two days before port. This would enable the analysts to only enact the decisions with a time criticality lower than a set threshold (for example, the number of days before the next anticipated run on that part), with the overall goal of waiting for as long as possible for the most recent data before committing to any decision.

4.1. Development
An important requirement for this optimization model was to capture the main trade-off involved in Dell’s supply-routing decisions, namely, the tension between repositioning transportation costs on one hand and shortage costs on the other hand (note that inventory holding costs were ignored, for reasons that will be explained in §4.1.2). Although expressing the repositioning costs incurred as a function of the routing decisions considered is relatively straightforward, as will be seen shortly, the critical modeling challenge was to quantify the benefits associated with these decisions, that is, the overall change of expected shortage costs in all of the sites where the projected inventory levels were affected. Our approach involved the formulation of an approximate expected shortage function depending on these inventory levels and the exogenous variability of demand forecasts for each location. Mathematical details for this function and its approximation are presented in §4.1.1, and the resulting MIP embedding this approximate expected shortage cost function is described in §4.1.2.

4.1.1. Expected Shortage Costs. We adopted a standard linear structure \( B \sum_{t \in \mathcal{T}, t \leq t_0} v_{lt} \) for the total expected shortage costs predicted in all facilities \( l \in \mathcal{L} \) {Austin, Nashville, Reno, Winston-Salem} for a specific part over the rolling horizon \( t \in \mathcal{T} \subseteq [1, \ldots, T] \) considered, where \( B \) is a unit daily shortage cost rate and \( v_{lt} \) is the expected shortage level for future day \( t \) in location \( l \). In practice, shortage costs stem from a variety of factors including: order cancellations by impatient customers; expedited shipping to customers with late orders; substitutions of more expensive components for the same price; lost profit from customers turned away by long posted lead times; price concessions on future orders… . We refer the reader to Kapuscinski et al. (2004) and Dhalla (2008) for more comprehensive and detailed descriptions. We discuss the constant \( B \) later in §4.2 and develop next an expression for \( v_{lt} \) as a function of the supply-routing decisions considered and the available inventory and demand data.

Our first step consisted of characterizing the distribution of actual demand relative to the forecast available for that quantity at the time when routing decisions need to be made, because this information was not available to us at the outset. This empirical study of the cumulative forecast error (see §A.1 in the online appendix) both suggested the structure and provided the standard deviation input data \( \sigma_{lt} \) for the stochastic model

\[
\sum_{k=1}^{l} d_{kl} \sim N \left( \sum_{k=1}^{l} f_{kl}, \sigma_{lt} \right),
\]
where $d_t$ is the random variable representing demand on day $t$ for a given part in a given location $l$, $\{\text{Austin, Nashville, Reno, Winston-Salem}\}$, as estimated at the beginning of the current day (always indexed by 1 in our rolling horizon model), so that $\bar{d}_t \triangleq \sum_{k=1}^{t} d_{tk}$ is the cumulative demand for the next $t$ days; $N(f, \sigma)$ refers to a normal distribution with mean $f$ and standard deviation $\sigma$; $f_{tl}$ is the (deterministic) forecast of the same quantity generated by Dell and provided to the supply chain analyst on day 1, so that $\bar{f}_{tl} \triangleq \sum_{k=1}^{t} f_{tk}$ is the corresponding cumulative forecast of demand up to day $t$; and finally $\sigma_{tl}$ is the standard deviation of the forecasting error $\bar{d}_t - \bar{f}_t$. Note that the forecasting error study mentioned above did identify some systematic biases. These biases were ignored, however, because they were relatively small and convincingly explained by the forecasting team. The notations $f_{tl}$ and $d_{tl} \triangleq \mathbb{E}[d_{tl}]$ will thus be used interchangeably from now on.

The inventory dynamics over the rolling horizon considered are described in our model by the following balance equation, which assumes that any unmet demand is backlogged:

$$I_{t+1,l} = I_{tl} + \sum_{k=1}^{t} s_{tk} - \bar{d}_{tk} \quad \text{for } t \geq 1, \quad (2)$$

where $I_{tl}$ is the (random) net inventory level available at the beginning of day $t$ in location $l$, as predicted at the beginning of day 1 (so that $I_{t1,l} = I_{tl}$ is deterministic input data), and $s_{tk}$ is the net result of deliveries into and transfers out of location $l$ on day $t$ (which is directly affected by the supply-routing decisions we seek to determine). Note that $s_{tk}$ is assumed to be deterministic in our model, which ignores supply uncertainty. This is justified by the fact that in Dell’s setting, supply uncertainty is small relative to demand uncertainty given the (daily) time granularity considered. As a result, the ranges of lead times appearing in Figure 1 are essentially driven by the differences across destinations, as opposed to any potential unpredictable variability affecting the lead time associated with a given transportation mode on a specific leg. Also, that assumption does not affect the operational applicability of the model output, as will be seen later.

Next, we approximate the shortage level $v_{tl}$ for day $t$ in location $l$ predicted at the beginning of day 1 as

$$v_{tl} \triangleq (I_{tl} - d_{tl})^{-}. \quad (3)$$

Note that the expression $I_{tl} - d_{tl}$ for the net inventory level during day $t$ that appears in (3) corresponds to the most pessimistic assumption for the daily schedule of supply and demand. That is, demand is assumed to occur entirely at the very beginning, and supply deliveries at the very end, of day $t$. This approach was followed because the expressions derived from other assumptions (say, continuous supply and demand processes) are less tractable, detailed hourly demand and supply data were not easily accessible, and because of Dell’s expressed desire to err on the conservative side when predicting shortages. Substituting (2) and (1) in (3) yields

$$v_{tl} \sim [N(I_{tl} - f_{tl}, \sigma_{tl})]^{-} \quad \text{for } t \geq 1, \quad (4)$$

which characterizes the distribution of these shortages in terms of decision variables and input data. The expectation of the random variable in (4) is thus given by the standard normal loss function

$$v_{tl} = \sigma_{tl}\Phi\left(\frac{f_{tl} - I_{tl}}{\sigma_{tl}}\right) + \left(f_{tl} - I_{tl}\right) \Phi\left(\frac{f_{tl} - I_{tl}}{\sigma_{tl}}\right), \quad (5)$$

where $\Phi$ and $\Phi$ are the standard normal probability density function and cumulative distribution function, respectively.

