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

John Foreman[1], Jérémie Gallien[2], Julie Alspaugh[3], Fernando Lopez[4], Rohit Bhatnagar[5], Chee Chong Teo[6], Charles Dubois[7]

July 22, 2008

Abstract

Dell's supply chain for desktops involves Asian vendors shipping components by sea to several U.S. plants. While suppliers are responsible for shipping enough inventory to meet total needs across all production sites, Dell can re-route 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 US-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 an MIP 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 re-positioning costs for monitors by about 60%.

1 Introduction

Dell's growth over the last ten 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 US-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.

In order to address this supply availability challenge, Dell has long relied on Vendor-Managed Inventory (VMI) relationships. That is, its suppliers are expected to maintain

[7] MIT Operations Research Center, Cambridge, MA 02142
[7] MIT Sloan School of Management, Cambridge, MA 02142 (corresponding author, e-mail: jgallien@mit.edu)
[7] Dell Computers, Inc., Round Rock, TX 78682
[7] Dell Computers, Inc., Round Rock, TX 78682
[7] Nanyang Technological University, Singapore
[7] Nanyang Technological University, Singapore
[7] Ecole des Mines de Paris, France

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 warehouse pick process), and suppliers are mostly free to follow any schedule of shipments as long as it meets some minimum service level[8]. In order 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 transhipments 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 US 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 transhipments (*transfers*) of on-hand inventory between its facilities. The available transportation modes for transfers include a set schedule of pre-contracted truck "milk runs" between Dell US

[8] Supplier DSI targets in individual locations constitute widely followed but non-contractual 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).

facilities, which have low relative cost but limited capacity, as well as special single or team trucks contracted on the spot[9]. Figure 1 illustrates the supply-chain structure and the associated supply routing decisions just defined, and also shows the four main locations in the US where chassis and monitors are shipped for assembly and/or inclusion into customer orders as well as typical transportation lead-times.

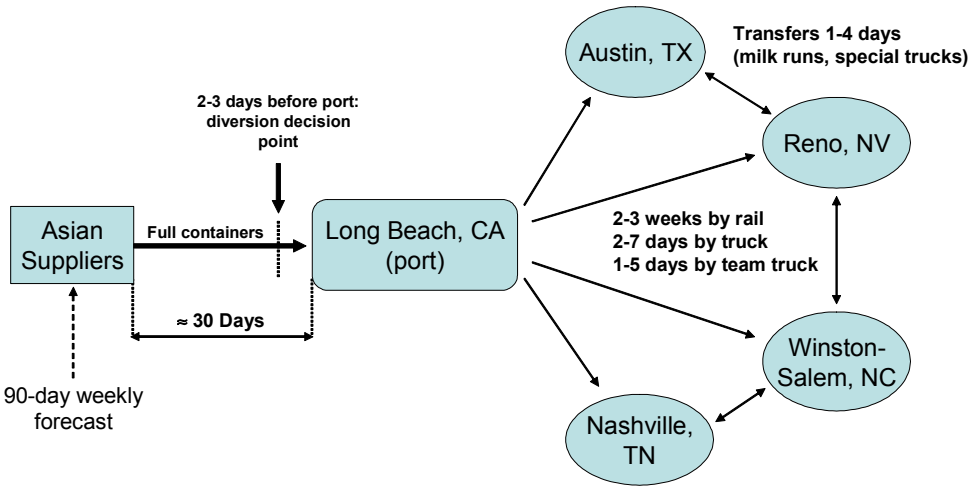


Figure 1: Dell's Supply-Chain Structure and Supply Routing Decisions for US-Bound Desktop Chassis and Monitors.

The volume of material continuously going through the supply chain just described is very significant: a rough estimate from Dell's public 10-K 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 US 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: while 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 [...]." 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 US-bound monitors. It contributes to the operations

[9] 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 through they all actually rely on the same type of transportation vehicle.

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 that 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* and 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 an 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, and thus capturing Dell's actual demand uncertainty in a model that is otherwise deterministic. Next, section §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., $d \triangleq \mathbb{E}[\mathbf{d}]$. 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 in order 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 *re-positioning 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 US destination originally intended) and are included in the "on dock" price per part charged by suppliers. Re-positioning 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, CA exactly matches the definition of a "stockless depot" considered in that paper[10]. Consequently, many of the results and insights described in the body of literature on multi-echelon inventory allocation that started with that seminal paper (see Axsäter, Marklund and Silver 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, Hsu and Sethi (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,

[10] An important difference however is that Dell only performs the *allocation* function (splitting incoming quantities among final destinations) and has delegated the *ordering* function (determining incoming quantities) to its suppliers.

in that it focuses on how the inventory ordering decisions, which we assume 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, Muckstadt and Rappold (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 newsboy-like 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 transhipments, considers a single expedited transportation mode, assumes a linear transportation cost structure and ignores transportation scheduling and capacity constraints. These differences are material, as the solution approach ultimately followed in Caggiano, Muckstadt and Rappold (2006) consists of developing heuristic solutions by exploiting the structural properties that can be established in their setting, while 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

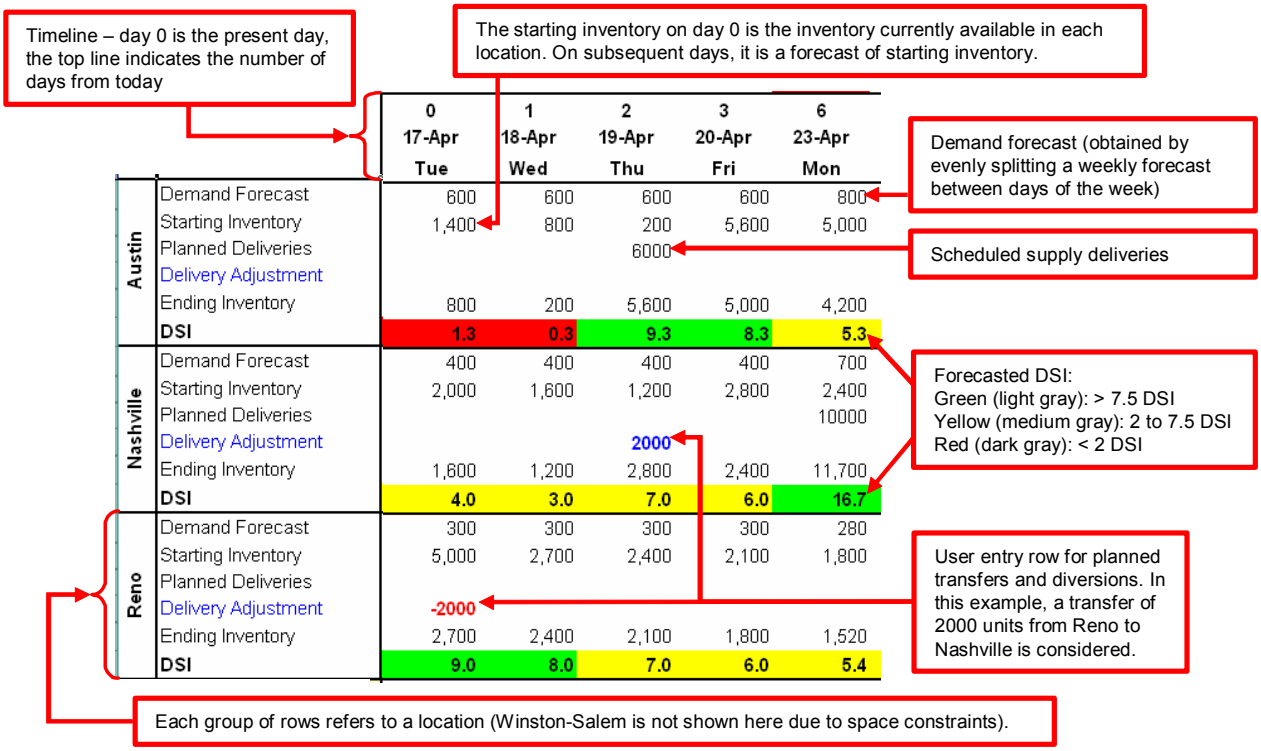
3.1 Development The first phase of this project, which is described more extensively in Reyner (2006), started in the Spring of 2005. Its goal was to correct an observed increase in expediting and transfer costs for large components by coordinating the associated decision process across the relevant 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 relevant information in order to make and implement all supply routing decisions. The primary objective specified for this new position consisted of reducing the re-positioning 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, which 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)[11]. 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