Our final goal is to embed the function of $f_{tl} - I_{tl}$ defined by (5) in the objective of a linear integer programming minimization model. Because this function is convex, we can use the following standard approximation method: For each location $l$ and time period $t$, we add linear constraints requiring that variable $v_{tl}$ exceed a number of supporting tangents to the

---

4 One referee speculates that Dell will at some point discover that it has more lead-time variability than it at first thought, and will consider switching back to having closer suppliers, in response to the costs induced by that uncertainty.

5 The alternative approximation $v_{tl} \triangleq (I_{tl} + s_{tl})^{-}$ is equally tractable and is the most optimistic in the sense just defined. It thus constitutes a bound that allowed us to verify that the assumption reflected by (3) was relatively immaterial.

6 Random variables and their distributions are used interchangeably in (4), because no related ambiguity arises here.
function defined by the right-hand side of (5), which amounts to approximating that function by the upper envelope of a finite number of its tangents. The resulting optimization model thus requires a precomputation of the slopes \( a_{ip} \) and intercepts \( b_{ip} \) of a set of tangents (indexed by \( p \)) to the expected shortage function (5) defined for each location \( l \) and time period \( t \). To this end, we first determine a relevant approximation range \([I_{il}^{LB}, I_{il}^{UB}]\) for \( I_{il} \) that depends only on input data. A tight upper bound \( I_{il}^{UB} \) follows from the observation that the maximum expected net inventory level in location \( l \) at the beginning of time \( t \) is obtained by instantly transferring to \( l \) all inventory from other facilities, and rerouting toward \( l \) with the fastest ground transportation mode (team truck) all “divertable” containers that can arrive at \( l \) by time \( t \). Likewise, a tight lower bound \( I_{il}^{LB} \) corresponds to the situation where all inventory available in \( l \) is transferred immediately to other facilities, and all containers initially bound to \( l \) are diverted away while demand continues to deplete this facility. Finally, we calculate iteratively a discrete set of sampling points \( P_{il} \subset [I_{il}^{LB}, I_{il}^{UB}] \) indexed by \( p \), and the corresponding slopes \( a_{ilp} \) and intercepts \( b_{ilp} \) of the tangents to the r.h.s. of (5) in those points, using numerical implementations of \( \phi \) and \( \Phi \), along with the maximum error rule method described in Rote (1992). In practice, we found that with this method calculating only four tangents for each time and location yields a very high accuracy.

4.1.2. Optimization Model Formulation. The optimization model we developed to generate supply-routing recommendations over a rolling horizon for each monitor type considered independently is the following MIP:

**Input Data:**

- **Time and Location.** The rolling horizon considered is \( \mathcal{T} \triangleq \{1, \ldots, T\} \), and the set of relevant locations is \( \mathcal{L} \triangleq \{\text{Austin, Nashville, Reno, Winston-Salem}\} \).
- **Part Characteristics.** For the part considered, the maximum number of parts per truck is denoted \( Q \), and \( W \) refers to the number of parts per pallet.
- **Supply Pipeline.** Incoming supply consists of a set \( \mathcal{C} \) of containers indexed by \( i \), each containing a quantity of parts \( q_i \) with a current destination \( l_i \in \mathcal{L} \) and an expected arrival date \( A_i \in \mathcal{T} \). Containers that are still divertable (typically all containers still on the ocean and at least two or three days away from port) form a subset \( \mathcal{C}^{RT} \subset \mathcal{C} \), whereas the containers in the complement set \( \mathcal{C}^{NRT} \triangleq \mathcal{C} \setminus \mathcal{C}^{RT} \) may no longer be rerouted before they arrive at their destination. The expected arrival date at the port (Long Beach, CA) of container \( i \) in \( \mathcal{C}^{RT} \) is denoted \( A_i^{LB} \in \mathcal{T} \). Often containers travel as a group of multiple containers all sharing the same bill of lading, and therefore the same destination, transportation mode, and expected arrival time. Containers with the same bill of lading may be split, however, provided they belong to \( \mathcal{C}^{RT} \). In this case, the carrier creates as many new bills of lading as the resulting number of container groups traveling together, incurring an administrative fee of \( c_{BL} \) times the number of new bills of lading created. Bills of lading are indexed by \( j \in \mathcal{J} \), and the subset of containers sharing each bill of lading \( j \) is denoted \( \mathcal{C}_j \), so that \( \mathcal{C} = \bigcup_{j \in \mathcal{J}} \mathcal{C}_j \).
- **Current Net Inventory.** The sum of on-hand inventory currently available (that is, at the beginning of day 1 of the rolling horizon) minus backorders in each location \( l \in \mathcal{L} \) is denoted \( I_{il} \).
- **Demand Forecast.** The forecast of demand in location \( l \in \mathcal{L} \) during day \( t \) is denoted \( f_{tl} \), whereas the cumulative forecast of demand from day 1 to day \( t \) (included) is denoted \( f_{t}^{NM} \).
- **Container Ground Transportation Modes.** Ground transportation modes between the port and Dell’s facilities are indexed by \( m \in \mathcal{M}^{RT} = \{\text{rail, single truck, team truck}\} \) and characterized for each destination \( l \in \mathcal{L} \) by a cost per container \( c_{im}^{RT} \) and an average lead

\(^7\) The equations for \( I_{il}^{LB} \) and \( I_{il}^{UB} \) in terms of input data are straightforward and omitted here.

\(^8\) This algorithm initiates with \( P_{il} = [I_{il}^{LB}, I_{il}^{UB}] \). In each iteration, tangents are constructed for each new point in \( P_{il} \), and the x-axis values of the intersection of tangents corresponding to adjacent points in \( P_{il} \) are added as new points. The algorithm terminates when the maximum difference between the y-axis values of these intersections and the corresponding function values reaches a specified upper bound.

\(^9\) The superscripts RT and NRT stand for *routable* and *nonroutable*, respectively.
time \( L_{it}^{RT} \) (expressed in days). Note that the repositioning transportation costs incurred when diverting a container \( i \) to a destination \( l \) with transportation mode \( m \) are \( c_{it}^{RT} - c_{i,t\text{rail}}^{RT} \) in addition to any bill of lading creation fee involved. That is, the embedded costs \( c_{i,t\text{rail}}^{RT} \) corresponding to a shipment by rail to the original destination \( l \) must be subtracted because they are reimbursed by the original rail carrier to Dell when a container is diverted (see §1 for definitions of transportation costs, and §4.2.2 for a discussion of related implementation issues). Finally, the potential expected delivery date at location \( l \) of any divertable container \( i \in C_{i}^{RT} \) is \( A_{i}^{LB} + L_{it}^{RT} \).