[11] The red, yellow and green colors from the original tool appear respectively as dark, medium and light gray cell backgrounds on Figure 2.

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 cut-off point of 2-3 days before port, assuming the longest possible ground transportation lead-time and then adding an additional time buffer.



		0 17-Apr Tue	1 18-Apr Wed	2 19-Apr Thu	3 20-Apr Fri	6 23-Apr Mon
Austin	Demand Forecast	600	600	600	600	800
	Starting Inventory	1,400	800	200	5,600	5,000
	Planned Deliveries			6000		
	Delivery Adjustment					
	Ending Inventory	800	200	5,600	5,000	4,200
	DSI	**1.3**	**0.3**	**9.3**	**8.3**	**5.3**
Nashville	Demand Forecast	400	400	400	400	700
	Starting Inventory	2,000	1,600	1,200	2,800	2,400
	Planned Deliveries					10000
	Delivery Adjustment			2000		
	Ending Inventory	1,600	1,200	2,800	2,400	11,700
	DSI	**4.0**	**3.0**	**7.0**	**6.0**	**16.7**
Reno	Demand Forecast	300	300	300	300	280
	Starting Inventory	5,000	2,700	2,400	2,100	1,800
	Planned Deliveries					
	Delivery Adjustment	-2000				
	Ending Inventory	2,700	2,400	2,100	1,800	1,520
	DSI	**9.0**	**8.0**	**7.0**	**6.0**	**5.4**

Figure 2: Visualization of Dynamic Routing Decisions with the Balance Tool (simplified version, adapted from Reyner 2006)

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 non-explicit 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 re-positioning transportation costs. Because the part shortages were generally perceived to have decreased during the pilot, the re-positioning 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 re-positioning 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 only involve 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: while 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 re-positioning 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 upfront 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. Since 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, since 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 be only 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 cutoff 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 re-positioning 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). While

expressing the re-positioning 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 $B \sum_{t\in\mathcal{T},\ell\in\mathcal{L}} v_{t\ell}$ for the total expected shortage costs predicted in all facilities $\ell \in \mathcal{L} \triangleq \{$Austin, Nashville, Reno, Winston-Salem$\}$ for a specific part over the rolling horizon $t \in \mathcal{T} \triangleq \{1, .., T\}$ considered, where B is a unit daily shortage cost rate and $v_{t\ell}$ 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 $v_{t\ell}$ 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, as 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 $\sigma_{t\ell}$ for the stochastic model

$$\sum_{k=1}^{t} \mathbf{d}_{k\ell} \sim N(\sum_{k=1}^{t} f_{k\ell}, \sigma_{t\ell}), \tag{1}$$

where: $\mathbf{d}_{t\ell}$ is the random variable representing demand on day t for a given part in a given location $\ell \in \{$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 $\bar{\mathbf{d}}_{t\ell} \triangleq \sum_{k=1}^{t} \mathbf{d}_{k\ell}$ is the

cumulative demand for the next t days; $N(f, \sigma)$ refers to a normal distribution with mean f and standard deviation σ; $f_{t\ell}$ is the (deterministic) forecast of the same quantity generated by Dell and provided to the supply-chain analyst on day 1, so that $\bar{f}_{t\ell} \triangleq \sum_{k=1}^{t} f_{k\ell}$ is the corresponding cumulative forecast of demand up to day t; and finally $\sigma_{t\ell}$ is the standard deviation of the forecasting error $\bar{\mathbf{d}}_{t\ell} - \bar{f}_{t\ell}$. Note that the forecasting error study mentioned above did identify some systematic biases. These biases were ignored however, since they were relatively small and convincingly explained by the forecasting team. The notations $f_{t\ell}$ and $d_{t\ell} \triangleq \mathbb{E}[\mathbf{d}_{t\ell}]$ 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:

$$\mathbf{I}_{(t+1)\ell} = I_{1\ell} + \sum_{k=1}^{t} s_{k\ell} - \bar{\mathbf{d}}_{t\ell} \text{ for } t \geq 1, \tag{2}$$

where: $\mathbf{I}_{t\ell}$ 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 $I_{1\ell} = \mathbf{I}_{1\ell}$ is deterministic input data), and $s_{t\ell}$ 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 $s_{t\ell}$ 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[12]. 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 further.

Next, we approximate the shortage level $\mathbf{v}_{t\ell}$ for day t in location ℓ predicted at the beginning of day 1 as

$$\mathbf{v}_{t\ell} \triangleq (\mathbf{I}_{t\ell} - \mathbf{d}_{t\ell})^{-}. \tag{3}$$

Note that the expression $\mathbf{I}_{t\ell} - \mathbf{d}_{t\ell}$ 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

[12] 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.

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[13]. Substituting (2) and (1) in (3) yields

$$\mathbf{v}_{t\ell} \sim [N(I_{t\ell} - f_{t\ell}, \sigma_{t\ell})]^- \text{ for } t \geq 1, \tag{4}$$

which characterizes the distribution of these shortages in terms of decision variables and input data[14]. The expectation of the random variable in (4) is thus given by the standard normal loss function

$$v_{t\ell} = \sigma_{t\ell}\phi\left(\frac{f_{t\ell} - I_{t\ell}}{\sigma_{t\ell}}\right) + (f_{t\ell} - I_{t\ell})\Phi\left(\frac{f_{t\ell} - I_{t\ell}}{\sigma_{t\ell}}\right), \tag{5}$$

where ϕ and Φ are the standard normal p.d.f. and c.d.f., respectively.