**Special Transfers.** Special transfers of inventory between two facilities \( l \) and \( l' \) in \( D \) are characterized by their expediting mode \( m \in \mathcal{M}^{SP} \) \( \neq \{ \text{single truck, team truck} \} \), their cost per truck \( c_{l'\text{t}}^{SP} \), and their lead time \( L_{ll'}^{SP} \).

**Milk-Run Transfers.** Milk-run transfers of inventory from facility \( l \) to facility \( l' \) are characterized by their schedule of departures \( S_{l}^{MR} \) (equal to 1 if a run from \( l \) to \( l' \) is scheduled on day \( t \) and 0 otherwise), their cost per pallet \( c_{l'}^{MR} \), the maximum number of pallets of a given part allowed in each run \( R \), and their lead time \( L_{ll'}^{SP} \). Note that the milk-run capacity limit \( R \) is part specific. Milk-run carriers have a contractual obligation to provide transportation capacity up to a specified number of trucks on every given run; however, that capacity is common to many different parts. In order to avoid potential capacity allocation conflicts between parts and incentives for each part manager to reserve some of that capacity before others, Dell has introduced these part-specific capacity limits, which are substantially smaller than the total truck capacity available.\(^{10}\)

**Decision Variables:**

**Container Routing.** Binary variables \( y_{ilm} \) are set to 1 if container \( i \in C_{i}^{RT} \) is routed from the port to facility \( l \) using transportation mode \( m \in \mathcal{M}^{RT} \), and 0 otherwise. In addition, binary variables \( z_{ilm} \) take the value 1 if at least one container \( i \in C_{i}^{RT} \) from bill of lading \( j \) is routed to facility \( l \) using transportation mode \( m \in \mathcal{M}^{RT} \), and 0 otherwise.

**Special Transfers.** Integer variables \( X_{il'l'm} \) represent the number of full trucks sent from facility \( l \) to facility \( l' \) on day \( t \) using expediting mode \( m \in \mathcal{M}^{SP} \), binary variables \( x_{il'l'm} \) are set to 1 if a less-than-full truck is used between \( l \) and \( l' \) on day \( t \) with mode \( m \) and 0 otherwise, and continuous variables \( w_{il'l'm} \leq Q \) represent the number of parts carried in that truck.\(^{11}\)

**Milk-Run Transfers.** Integer variables \( r_{il'l'} \) represent the number of pallets included in the run from facility \( l \) to facility \( l' \) on day \( t \).

**Inventory Variables.** Continuous variable \( I_{it} \) denotes the expected net inventory level at the beginning of day \( t > 1 \) in location \( l \), associated variables include its positive part \( I_{it}^{+} \) and negative part \( I_{it}^{-} \), and a binary indicator variable \( I_{it}^{0} = 1 \{ I_{it} \geq 0 \} \).

**Expected Shortages.** Continuous variables \( v_{it} \) approximate the predicted expected shortages during each day \( t \) in each location \( l \) (see §4.1.1).

**Formulation:**

\[
\min \sum_{i \in C_{i}^{RT}, l, m \in \mathcal{M}^{RT}} (c_{i}^{RT} - c_{i,t\text{rail}}^{RT})y_{ilm} + \sum_{j \in J} c_{l\text{t}}^{BL} \left( \sum_{l, m \in \mathcal{M}^{RT}} z_{ilm} - 1 \right) + \sum_{l, l', m \in \mathcal{M}^{SP}, t \in T} c_{l'l'm}^{SP} (X_{il'l'm} + x_{il'l'm}) + \sum_{l, l', t \in T} c_{l'}^{MR} r_{il'l'} + B \sum_{t \in T} v_{it} \tag{6}
\]

subject to:

\[
I_{it} = I_{it}^{0} - \bar{f}(t-1) + \sum_{i \in C_{i}^{RT}, l, m \in \mathcal{M}^{RT}} q_{i} y_{ilm} + \sum_{i \in C_{i}^{RT}, l, m \in \mathcal{M}^{RT}} q_{i} y_{ilm}^{1-1} \leq -1 \]

\(^{10}\) These static part-specific capacity limits may not be optimal, and researching improved mechanisms with dynamic, situation-specific allocation limits, for example, seems worthwhile. However, the current practice is simple and relies on tools and processes that are readily available.

\(^{11}\) The integrality of \( w_{il'l'm} \) is immaterial in light of the quantities at stake here.
Foreman et al.: Implementing Supply-Routing Optimization in a Make-to-Order Manufacturing Network

The objective (6) is the sum of all repositioning transportation costs associated with the decisions considered, including container diversions (first term), bill of lading fees (second term), special trucks (third term), and milk runs (fourth term), along with the corresponding expected shortage costs (last term). Note that our choice of minimizing total costs, as opposed to, say, minimizing repositioning costs subject to a service-level constraint on total expected shortages, is dictated by context. Specifically, Dell’s suppliers are responsible for all initial shipment decisions (see §1), which are thus exogenous to the routing problem considered. As a result, such a service-level constraint could lead to infeasibility problems. Also, (6) does not account for any inventory costs that could arise from excessive inventory in a given location. Although it would be straightforward to add a term summing the on-hand inventory variables $I^+_t$ multiplied by an inventory holding cost rate, it turns out that the relevant costs associated with excessive inventory mostly stem here from the additional storage required in the warehouses adjacent to its factories when the overall amount of inventory across all parts exceeds a threshold. Although the inventory holding costs incurred by Dell’s suppliers in those warehouses may in turn affect Dell in important ways, these primarily depend on the overall quantity of inventory shipped (as opposed to the allocation of this inventory across sites), which is exogenous. In light of these considerations and because the inventory storage costs incurred historically represent only a very small fraction of the repositioning costs, it was decided to leave them out of the optimization model.