Our final goal is to embed the function of $f_{t\ell} - I_{t\ell}$ 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 $v_{t\ell}$ exceed a number of supporting tangents to the 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 pre-computation of the slopes $a_{t\ell p}$ and intercepts $b_{t\ell p}$ 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 $[I_{t\ell}^{LB}, I_{t\ell}^{UB}]$ for $I_{t\ell}$ that only depends on input data. A tight upper bound $I_{t\ell}^{UB}$ 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 re-routing towards 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 $I_{t\ell}^{LB}$ 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[15].

[13] The alternative approximation $\mathbf{v}_{t\ell} \triangleq (\mathbf{I}_{t\ell} + s_{t\ell})^-$ is equally tractable and is the most optimistic in the sense just defined. It thus constitutes a bound which allowed us to verify that the assumption reflected by (3) was relatively immaterial.

[14] Random variables and their distributions are used interchangeably in (4), as no related ambiguity arises here.

[15] The equations for $I_{t\ell}^{LB}$ and $I_{t\ell}^{UB}$ in terms of input data are straightforward and omitted here.

Finally, we calculate iteratively a discrete set of sampling points $\mathcal{P}_{t\ell} \subset [I_{t\ell}^{LB}, I_{t\ell}^{UB}]$ indexed by p, and the corresponding slopes $a_{t\ell p}$ and intercepts $b_{t\ell p}$ of the tangents to the r.h.s. of (5) in those points, using numerical implementations of ϕ and Φ along with the maximum error rule method described in Rote (1992)[16]. 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 $\mathcal{T} \triangleq \{1, .., T\}$, and the set of relevant locations is $\mathcal{L} \triangleq \{$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 $\ell_i \in \mathcal{L}$ 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 2 or 3 days away from port) form a subset $\mathcal{C}^{RT} \subset \mathcal{C}$, while the containers in the complement set $\mathcal{C}^{NRT} \triangleq \mathcal{C} \backslash \mathcal{C}^{RT}$ may no longer be re-routed before they arrive at their destination[17]. 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 new 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 $\mathcal{C} = \cup_{j \in \mathcal{J}} \mathcal{C}_j$.

Current Net Inventory The sum of on-hand inventory currently available (that is at the

[16] This algorithm initiates with $\mathcal{P}_{t\ell} = \{I_{t\ell}^{LB}, I_{t\ell}^{UB}\}$. In each iteraction, tangents are constructed for each new point in $\mathcal{P}_{t\ell}$, and the x-axis values of the intersection of tangents corresponding to adjacent points in $\mathcal{P}_{t\ell}$ 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.

[17] The superscripts RT and NRT stand for *routable* and *non-routable*, respectively.

beginning of day 1 of the rolling horizon) minus backorders in each location $\ell \in \mathcal{L}$ is denoted $I_{1\ell}$.

Demand Forecast The forecast of demand in location $\ell \in \mathcal{L}$ during day t is denoted $f_{t\ell}$, while the cumulative forecast of demand from day 1 to day t (included) is denoted $\bar{f}_{t\ell}$.

Container Ground Transportation Modes Ground transportation modes between the port and Dell's facilities are indexed by $m \in \mathcal{M}^{RT} \triangleq \{$rail,single truck,team truck$\}$ and characterized for each destination $\ell \in \mathcal{L}$ by a cost per container $c^{RT}_{\ell m}$ and an average lead-time $L^{RT}_{\ell m}$ (expressed in days). Note that the re-positioning transportation costs incurred when diverting a container i to a destination ℓ with transportation mode m are $c^{RT}_{\ell m} - c^{RT}_{\ell_i \text{rail}}$ in addition to any bill of lading creation fee involved. That is, the embedded costs $c^{RT}_{\ell_i \text{rail}}$ corresponding to a shipment by rail to the original destination ℓ_i must be subtracted since 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 $i \in \mathcal{C}^{RT}$ is $A^{LB}_i + L^{RT}_{\ell m}$.

Special Transfers Special transfers of inventory between two facilities ℓ and ℓ' in $\mathcal{L}$ are characterized by their expediting mode $m \in \mathcal{M}^{SP} \triangleq \{$single truck,team truck$\}$, their cost per truck $c^{SP}_{\ell\ell' m}$ and their lead-time $L^{SP}_{\ell\ell' m}$.

Milk Run Transfers Milk run transfers of inventory from facility ℓ to facility ℓ' are characterized by their schedule of departures $S^{MR}_{t\ell\ell'}$ (equal to 1 if a run from ℓ to ℓ' is scheduled on day t and 0 otherwise), their cost per pallet $c^{MR}_{\ell\ell'}$, the maximum number of pallets of a given part allowed in each run R, and their lead-time $L^{MR}_{\ell\ell'}$. 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[18].

Shortage Cost Parameters They include the unit daily shortage cost rate B as well as

[18] 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.

each slope $a_{t\ell p}$ and intercept $b_{t\ell p}$of the approximating tangents to the expected shortage cost function indexed by $p \in \mathcal{P}_{t\ell}$ for each day t and location ℓ. An associated upper bound for the absolute value of the expected net inventory level at the beginning of day t in location ℓ is $M_{t\ell} = \max(|I_{t\ell}^{UB}|, |I_{t\ell}^{LB}|)$ (see §4.1.1).

Decision Variables:

Container Routing Binary variables $y_{i\ell m}$ are set to 1 if container $i \in \mathcal{C}^{RT}$ is routed from the port to facility ℓ using transportation mode $m \in \mathcal{M}^{RT}$, and 0 otherwise. In addition, binary variables $z_{j\ell m}$ take the value 1 if at least one container $i \in \mathcal{C}_j$ 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 $X_{t\ell\ell' m}$ 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 $x_{t\ell\ell' m}$ 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 $w_{t\ell\ell' m} \leq Q$ represent the number of parts carried in that truck[19].

Milk Run Transfers Integer variables $r_{t\ell\ell'}$ represent the number of pallets included in the run from facility ℓ to facility ℓ' on day t.

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

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

[19] The integrality of $w_{t\ell\ell' m}$ is immaterial in light of the quantities at stake here.

Formulation:

$$\text{Min} \sum_{i\in\mathcal{C}^{RT},\ell,m\in\mathcal{M}^{RT}} (c^{RT}_{\ell m} - c^{RT}_{\ell_i \text{rail}}) y_{i\ell m} + \sum_{j\in\mathcal{J}} c^{BL} \left(\sum_{\ell,m\in\mathcal{M}^{RT}} z_{j\ell m} - 1 \right)$$
$$+ \sum_{\{t,\ell,\ell',m\in\mathcal{M}^{SP}:\ell\neq\ell'\}} c^{SP}_{\ell\ell' m} \left(X_{t\ell\ell' m} + x_{t\ell\ell' m}\right) + \sum_{\{t,\ell,\ell':\ell\neq\ell'\}} c^{MR}_{\ell\ell'} r_{t\ell\ell'} + B \sum_{t,\ell} v_{t\ell} \tag{6}$$

subject to:

$$I_{t\ell} = I_{1\ell} - \bar{f}_{(t-1)\ell} + \sum_{\{i\in\mathcal{C}^{NRT}:\ell_i=\ell, A_i\leq t-1\}} q_i + \sum_{\{(i,m)\in\mathcal{C}^{RT}\times\mathcal{M}^{RT}:A^{LB}_i + L^{RT}_{\ell m}\leq t-1\}} q_i y_{i\ell m}$$
$$+ \sum_{\{(\tau,\ell',m)\in\mathcal{T}\times\mathcal{L}\times\mathcal{M}^{SP}:\ell'\neq\ell,\tau+L^{SP}_{\ell'\ell m}\leq t-1\}} (QX_{\tau\ell'\ell m} + w_{\tau\ell'\ell m}) + \sum_{\{(\tau,\ell')\in\mathcal{T}\times\mathcal{L}:\ell'\neq\ell,\tau+L^{MR}_{\ell'\ell}\leq t-1\}} Wr_{\tau\ell'\ell}$$
$$- \sum_{\{(\tau,\ell',m)\in\mathcal{T}\times\mathcal{L}\times\mathcal{M}^{SP}:\ell'\neq\ell,\tau\leq t-1\}} (QX_{\tau\ell\ell' m} + w_{\tau\ell\ell' m}) - \sum_{\{(\tau,\ell')\in\mathcal{T}\times\mathcal{L}:\ell'\neq\ell,\tau\leq t-1\}} Wr_{\tau\ell\ell'} \tag{7}$$

$$\sum_{m\in\mathcal{M}^{RT},\ell} y_{i\ell m} = 1 \qquad \forall i\in\mathcal{C}^{RT} \tag{8}$$

$$z_{j\ell m} \geq y_{i\ell m} \qquad \forall j\in\mathcal{J}, \ell\in\mathcal{L}, m\in\mathcal{M}^{RT}, i\in\mathcal{C}_j \tag{9}$$

$$I_{t\ell} = I^+_{t\ell} - I^-_{t\ell} \qquad \forall t\in\mathcal{T}, \ell\in\mathcal{L} \tag{10}$$

$$I^+_{t\ell} \leq M_{t\ell} I^1_{t\ell} \qquad \forall t\in\mathcal{T}, \ell\in\mathcal{L} \tag{11}$$

$$I^-_{t\ell} \leq M_{t\ell}(1 - I^1_{t\ell}) \qquad \forall t\in\mathcal{T}, \ell\in\mathcal{L} \tag{12}$$

$$\sum_{\{(\ell',m)\in\mathcal{T}\times\mathcal{L}\times\mathcal{M}^{SP}:\ell'\neq\ell\}} (QX_{t\ell\ell' m} + w_{t\ell\ell' m}) + \sum_{\{\ell'\in\mathcal{L}:\ell'\neq\ell\}} Wr_{t\ell\ell'} \leq I^+_{t\ell} \qquad \forall t\in\mathcal{T}, \ell\in\mathcal{L} \tag{13}$$

$$w_{t\ell\ell' m} \leq Qx_{t\ell\ell' m} \qquad \forall t\in\mathcal{T}, (\ell,\ell')\in\mathcal{L}^2, m\in\mathcal{M}^{SP} \tag{14}$$

$$r_{t\ell\ell'} \leq RS^{MR}_{t\ell\ell'} \qquad \forall t\in\mathcal{T}, (\ell,\ell')\in\mathcal{L}^2 \tag{15}$$

$$v_{t\ell} \geq a_{t\ell p}(f_{tl} - I_{t\ell}) + b_{t\ell p} \qquad \forall t\in\mathcal{T}, \ell\in\mathcal{L}, p\in\mathcal{P}_{t\ell} \tag{16}$$

$$y_{i\ell m}, z_{j\ell m}, x_{t\ell\ell' m}, I^1_{t\ell} \in \{0,1\} \tag{17}$$

$$X_{t\ell\ell' m}, r_{t\ell\ell'} \in \mathbb{N} \tag{18}$$

$$w_{t\ell\ell' m}, I_{t\ell}, I^+_{t\ell}, I^-_{t\ell} \geq 0 \tag{19}$$

The objective (6) is the sum of all re-positioning 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 re-positioning 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. While it would be straightforward to add a term summing the on-hand inventory variables $I^+_{t\ell}$ 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. While 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 re-positioning 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 $I_{t\ell}$ and the inventory currently available ($I_{1\ell}$), the demand forecasts ($\bar{f}_{(t-1)\ell}$), the pipeline of non-routable containers ($\sum_{\{i\in\mathcal{C}^{NRT}:\ell_i=\ell,:A_i\leq t-1\}} q_i$), and all the supply routing decisions considered (all subsequent terms in the r.h.s.). Constraints (8) ensure that every container is routed to exactly one destination through one transportation mode. Constraints (9) ensure that the term $\sum_{\ell,m} z_{j\ell m} - 1$ 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 $I^+_{t\ell}$, $I^-_{t\ell}$ and $I^1_{t\ell}$ correspond to the positive part, negative part and non-negativity indicator of variable $I_{t\ell}$, 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 $S^{MR}_{t\ell\ell'}$ are only introduced here to simplify exposition, as for implementation purposes it is more computationally efficient to only de-

fine variables $r_{t\ell\ell'}$ 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 $v_{t\ell}$ 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[20]. 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 pre-processing necessary to compute the piecewise linear approximations to the expected shortage cost functions (see §4.1.1) was implemented using Microsoft Visual Basic[21]. The creation of this software tool from complete specifications required approximately 6 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

[20] Figure 3 notes: The entries "Red Ball" appearing in the 7th column of the table under the heading "Mode" refer to the name of a carrier contracted by Dell to perform the milk runs between sites described in §1, which has become synonymous with that transportation mode within Dell. The different container quantities seen for part number FG645 stem from the use of both 20' and 40' containers.

[21] The daily execution of this process at Dell for all monitors requires requires approximately 5 minutes for accessing and loading all the required data and performing the pre-processing necessary to create the updated optimization problem instances, and another 5 minutes for computing solutions to these problems.