Constraints (7) are inventory balance equations defining the relationship between the expected net inventory variables $I^*_t$ and the inventory currently available ($I^0_t$), the demand forecasts ($\bar{I}^0_t - I^0_t$), the pipeline of nonroutable containers ($\sum_{l \in \mathcal{L}} q^l_t$), and all the supply-routing decisions considered (all subsequent terms in the right-hand side). Constraints (8) ensure that every container is routed to exactly one destination through one transportation mode. Constraints (9) ensure that the term $\sum_{l \in \mathcal{L}} z^l_{lm} - 1$ appearing in the objective corresponds to the number of new bills of lading created for the containers initially included in bill of lading $j$ as a result of the routing decisions. Constraints (10)–(12) ensure that variables $I^*_t$, $I^0_t$, and $I^+_t$ correspond to the positive part, negative part, and nonnegativity indicator of variable $I_t$, respectively. Constraints (13) state that the total inventory transferred out of any facility $l$ during a given day $t$, either through special trucks or a milk run, may not exceed the inventory on hand expected to be available in that facility at the beginning of that day. Constraints (14) ensure both that the quantity of parts recommended for transportation aboard a less-than-full special transfer truck does not exceed its capacity, and that the binary variables signalling the existence of such trucks take values consistent with their definition. Similarly, (15) enforces both the capacity and the scheduling restrictions of milk runs between facilities.
Note that the variables $S^\mathrm{MR}_{l,t}$ are only introduced here to simplify exposition, because for implementation purposes it is more computationally efficient to only define variables $t_{l,t}$ over the set of indices $(t, l, l')$, such that there exists a run from $l$ to $l'$ on day $t$. Finally, constraints (16), together with the minimization objective, ensure that in any optimal solution (and any solution computed through a branch-and-bound MIP algorithm) the variables $t_{l,t}$ implement, indeed, the approximate expected shortage level during day $t$ in location $l$, which is described in §4.1.1.

4.2. Implementation

We discuss, in turn, software development and computational performance (§4.2.1), input data collection and shortage cost estimation (§4.2.2), and, finally, pilot testing (§4.2.3).

4.2.1. Software Development and Computational Performance. The software implementation of the model described in §4.1 was performed using the development environment OPL Studio linked with the optimization engine CPLEX 9.1, using Microsoft Excel as a repository for the static input data (costs, lead times, forecast accuracy parameters, shortage costs) and also to visualize the output data. As illustrated in Figure 3, these output data include not only the individual recommended decisions for all monitors, but also their time sensitivity (see definition at the beginning of §4), and the associated visualization interface can sort all decisions generated accordingly. In addition, links were created with some of Dell’s existing databases in order to automatically import the dynamic input data (current inventory levels, forecasts, pipeline inventory) whenever required. Finally, the preprocessing necessary to compute the piecewise-linear approximations to the expected shortage cost functions (see §4.1.1) was implemented using Microsoft Visual Basic. The creation of this software tool from complete specifications required approximately six months of full-time work by an experienced developer familiar with optimization theory at an introductory graduate course level. We refer the reader to §A.2 in the online appendix for a more detailed description of this software (including additional screen copies of its interface), and to Foreman (2008) for the source code.

Our next step was to evaluate the computational time associated with executing the branch-and-bound algorithm on realistic problem instances. To this end, we gathered a large and representative collection of input data sets on which we performed many

---

**Figure 3 Output Interface of the Model Software Implementation**

<table>
<thead>
<tr>
<th>Part #</th>
<th>Decision</th>
<th>Cord #</th>
<th>BOL #</th>
<th>Org</th>
<th>Dest</th>
<th>Mode</th>
<th>Qty</th>
<th>Qty parts</th>
<th>Time sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>GY946</td>
<td>DIV</td>
<td>MERW74948502</td>
<td>YO1966972</td>
<td>Nashville</td>
<td>Austin</td>
<td>Rail</td>
<td>Container</td>
<td>1580</td>
<td>3</td>
</tr>
<tr>
<td>GY946</td>
<td>DIV</td>
<td>JDRB84862765</td>
<td>YO1966972</td>
<td>Nashville</td>
<td>Austin</td>
<td>Rail</td>
<td>Container</td>
<td>1580</td>
<td>3</td>
</tr>
<tr>
<td>D9923</td>
<td>DIV</td>
<td>WDL66862665</td>
<td>JDR1741858</td>
<td>Austin</td>
<td>Austin</td>
<td>Team truck</td>
<td>Container</td>
<td>2184</td>
<td>4</td>
</tr>
<tr>
<td>D9923</td>
<td>DIV</td>
<td>JDRB84862765</td>
<td>JDR1741858</td>
<td>Austin</td>
<td>Austin</td>
<td>Team truck</td>
<td>Container</td>
<td>2184</td>
<td>4</td>
</tr>
<tr>
<td>D9923</td>
<td>DIV</td>
<td>CEGS97918425</td>
<td>JDR1741858</td>
<td>Austin</td>
<td>Austin</td>
<td>Team truck</td>
<td>Container</td>
<td>2184</td>
<td>4</td>
</tr>
<tr>
<td>D9923</td>
<td>RB</td>
<td>N/A</td>
<td>N/A</td>
<td>Nashville</td>
<td>Austin</td>
<td>Red ball</td>
<td>5 pallets</td>
<td>420</td>
<td>3</td>
</tr>
<tr>
<td>D9923</td>
<td>RB</td>
<td>N/A</td>
<td>N/A</td>
<td>Nashville</td>
<td>Austin</td>
<td>Red ball</td>
<td>8 pallets</td>
<td>672</td>
<td>7</td>
</tr>
<tr>
<td>D9923</td>
<td>RB</td>
<td>N/A</td>
<td>N/A</td>
<td>Nashville</td>
<td>Austin</td>
<td>Red ball</td>
<td>8 pallets</td>
<td>672</td>
<td>10</td>
</tr>
<tr>
<td>D9923</td>
<td>TRK TRNS</td>
<td>N/A</td>
<td>N/A</td>
<td>Winston salem</td>
<td>Austin</td>
<td>Truck</td>
<td>1 ft</td>
<td>2173</td>
<td>6</td>
</tr>
<tr>
<td>D9923</td>
<td>TRK TRNS</td>
<td>N/A</td>
<td>N/A</td>
<td>Winston salem</td>
<td>Austin</td>
<td>Team truck</td>
<td>3 ft</td>
<td>8620</td>
<td>6</td>
</tr>
<tr>
<td>D9923</td>
<td>TRK TRNS</td>
<td>N/A</td>
<td>N/A</td>
<td>Winston salem</td>
<td>Austin</td>
<td>Team truck</td>
<td>3 ft</td>
<td>2184</td>
<td>11</td>
</tr>
<tr>
<td>FG645</td>
<td>DIV</td>
<td>MERW1418486</td>
<td>GUY269444</td>
<td>Austin</td>
<td>Rano</td>
<td>Rail</td>
<td>Container</td>
<td>616</td>
<td>1</td>
</tr>
<tr>
<td>FG645</td>
<td>DIV</td>
<td>FWGR5894867</td>
<td>GUY269444</td>
<td>Austin</td>
<td>Rano</td>
<td>Rail</td>
<td>Container</td>
<td>616</td>
<td>1</td>
</tr>
<tr>
<td>FG645</td>
<td>DIV</td>
<td>JDRB8174585</td>
<td>GUY269444</td>
<td>Austin</td>
<td>Rano</td>
<td>Rail</td>
<td>Container</td>
<td>1392</td>
<td>1</td>
</tr>
</tbody>
</table>