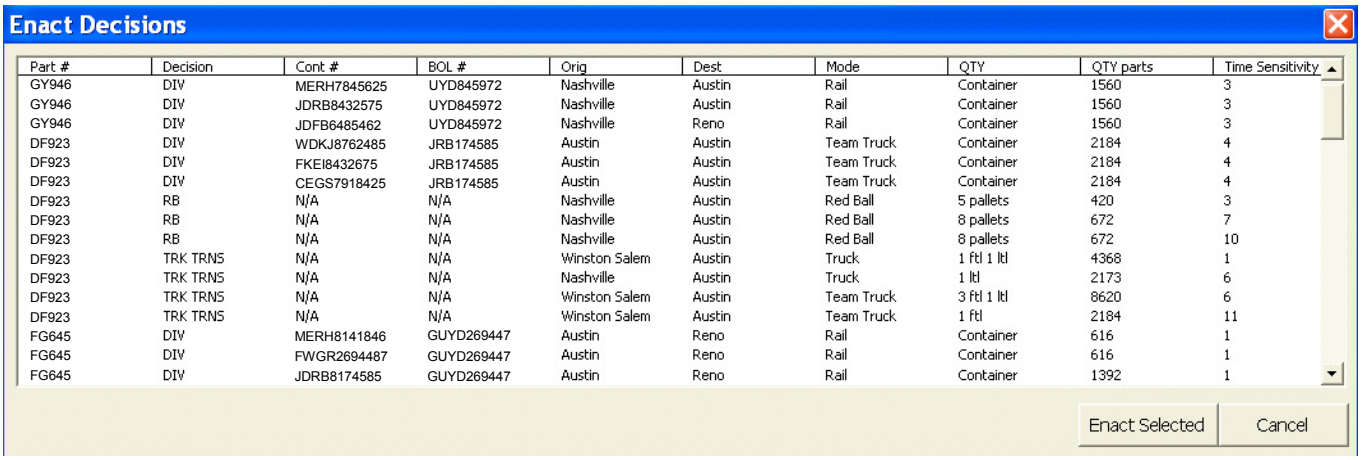


Figure 3: Output Interface of the Model Software Implementation

and representative collection of input data sets on which we performed many optimization runs. This demonstrated that while 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 criteria, and we did not further investigate the computational solution time for this problem[22].

4.2.2 Input Data Collection and Shortage Cost Estimation While 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 re-positioning costs involved, which justified the approximations $c^{RT}_{\ell m} - c^{RT}_{\ell_i \text{rail}} \approx c^{RT}_{\ell m}$ for $m \in \mathcal{M}^{RT} \backslash \{\text{rail}\}$, and $c^{RT}_{\ell \text{rail}} - c^{RT}_{\ell_i \text{rail}} \approx 0$ 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 $c^{RT}_{\ell m} = 0$ for $m =$"rail".

The most important implementation hurdle faced at this point however was to determine

[22] While 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).

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 re-positioning 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, while 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 towards an objective value for that parameter. The strategy decided then was two-fold: 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 Dhalla 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. While this short term strategy would not be necessarily 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 re-positioning 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 re-positioning 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 re-positioning costs incurred (independent variable) for each part over that dataset, which spanned several weeks of decisions. An interesting aspect of these regressions is that their fit provided a measure

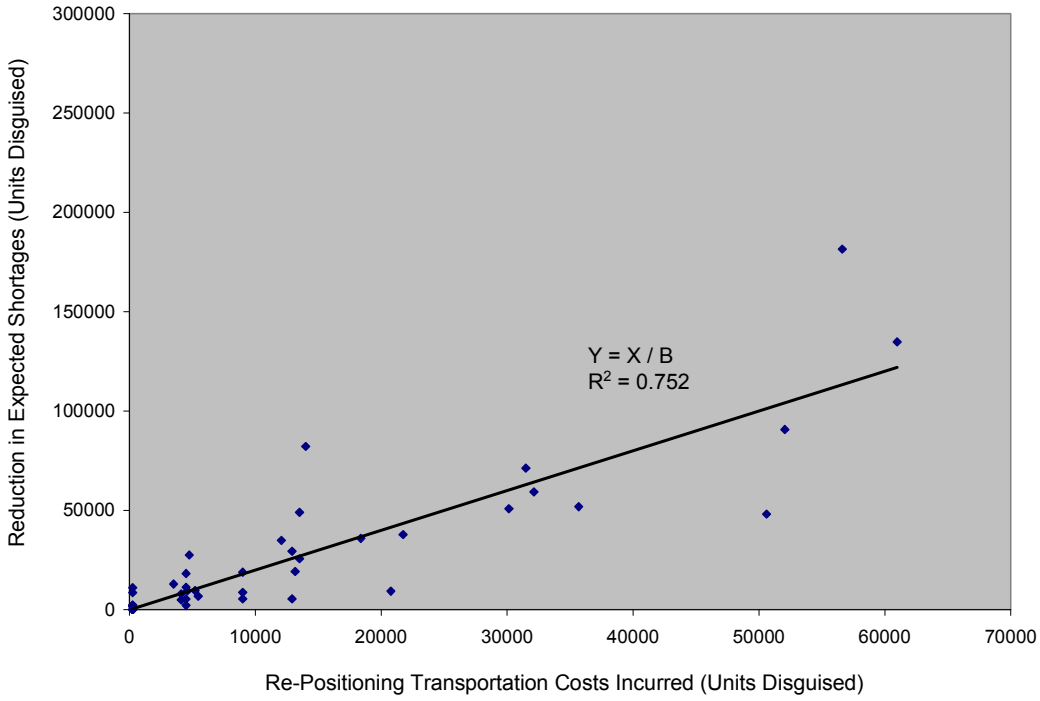


Figure 4: Linear Regression of Expected Shortage Reductions Achieved Against Re-Routing Transportation Costs Incurred

for the consistency of the analysts' historical decisions with respect to the trade-off between re-positioning 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 .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 to use 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 this does not impact re-positioning transportation costs. To capture this preference we introduced the additional objective function term $\sum_{i \in \mathcal{C}^{RT}, \ell \in \mathcal{L} \setminus \{\ell_i\}} y_{i\ell m}$, which essentially adds a dollar penalty for such destination changes. This modification indeed eliminated all such "flipped expedites" without affecting the re-positioning costs of computed solutions, and thus improved the simplicity of the model output.