Deactivate selected | Cancel
optimization runs. This demonstrated that although achieving optimality occasionally required more than an hour for these problems, a suboptimality gap equivalent to a bill-of-lading creation fee (the smallest individual transportation cost component appearing in the objective function (6)) was almost always achieved in a matter of seconds using standard search strategies. As a result, the achievement of a suboptimality gap equal to that amount was set as our algorithm termination criterion, and we did not further investigate the computational solution time for this problem.14

4.2.2. Input Data Collection and Shortage Cost Estimation. Although the input data requiring frequent updates (current inventory levels, forecasts, pipeline inventory) could be readily obtained from Dell’s existing databases, it proved difficult to obtain accurate rail transportation costs. It was known, however, that rail transportation costs are very small relative to all other repositioning costs involved, which justified the approximations $c_{lm}^{RT} - c_{rail}^{RT} \approx c_{lm}^{RT}$ for $m \in M^{RT} \setminus \text{[rail]}$, and $c_{rail}^{RT} - c_{l,rail}^{RT} \approx 0$ in the first term of the objective. After checking through sensitivity analysis that this would have little impact, if any, on the decisions recommended, we thus decided to ignore rail transportation costs, i.e., we set $c_{lm}^{RT} = 0$ for $m = \text{“rail.”}$

The most important implementation hurdle faced at this point, however, was to determine what value(s) should be used in practice for the unit shortage cost rate $B$ introduced in §4.1. As the model’s main input parameter for resolving the trade-off between shortage and repositioning transportation costs, it had a significant impact on the output: With a low value of $B$, most decisions generated could be container diversions with no expediting and some milk-run transfers, whereas in the same situation a high value of $B$ could generate much expediting and many transfers through team trucks. However, no study previously performed within Dell was available to guide the implementation team toward an objective value for that parameter. The strategy decided on, then, was twofold: For the long term, an in-depth study of Dell’s shortage costs was initiated, following a methodology similar to that described by Oral et al. (1972) (see Dhall 2008 for more details); in the short term, $B$ was to be treated as a control lever that the supply-routing analyst could initially adjust, with the goal of achieving through experimentation the same trade-off between transportation costs and service level as the one that was implicitly associated with the decisions made to date. Although this short-term strategy would not necessarily be optimal, it would still hopefully generate consistent supply-routing decisions in an efficient manner. In addition, these decisions could still possibly produce substantial savings in repositioning costs.

Unfortunately, this empirical determination of $B$ proved more difficult than anticipated. This is because the analyst would primarily evaluate the criticality of a given supply situation by inspecting on the Balance Tool the DSI levels projected in all of Dell’s facilities over the planning horizon, and then relied on a subjective and empirical notion of the relationship between these DSI levels and the corresponding expected shortages—we refer the reader to §§4.3.2 and 4.3.3 for a more detailed discussion of the analyst’s heuristics and their implications.

In the end, the implementation team resolved the question of what initial values for $B$ should be chosen through a study of historical data. We implemented the idea of constructing management decision rules based on a regression of past managerial actions, which goes back at least as far as Bowman (1963). More specifically, we constructed a database where each entry corresponds to a set of routing decisions made by the analyst for a given part on a given day, and includes both the associated repositioning transportation costs as well as the corresponding reduction in total expected shortages over the planning horizon, as estimated by the model using all relevant input data available at the time. As Figure 4 illustrates, we then performed a linear regression with forced zero intercept of the reduction in expected shortages achieved (dependent variable) as a function of the repositioning costs incurred (independent variable).

14 Although we do not offer a full explanation for this good computational performance, we believe that it is due to the relatively high structural similarity between our model and a time-space network formulation where nodes represent every day and location, flows represent inventory, and arcs joining these nodes represent either the passing of time or the routing decisions considered. Note that flow conservation constraints in such a network formulation correspond exactly to our inventory balance Equation (7).
for each part over that data set, which spanned several weeks of decisions. An interesting aspect of these regressions is that their fit provided a measure for the consistency of the analysts’ historical decisions with respect to the trade-off between repositioning costs and expected shortages, as determined through our stochastic evaluation model. From this standpoint, it was found that these regressions yielded a better fit with the data than was expected, as reflected by their relatively high \( R^2 \) values (the value of 0.75 reported in Figure 4 is typical). Consequently, we decided to use their slopes as an (inverse) estimate of the unit shortage cost rate \( B \) corresponding to the current implicit trade-off. This regression study greatly facilitated the determination of what unit shortage cost rate values should be used initially.

4.2.3. Pilot Test. A key aspect of the implementation was to first go through a pilot period of several weeks before the full deployment of the new tool, during which the model output was to be systematically compared with the supply-routing decisions generated manually with the Balance Tool. The two main objectives pursued in this pilot were: (i) improve the software interface and functionalities with observations grounded in practice; and (ii) build an archive of input and output data in order to evaluate the qualitative and quantitative impact of the model. We now review the improvements that resulted from this pilot, and discuss next in §4.3 our quantitative impact assessment.

A first improvement consisted of eliminating the “flipped expedites” initially observed as part of the model output. This would arise when two sites in short supply were scheduled to receive at some point in the future some containers loaded by a common supplier in the same ship, and therefore with the same expected port arrival date. As illustrated in Figure 5, the model could then recommend use of expedited ground transportation (e.g., team truck) for all containers, but also switch the containers’ destinations. We found out, however, that for reasons not captured by the initial model (an expediting decision entails a bill of lading creation expense independently of the chosen destination), both the carriers and the supply-routing analysts prefer the simpler communications associated with a small number of destination changes, provided that this does not impact repositioning transportation costs. To capture this preference, we introduced the additional objective function term \( \sum_{i \in \mathcal{S}_t, j \in \mathcal{L}
abla \mathcal{C}_i} \gamma_{ij} m \), which essentially adds a dollar penalty for such destination changes. This modification indeed eliminated all such “flipped expedites” without affecting the repositioning costs of computed solutions, and thus improved the simplicity of the model output.

Another feature addition was motivated by issues occasionally found with the demand forecasts, in particular those covering the next seven days of demand. Because these were only updated once a week by the forecasting team using a fairly coarse method for disaggregating weekly forecasts into daily forecasts (see §3.1), the supply-routing analysts had designed an alternative forecasting method based on a simple time-series analysis of the actual parts consumption patterns observed in all relevant locations over the previous days. Whenever the predictions for the
next seven days of demand provided by that alternative method differed substantially from those provided by the forecasting team, the analysts tended to substitute their own daily forecast for the upcoming week. From an organizational standpoint, we believe that such forecast corrections are better done centrally by the forecasting team, perhaps by making better use of relevant decentralized input data such as these recent actual parts consumption patterns. However, we also recognized that some hurdles for implementing such coordination between the forecasting and the procurement teams would likely take time to overcome. This created a need for the software to support the ad hoc forecast correction practice just described. Specifically, we added a feature whereby the historical consumption of each part in every site is stored in a database covering the past 10 days of actual demand, and any major discrepancy between time-series based forecasts constructed from that database and those provided by the forecasting team is automatically highlighted. The analyst could then decide to automatically modify the model input data \( \bar{f}_t \) by replacing the original forecast for the next seven days of the horizon with the alternative one based on time series calculations.

Finally, an important implementation issue was to determine how large orders from retailers distributing Dell’s computers should be captured by the model. That question arose in a context of strategic change for Dell, which in 2007 started to develop distribution partnerships with large retailers in addition to its existing direct-sales channels. As a result, large customer orders for a single type of computer became more frequent. In particular, the supply-routing analysts were starting to receive notes informing them of committed schedules of large retailer deliveries for specific parts, which they were asked to plan for in addition to the existing forecasts for direct channels. The approach followed to account for these special orders consisted initially of simply adding these large customer orders to the existing forecasts. That method, however, resulted in transfers and diversions that were sometimes thought to be excessive. We determined that this resulted from a substantial overestimation of demand variability (and therefore expected shortages) in those sites, because the original demand model resulting from our forecast accuracy study evaluated the standard deviation of (cumulative) demand \( \sigma_t \) as a specified coefficient of variation times the corresponding forecast value \( \bar{f}_t \) (see §A.1 in the online appendix). This did not reflect the fact that these special retail orders have a substantially lower associated uncertainty than the direct channel orders. In order to address this issue, we created a feature to capture these special orders by modifying the means of demand forecasts \( \bar{f}_t \) correspondingly, but without affecting the forecast standard deviations \( \sigma_t \) (see §§4.1.1 and A.1 in the online appendix). This substantially reduced the seemingly unnecessary diversions and transfers.

4.3. Impact

4.3.1. Financial Impact Assessment. The quantitative impact evaluation of the model implementation described in §4.2 had to account for any effects on both repositioning transportation costs and part shortages—a reduction in repositioning costs alone is easily obtained by eliminating all ground expediting modes, for example, and may thus not represent an improvement if it is associated with an increase of shortages. Conversely, using only team trucks for all ground transportation would likely reduce part shortages, but also substantially increase repositioning costs. In order to construct an unambiguous measure of overall impact, one method considered was to use the current implicit shortage cost rate \( B \) (see §4.2) in order to estimate shortage costs, and then measure any changes in the sum of repositioning and shortage costs. Out of concern that the shortage cost rate was affected by subjective factors, Dell executives suggested that it would be desirable not to rely on its inferred value for impact evaluation purposes.

For this reason, we followed an alternative methodology consisting of computing a posteriori the reduction of repositioning transportation costs achieved by the optimization model relative to the legacy process, under the additional constraint that its output should result in shortages no higher than that achieved historically. We note that the underlying idea of constraining for comparison purposes a subset of performance dimensions for which the cost coefficients are hard to estimate in practice (e.g., backlog and ordering cost) has already been used by previous authors (e.g., Hopp et al. 1997). More specifically,
our assessment study is based on a representative group of monitors $K$ accounting for approximately half of total monitor sales over a period of 14 weeks in 2007 that preceded the implementation of the optimization-based process. We constructed a data set including every corresponding individual routing decision made by the analysts using the existing manual process and the Balance Tool described in §3, along with all the corresponding input data (inventory, forecasts, supply line) available at the time when these decisions were made. From that data set, we were able to construct an instance of the optimization problem (6)–(19) for every week that the analysts made a set of routing decisions for each monitor $k$ within that group. Note that the set of historical routing decisions recorded each week along with their corresponding expected shortage variables\textsuperscript{15} $\hat{v}_{il}$ (and associated secondary variables) constitute a feasible solution to that problem instance, with repositioning transportation cost $\hat{C}_k$ and total objective value $\hat{C}_k + B \sum_{l} \hat{v}_{il}$. Our impact assessment was then based on the solution to the modified optimization problem obtained by minimizing only the repositioning transportation cost components of (6) subject to the previous constraints (7)–(19), along with the additional constraint that $\sum_{l} v_{il} \leq \sum_{l} \hat{v}_{il}$. Denoting by $C_k$ the optimal value of the modified objective (i.e., the lowest repositioning transportation costs achievable when allowing no more expected shortages than achieved historically), Figure 6 contains a plot of the weekly repositioning transportation costs $\sum_{k \in K} \hat{C}_k$ incurred historically for all these parts as well as data labels indicating the corresponding relative total reduction $(\sum_{k \in K} \hat{C}_k - C_k)/(\sum_{k \in K} \hat{C}_k)$ achieved by the optimization model.

When summed over all 14 weeks of the data collection period defined above, the cumulative repositioning transportation cost savings associated with these optimization model runs represent approximately 46% of the total incurred historically, which provides an aggregate measure for the impact of this implementation. However, these relative savings seem to depend on the overall scarcity of supply, which is driven by the total quantities of components shipped by suppliers relative to demand and is thus exogenous to the routing model considered here. This can be seen from Figure 6, where the average weekly transportation costs plotted increase substantially in the second half of the data collection period (April 16–June 1) compared to its first half (February 26–April 13). This increase corresponds to an industry-wide shortage of glass substrates and color filters that began to impact the deliveries of flat panel monitors by Dell's suppliers in the middle of April of that year (Uno 2008), and in turn resulted in additional repositioning costs (in particular expediting). This affected the corresponding relative repositioning transportation cost savings, which can be evaluated independently for the first and second halves of the data collection period at 38% and 48%, respectively. These observations suggest that the lower of these last two numbers constitutes a better estimation for the relative repositioning transportation cost savings attributable to the optimization model during normal periods characterized by appropriate overall supply quantities. It should be noted, however, that the relative benefits derived from the optimization model seem to increase during severe shortage situations. Our explanation of this observation is that under the legacy process, the analysts are typically required then to execute a higher number of routing decisions every day, which leaves them less time for analysis. More generally, we wanted to identify

\textsuperscript{15} The hat symbol used in $\hat{v}_{il}$ and $\hat{C}_k$ emphasizes that these notations refer to the historical routing solution implemented by the analyst and its objective value.
the main qualitative reasons explaining the cost savings attributed to the optimization model. This led us to inspect the output of many of the optimization runs we conducted a posteriori as described above, and compare them with the historical supply-routing decisions made by the analysts with the same input data. Although we cannot provide an exhaustive description of these qualitative comparisons due to space constraints, the two representative examples discussed next in §§4.3.2 and 4.3.3 convey the main insights we obtained.