Part #	Decision	Cont #	BOL #	Orig	Dest	Mode	QTY	QTY parts	Time Sensitivity
657FT	DIV	YRLF8463748	RYT549786	Nashville	Austin	Team Truck	Container	2184	0
657FT	DIV	HCEF5798165	RYT549722	Austin	Nashville	Team Truck	Container	2184	0

Figure 5: Example of Model Output with Swapped Container Destinations

Another feature addition was motivated by issues occasionally found with the demand forecasts, in particular those covering the next 7 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 ten days of actual demand, and any major discrepancy between a 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 $\bar{f}_{t\ell}$ by replacing the original forecast for the next 7 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, as the original demand model resulting from our forecast accuracy study evaluated the standard deviation of (cumulative) demand $\sigma_{t\ell}$ as a specified coefficient of variation times the corresponding forecast value $\bar{f}_{t\ell}$ (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 $\bar{f}_{t\ell}$ correspondingly, but without affecting the forecast

standard deviations $\sigma_{t\ell}$ (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 re-positioning transportation costs and part shortages – a reduction in re-positioning 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 re-positioning 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 re-positioning and shortage costs. Out of concern that the shortage cost rate was affected by subjective factors, Dell executives suggested that it would be desirable to not 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 re-positioning 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 $\mathcal{K}$ 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 dataset 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 dataset, 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 corre-

sponding expected shortage variables[23] $\hat{v}^k_{t\ell}$ (and associated secondary variables) constitute a feasible solution to that problem instance, with re-positioning transportation cost $\hat{C}_k$ and total objective value $\hat{C}_k + B\sum_{t,\ell}\hat{v}^k_{t\ell}$. Our impact assessment was then based on the solution to the modified optimization problem obtained by minimizing only the re-positioning transportation cost components of (6) subject to the previous constraints (7)-(19) along with the additional constraint that $\sum_{t,\ell} v^k_{t\ell} \leq \sum_{t,\ell}\hat{v}^k_{t\ell}$. Denoting by C_k the optimal value of the modified objective (i.e. the lowest re-positioning transportation costs achievable when allowing no more expected shortages than achieved historically), Figure 6 contains a plot of the weekly re-positioning transportation costs $\sum_{k\in\mathcal{K}}\hat{C}_k$ incurred historically for all these parts as well as data labels indicating the corresponding relative total reduction $\frac{\sum_{k\in\mathcal{K}}\hat{C}_k - C_k}{\sum_{k\in\mathcal{K}}\hat{C}_k}$ achieved by the optimization model.

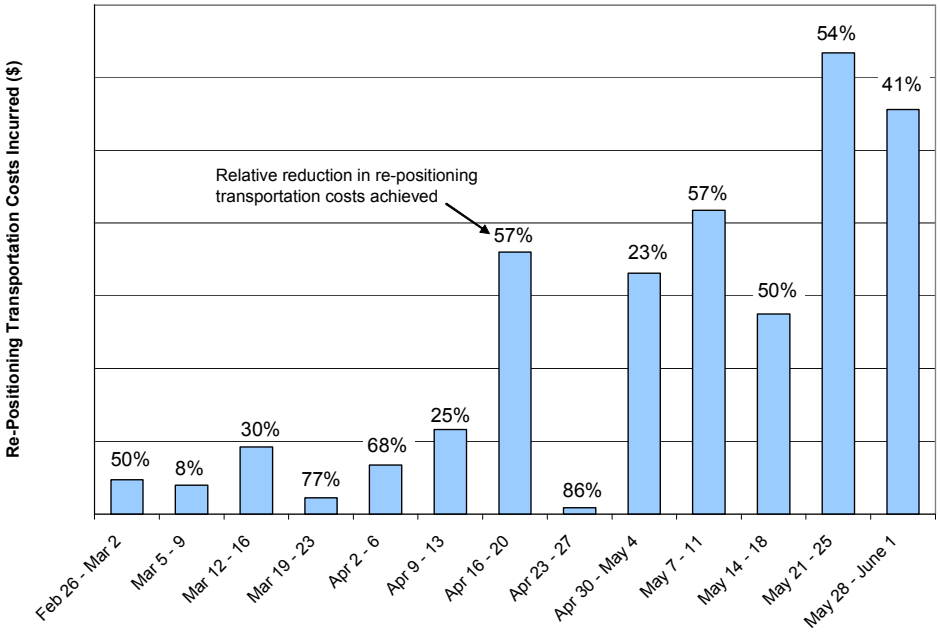


Figure 6: Weekly Re-Positioning Transportation Costs Incurred and Relative Reduction Achieved by the Optimization Model for Six Monitors from February 26 to June 1, 2007

When summed over all 14 weeks of the data collection period defined above, the cumulative re-positioning 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

[23] The hat symbol used in $\hat{v}^k_{t\ell}$ and $\hat{C}_k$ emphasizes that these notations refer to the historical routing solution implemented by the analyst and its objective value.

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 that year (Uno 2008), and in turn resulted in additional re-positioning costs (in particular expediting). This affected the corresponding relative re-positioning 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 re-positioning 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 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 but qualitatively representative 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, as 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.

		0 14-Mar Wed	1 15-Mar Thu	2 16-Mar Fri	5 19-Mar Mon	6 20-Mar Tue	7 21-Mar Wed	8 22-Mar Thu	9 23-Mar Fri	12 26-Mar Mon	13 27-Mar Tue	14 28-Mar Wed	15 29-Mar Thu	16 30-Mar Fri	19 2-Apr Mon	20 3-Apr Tue
Austin	Demand Forecast	566	566	566	656	656	656	656	656	656	656	656	656	656	656	656
	Starting Inventory	6,000	5,434	4,868	4,302	3,646	2,990	2,335	1,679	1,023	367	-289	-945	-1,601	-2,257	-2,913
	Planned Deliveries															
	Delivery Adjustment															
	Ending Inventory	5,434	4,868	4,302	3,646	2,990	2,335	1,679	1,023	367	-289	-945	-1,601	-2,257	-2,913	-3,569
	DSI	**9.6**	**8.6**	**7.6**	**5.6**	**4.6**	**3.6**	**2.6**	**1.6**	**0.6**	**-0.4**	**-1.4**	**-2.4**	**-3.4**	**-4.4**	**-5.4**
Nashville	Demand Forecast	348	348	348	393	393	393	393	393	393	393	393	393	393	393	393
	Starting Inventory	19,320	18,972	18,624	18,276	17,883	17,490	17,097	16,704	16,311	15,918	15,526	15,133	14,740	14,347	13,954
	Planned Deliveries															
	Delivery Adjustment															
	Ending Inventory	18,972	18,624	18,276	17,883	17,490	17,097	16,704	16,311	15,918	15,526	15,133	14,740	14,347	13,954	13,561
	DSI	**54.5**	**53.5**	**52.5**	**45.5**	**44.5**	**43.5**	**42.5**	**41.5**	**40.5**	**39.5**	**38.5**	**37.5**	**36.5**	**35.5**	**34.5**
Reno	Demand Forecast	33	33	33	41	41	41	41	41	41	41	41	41	41	41	41
	Starting Inventory	639	606	573	540	499	458	417	375	334	293	252	211	169	128	87
	Planned Deliveries															
	Delivery Adjustment															
	Ending Inventory	606	573	540	499	458	417	375	334	293	252	211	169	128	87	46
	DSI	**18.5**	**17.5**	**16.5**	**12.1**	**11.1**	**10.1**	**9.1**	**8.1**	**7.1**	**6.1**	**5.1**	**4.1**	**3.1**	**2.1**	**1.1**
Winston Salem	Demand Forecast	226	226	226	240	240	240	240	240	240	240	240	240	240	240	240
	Starting Inventory	21,209	20,983	20,757	20,530	20,290	20,050	19,810	19,570	19,330	19,090	18,850	18,609	18,369	18,129	17,889
	Planned Deliveries															
	Delivery Adjustment															
	Ending Inventory	20,983	20,757	20,530	20,290	20,050	19,810	19,570	19,330	19,090	18,850	18,609	18,369	18,129	17,889	17,649
	DSI	**92.8**	**91.8**	**90.8**	**84.5**	**83.5**	**82.5**	**81.5**	**80.5**	**79.5**	**78.5**	**77.5**	**76.5**	**75.5**	**74.5**	**73.5**

Figure 7: Disguised and Simplified Copy of the Balance Tool Interface for a 15 inch Monitor on March 14, 2007

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 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 re-positioning 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

		0 14-Mar Wed	1 15-Mar Thu	2 16-Mar Fri	5 19-Mar Mon	6 20-Mar Tue	7 21-Mar Wed	8 22-Mar Thu	9 23-Mar Fri	12 26-Mar Mon	13 27-Mar Tue	14 28-Mar Wed	15 29-Mar Thu	16 30-Mar Fri	19 2-Apr Mon	20 3-Apr Tue
Austin	Demand Forecast	566	566	566	656	656	656	656	656	656	656	656	656	656	656	656
	Starting Inventory	6,000	5,434	4,868	7,992	7,336	7,040	6,385	5,729	5,433	4,777	4,481	3,825	3,169	2,873	2,217
	Planned Deliveries															
	Delivery Adjustment			3690		360			360		360			360		45
	Ending Inventory	5,434	4,868	7,992	7,336	7,040	6,385	5,729	5,433	4,777	4,481	3,825	3,169	2,873	2,217	1,606
	DSI	9.6	8.6	14.1	11.2	10.7	9.7	8.7	8.3	7.3	6.8	5.8	4.8	4.4	3.4	2.4
Nashville	Demand Forecast	348	348	348	393	393	393	393	393	393	393	393	393	393	393	393
	Starting Inventory	19,320	17,307	16,599	16,251	15,498	15,105	14,712	13,959	13,566	12,813	12,421	12,028	11,275	10,882	10,444
	Planned Deliveries	Truck														
	Delivery Adjustment	-1665	-360	Milk Run	-360	Milk Run		-360	Milk Run	-360	Milk Run		-360	Milk Run	-45	Milk Run
	Ending Inventory	17,307	16,599	16,251	15,498	15,105	14,712	13,959	13,566	12,813	12,421	12,028	11,275	10,882	10,444	10,051
	DSI	49.7	47.7	46.7	39.4	38.4	37.4	35.5	34.5	32.6	31.6	30.6	28.7	27.7	26.6	25.6
Reno	Demand Forecast	33	33	33	41	41	41	41	41	41	41	41	41	41	41	41
	Starting Inventory	639	606	573	540	499	458	417	375	334	293	252	211	169	128	87
	Planned Deliveries															
	Delivery Adjustment															
	Ending Inventory	606	573	540	499	458	417	375	334	293	252	211	169	128	87	46
	DSI	18.5	17.5	16.5	12.1	11.1	10.1	9.1	8.1	7.1	6.1	5.1	4.1	3.1	2.1	1.1
Winston Salem	Demand Forecast	226	226	226	240	240	240	240	240	240	240	240	240	240	240	240
	Starting Inventory	21,209	19,318	19,092	18,865	18,625	18,385	18,145	17,905	17,665	17,425	17,185	16,944	16,704	16,464	16,224
	Planned Deliveries		Truck													
	Delivery Adjustment	-1665														
	Ending Inventory	19,318	19,092	18,865	18,625	18,385	18,145	17,905	17,665	17,425	17,185	16,944	16,704	16,464	16,224	15,984
	DSI	85.4	84.4	83.4	77.6	76.6	75.6	74.6	73.6	72.6	71.6	70.6	69.6	68.6	67.6	66.6

Figure 8: Routing Decisions Recommended by the Optimization Model for the Example Illustrated by Figure 7

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 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 (3564 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 5000 parts from Austin to Nashville using 4 team trucks, and a ground transportation expediting by team truck of all 3564 parts (3 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

		0 17-Apr Tue	1 18-Apr Wed	2 19-Apr Thu	3 20-Apr Fri	6 23-Apr Mon	7 24-Apr Tue	8 25-Apr Wed	9 26-Apr Thu	10 27-Apr Fri	13 30-Apr Mon	14 1-May Tue	15 2-May Wed	16 3-May Thu	17 4-May Fri	20 7-May Mon	21 8-May Tue
Austin	Demand Forecast	612	612	612	612	758	758	758	758	758	775	775	775	775	775	803	803
	Starting Inventory	19,569	18,957	18,345	17,733	17,121	16,363	15,605	14,846	14,088	13,330	12,555	11,781	11,006	10,231	9,457	8,654
	Planned Deliveries																
	Delivery Adjustment																
	Ending Inventory	18,957	18,345	17,733	17,121	16,363	15,605	14,846	14,088	13,330	12,555	11,781	11,006	10,231	9,457	8,654	7,851
	DSI	**31.0**	**30.0**	**29.0**	**28.0**	**21.6**	**20.6**	**19.6**	**18.6**	**17.6**	**16.2**	**15.2**	**14.2**	**13.2**	**12.2**	**10.8**	**9.8**
Nashville	Demand Forecast	478	478	478	478	706	706	706	706	706	724	724	724	724	724	741	741
	Starting Inventory	3,832	3,354	2,876	2,398	1,920	1,214	509	-197	-902	-1,608	-2,332	-3,057	-3,781	-4,506	-5,230	-5,971
	Planned Deliveries																
	Delivery Adjustment																
	Ending Inventory	3,354	2,876	2,398	1,920	1,214	509	-197	-902	-1,608	-2,332	-3,057	-3,781	-4,506	-5,230	-5,971	-6,712
	DSI	**7.0**	**6.0**	**5.0**	**4.0**	**1.7**	**0.7**	**-0.3**	**-1.3**	**-2.3**	**-3.2**	**-4.2**	**-5.2**	**-6.2**	**-7.2**	**-8.1**	**-9.1**
Reno	Demand Forecast	25	25	25	25	30	30	30	30	30	30	30	30	30	30	25	25
	Starting Inventory	663	638	614	589	564	534	504	474	444	414	384	354	324	294	264	1,199
	Planned Deliveries															960	
	Delivery Adjustment																
	Ending Inventory	638	614	589	564	534	504	474	444	414	384	354	324	294	264	1,199	1,174
	DSI	**25.8**	**24.8**	**23.8**	**22.8**	**17.8**	**16.8**	**15.8**	**14.8**	**13.8**	**12.8**	**11.8**	**10.8**	**9.8**	**8.8**	**48.0**	**47.0**
Winston Salem	Demand Forecast	320	320	320	320	467	467	467	467	467	486	486	486	486	486	537	537
	Starting Inventory	12,311	11,991	11,670	11,350	11,029	10,562	10,095	9,628	9,160	8,693	8,207	7,721	7,235	6,749	6,263	5,726
	Planned Deliveries																
	Delivery Adjustment																
	Ending Inventory	11,991	11,670	11,350	11,029	10,562	10,095	9,628	9,160	8,693	8,207	7,721	7,235	6,749	6,263	5,726	5,190
	DSI	**37.4**	**36.4**	**35.4**	**34.4**	**22.6**	**21.6**	**20.6**	**19.6**	**18.6**	**16.9**	**15.9**	**14.9**	**13.9**	**12.9**	**10.7**	**9.7**