4.3.2. Qualitative Impact Assessment: First Example. Figure 7 shows a disguised and simplified version of the Balance Tool interface for a specific 15-inch monitor and a portion of the planning horizon as it appeared to the analyst on March 13, 2007. It shows a situation with an apparent excess of inventory relative to predicted demand in Nashville and Winston-Salem, and a shortage of inventory appearing in Austin and Reno at some point over the horizon considered. The situation in Austin would be particularly preoccupying at that point, because the shortages there are predicted to be higher and occur sooner than in Reno, which is only attributed a small demand forecast. Indeed, the (disguised) total number of expected shortages across all sites and days in the (complete) horizon predicted by our shortage model in the case where no action would be taken then is 50,000 unit-days of shortages (i.e., a measurement corresponding for example to predicted shortages of 2,500 units across all Dell sites on each day of a 20-day horizon). Note also that no upcoming deliveries of containers by suppliers for that component are visible within the planning horizon, leaving transfers as the only supply-routing decisions available.

On that day, the analyst ordered a transfer of 5,000 parts from Winston-Salem to Austin with three full special team trucks, for a (disguised) cost of $30,000. Winston-Salem was chosen as the location providing inventory because it had the largest amount of inventory available, both in absolute terms and when evaluated through DSI levels. Also, note that Winston-Salem has a forecasted demand about 30% lower than that of Nashville over the horizon considered, so that a transfer of a given quantity out of that facility results in a larger decrease of its DSI level. Finally, observe that no inventory was transferred to Reno, presumably because the potential corresponding transportation costs were not justified by the minor and distant predicted shortages at stake in that location. These decisions therefore suggest a good appreciation by the analyst of the overall directions, criticality, and time sensitivity of inventory imbalances across sites, and indeed decreased by 59% the total expected shortages predicted by our stochastic model, down to about 0.5%.
20,450 unit-days of shortages (note that because the overall supply quantity is exogenous, in many situations such as this one routing decisions may not reduce expected shortages below a certain level).

In the same situation, however, the optimization model recommended two regular truck transfers of 1,665 parts each (this quantity corresponds to a full truckload for that part) from Nashville and Winston-Salem, respectively, along with a schedule of subsequent milk-run transfers from Nashville to Austin containing each the maximum number of parts allowed—this solution is illustrated by Figure 8, which also shows the impact of these decisions on the predicted inventory and DSI levels. By construction, that solution achieved the same total expected shortages as the analyst’s; however, its total repositioning transportation cost amounts to $20,010, which represents a 33% reduction relative to the cost incurred historically. Remarkably, the total quantity of inventory transferred to Austin according to that solution (5,175) is very similar to that decided by the analyst, which is a by-product of the additional constraint on expected shortages. However, it exploits the lower transfer cost to Austin from Nashville than that from Winston-Salem, and is immune to considerations about the potential perceptions of high DSI levels in Winston-Salem—the reason here why the model does not recommend all inventory to be transferred from Nashville is that this would generate more expected shortages for that facility in the later part of the horizon, which is not shown in Figure 8. Another source of cost difference is the use of regular trucks as opposed to team trucks, which results from the model’s calculation that the corresponding lead-time difference of one day (delivery on March 15 instead of March 16) does not justify this additional cost in light of the predicted inventory situation in Austin over these couple of days—as seen in Figure 7, Austin is still predicted to have 5.6 DSI on March 19 absent any transfer decisions, also, this time period (March 15–19) is situated very early in the rolling horizon. As mentioned earlier, the analysts tend to infer the criticality of shortages based on DSI levels alone, whereas the model also takes into account whether that level is predicted early or late in the planning horizon, which affects the variability of the corresponding cumulative demand forecast, and therefore the estimation of expected shortages. As a result, for a given DSI level the analysts tend to overestimate expected shortages relative to the model in the early part of the horizon, and underestimate them in the more distant part. Finally, the model solution also exploits the lower transportation costs associated with milk-run transfers (RB) than with special trucks, even though the capacity restrictions of milk-run transfers result in a higher number of individual transfer decisions. In addition, milk-run transfers for a given leg are
only available on specific days, and therefore require the additional step of checking their current weekly schedule. These last observations explain why the analysts, who are subject to time pressure and human cognitive limitations, are unlikely to devise this type of transportation plan, which is more cost effective, but also more complex.

4.3.3. Qualitative Impact Assessment: Second Example. Figure 9 shows a disguised portion of the Balance Tool interface for a 20-inch monitor on the morning of April 17, 2007. That initial situation is characterized by insufficient inventory in Nashville, with the other facilities showing sufficient inventory levels that are initially comparable in terms of DSI. Also, there are planned container arrivals in Reno on May 7 (960 parts), and in Nashville on May 10 (3,564 parts, not visible in Figure 9). Absent any routing decisions in that initial situation, our stochastic model predicts a (disguised) total of 80,000 expected unit-days of shortages over the complete rolling horizon.

On that day, however, the analyst ordered an immediate transfer of 5,000 parts from Austin to Nashville using four team trucks, and a ground transportation expediting by team truck of all 3,564 parts (three containers) initially scheduled to arrive in Nashville on May 10. This advanced the arrival date of these parts to April 30, and thus mitigated the predicted shortages in Nashville from April 30 to May 10. The (disguised) total repositioning transportation cost of these decisions was $71,400. The optimization model solution for the same situation is illustrated by Figure 10, and consists of two immediate regular truck transfers of two full trucks each (2,500 parts) from Austin and Winston-Salem to Nashville, two milk-run transfers from Austin to Nashville, and a diversion to Nashville by rail of the 980 parts initially scheduled to arrive in Reno on May 7 (which postponed their arrival date to May 14 because of the longer lead time from California to Nashville). It achieves by construction the same number of expected unit-days of shortages, but costs 53% less in transportation than the manual solution implemented historically (or $33,450).

Observe that both the manual and the model solutions involve initial transfers to Nashville of the same quantity of parts (5,000). However, the model does not use the more costly team trucks for these transfers. Also, it spreads the origins of these transfers across two locations (Austin and Winston-Salem), which saves many expected shortages in the later part of the horizon in Austin: Note that with only 2,875 parts withdrawn from Austin in the model’s solution (against 5,000 for the manual one), the last day of the horizon portion shown in Figure 10 (May 8) shows only 6.2 predicted DSI, with continued demand and no subsequent container arrival in Austin in the time horizon beyond that—the situation in Austin from

<table>
<thead>
<tr>
<th>Demand forecast</th>
<th>Starting inventory</th>
<th>Planned deliveries</th>
<th>Delivery adjustment</th>
<th>Ending inventory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
<tr>
<td>Demand forecast</td>
<td>Starting inventory</td>
<td>Planned deliveries</td>
<td>Delivery adjustment</td>
<td>Ending inventory</td>
</tr>
</tbody>
</table>
then on is thus significantly worse with the analyst’s solution.

The recommendation of transfers from both Austin and Winston-Salem results from the convexity of expected shortages as a function of the negative of the inventory level (see discussion following (5)); from that property total expected shortages are lower when the “pain” (that is, low inventory levels) is shared across several locations rather than concentrated in one location only. Note also that, in contrast with the model’s decisions, the expediting decision by the analyst does not affect the shortages in Nashville beyond May 10 (the original container arrival date), a distant time period with high cumulative demand forecast variability. Finally, this example illustrates another important difference between the analysts’ heuristic and the model output. Specifically, analysts often needed to quickly evaluate the situation for many different parts and quickly determine whether any specific one deserved some attention. When doing so, they tended to inspect the total number of cells showing in red or yellow on each part’s Balance Tool for any day and location because of a low predicted DSI level (see Figure 2), and use that number as an overall indicator of criticality. By extension, they had come to also use that metric as a proxy for total expected shortages when making routing decisions.

Indeed, the first reaction of an analyst with whom we shared the model solution shown in Figure 10 was that it was worse than the one determined manually because it entails a larger area of the Balance Tool showing in red. Because of the convexity property just discussed, however, that metric can in fact lead to an increase of total expected shortages, as is shown by the simple example of two locations facing the same demand on a given day with a total of 3 DSI available for both (allocating 1.5 DSI to each minimizes total expected shortages but results in both location showing in red on the Balance Tool, whereas allocating all 3 DSI to a single location only puts the other one in the red). Finally, we believe that the performance differentials observed between the sensible but simple solutions generated by an experienced analyst and the corresponding model recommendations illustrated in Figures 8 and 10 make it implausible that a single near-optimal heuristic not relying on optimization methods may be developed for this problem.

5. Conclusion

At the time of writing, the process and optimization model for supply routing described in this paper have been continuously used by Dell for several quarters, with no plans for any significant changes. There are still several important improvement opportunities associated with this work, however, all of which motivate ongoing or future research. A first path is the implementation of unit shortage costs resulting
from a rigorous evaluation of the main cost components involved. The related study mentioned in §4.2 (Dhalla 2008) is now completed, and has already been used to generate more objective estimates for the value of the unit shortage cost rate $B$ that should be used in optimization model runs. In particular, that study has shown how $B$ should depend not only on the part, but also the location considered—a key factor is that one of the facilities in Dell’s network receives a larger proportion of option orders (e.g., for monitors only), for which the cost consequences of delays are milder than for complete system orders. That study also showed that in some cases our (standard) assumption of a linear structure for shortage costs (see §4.1.1) was fairly coarse, in part because the likelihood of order cancellation by a customer does not seem to increase linearly with the number of days of delay relative to the promised delivery date. This motivates ongoing efforts to develop and test a more realistic optimization model. Because of the likely associated increases in complexity and data maintenance requirements, however, it is not clear yet that this work will ultimately affect Dell’s practice.

Another opportunity would be to capture the dependencies across different parts when generating supply-routing decisions. A first avenue would be to extend the current model structure to components that, unlike monitors and chassis, are shipped in mixed containers of several part types. Although we did not focus on these “mixed” parts initially because they account for less transportation costs, that extension may still generate substantial savings over time. A more ambitious goal would be to take into account the inventory situation of several components likely to be required by the same customer orders when determining supply-routing (and more generally ordering) decisions for each. Interestingly, although the academic literature discusses the potential benefits of this practice (see Song and Zipkin 2003), it does not seem to have impacted operations at Dell yet, in part because of concerns linked to organizational incentives (e.g., two managers responsible for the supply of different components, both saving on expediting costs because of a simultaneous belief that the other manager’s component will be short anyway). Finally, another opportunity is to relax the assumption that demand in individual sites is exogenous, i.e., to jointly optimize the allocation of customer orders to manufacturing sites and inventory transfer decisions. The approach followed in the present paper seems correct as a first approximation because Dell ships directly to most of its customers. Therefore, the differences in (unit) outbound shipping costs for complete systems across different manufacturing sites are often substantially larger than the average (bulk) inbound transportation costs for individual components. In certain situations, however, for example, when transferring customer orders to a different factory may avoid some overtime, such joint optimization could prove profitable.

Despite all these limitations, the financial impact assessment presented earlier (the relative cost reduction estimates of 40% and 38% discussed in §§3.3 and 4.3 amount to a cumulative reduction of repositioning transportation costs for monitors by about 60% since the beginning of this collaboration) suggest that the model described in the present paper is already quite valuable for operational purposes. This is also supported by several recent developments at Dell. Specifically, Dell has committed some resources to implement that model in its European manufacturing network, where the supply chain structure is more complex because it involves several disembarkation ports where inventory can be held and rerouted. In addition, Dell is funding an effort to develop and test an extension of that model to compute recommended quantities, timing, and transportation modes for all component shipments between a global Asian warehouse and all of its manufacturing sites worldwide (see Foreman 2008). Finally, we note that many features of the model defined in §4 do not seem specific to Dell, so that part or all of it may also be useful in the future to other firms facing supply-routing and/or transportation-mode decisions.

Acknowledgments
J. Gallien’s current affiliation is London Business School, London NW1 4SA, United Kingdom.

Electronic Companion
An electronic companion to this paper is available on the Manufacturing & Service Operations Management website (http://msom.pubs.informs.org/ecompanion.html).

References