Figure 9: Disguised and Simplified Copy of the Balance Tool Interface for a 20 inch Monitor on April 17, 2007

Nashville from April 30 to May 10. The (disguised) total re-positioning 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 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

		0 17-Apr Tue	1 18-Apr Wed	2 19-Apr Thu	3 20-Apr Fri	6 23-Apr Mon	7 24-Apr Tue	8 25-Apr Wed	9 26-Apr Thu	10 27-Apr Fri	13 30-Apr Mon	14 1-May Tue	15 2-May Wed	16 3-May Thu	17 4-May Fri	20 7-May Mon	21 8-May Tue
Austin	Demand Forecast	612	612	612	612	758	758	758	758	758	775	775	775	775	775	803	803
	Starting Inventory	19,569	16,457	15,845	15,058	14,446	13,688	12,930	12,171	11,213	10,455	9,680	8,906	8,131	7,356	6,582	5,779
	Planned Deliveries																
	Delivery Adjustment	-2500		-175					-200								
	Ending Inventory	16,457	15,845	15,058	14,446	13,688	12,930	12,171	11,213	10,455	9,680	8,906	8,131	7,356	6,582	5,779	4,976
	DSI	26.9	25.9	24.6	23.6	18.0	17.0	16.0	14.8	13.8	12.5	11.5	10.5	9.5	8.5	7.2	6.2
Nashville	Demand Forecast	478	478	478	478	706	706	706	706	706	724	724	724	724	724	741	741
	Starting Inventory	3,832	3,354	5,376	7,398	7,095	6,389	5,684	4,978	4,273	3,767	3,043	2,318	1,594	869	145	-596
	Planned Deliveries																
	Delivery Adjustment		2500	2500	175					200							
	Ending Inventory	3,354	5,376	7,398	7,095	6,389	5,684	4,978	4,273	3,767	3,043	2,318	1,594	869	145	-596	-1,337
	DSI	7.0	11.2	15.5	14.8	9.1	8.1	7.1	6.1	5.3	4.2	3.2	2.2	1.2	0.2	-0.8	-1.8
Reno	Demand Forecast	25	25	25	25	30	30	30	30	30	30	30	30	30	30	25	25
	Starting Inventory	663	638	614	589	564	534	504	474	444	414	384	354	324	294	264	239
	Planned Deliveries															960	
	Delivery Adjustment															-960	
	Ending Inventory	638	614	589	564	534	504	474	444	414	384	354	324	294	264	239	214
	DSI	25.8	24.8	23.8	22.8	17.8	16.8	15.8	14.8	13.8	12.8	11.8	10.8	9.8	8.8	9.6	8.6
Winston Salem	Demand Forecast	320	320	320	320	467	467	467	467	467	486	486	486	486	486	537	537
	Starting Inventory	12,311	9,491	9,170	8,850	8,529	8,062	7,595	7,128	6,660	6,193	5,707	5,221	4,735	4,249	3,763	3,226
	Planned Deliveries																
	Delivery Adjustment	-2500															
	Ending Inventory	9,491	9,170	8,850	8,529	8,062	7,595	7,128	6,660	6,193	5,707	5,221	4,735	4,249	3,763	3,226	2,690
	DSI	29.6	28.6	27.6	26.6	17.3	16.3	15.3	14.3	13.3	11.7	10.7	9.7	8.7	7.7	6.0	5.0

Figure 10: Routing Decisions Recommended by the Optimization Model for the Example Illustrated by Figure 9

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 Figure 8 and 10 make it implausible that a simple 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 used continuously 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. While 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, while 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. 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 transfering 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 re-positioning 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 at and re-routed from. 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.

References

Axsäter, S., J. Marklund, and E. A. Silver (2002). Heuristic methods for centralized control of one-warehouse, n-retailer inventory systems. *Manufacturing and Service Operations*

Management 4(1), 75–97.

Bowman, E. H. (1963). Consistency and optimality in managerial decision making. *Management Science 9*(2), 310–321.

Caggiano, K. E., J. A. Muckstadt, and J. A. Rappold (2006). Integrated real-time capacity and inventory allocation for reparable service parts in a two-echelon supply system. *Manufacturing and Service Operations Management 8*(3), 292–319.

Chand, S., V. N. Hsu, and S. Sethi (2002). Forecast, solution, and rolling horizons in operations management problems: A classified bibliography. *Manufacturing and Service Operations Management 4*(1), 25–43.

Dhalla, N. (2008). Evaluating shortage costs in a dynamic environment. Master's thesis, Massachusetts Institute of Technology.

Eppen, G. and L. Schrage (1981). Centralized ordering policies in a multi-warehouse system with leadtimes and random demand. In L. Schwarz (Ed.), *Multi Level Production / Inventory Control Systems: Theory and Practice*, pp. 51–69. North Holland, Amsterdam, The Netherlands.

Foreman, J. (2008). Optimized supply routing at Dell under non-stationary demand. Master's thesis, Massachusetts Institute of Technology.

Hopp, W. J., M. L. Spearman, and R. Q. Zhang (1997). Easily implementable inventory control policies. *Operations Research 45*(3), 327–340.

Kapuscinski, R., R. Q. Zhang, P. Carbonneau, R. Moore, and B. Reeves (2004, May-June). Inventory decisions in Dell's supply chain. *Interfaces 34*(3), 191–205.

Oral, M., M. S. Salvador, A. Reisman, and B. V. Dean (1972). On the evaluation of shortage costs for inventory control of finished goods. *Management Science 18*(6), B344–B351.

Reyner, A. (2006). Multi-site inventory balancing in a global extended supply chain. Master's thesis, Massachusetts Institute of Technology.

Rote, G. (1992). The convergence rate of the sandwich algorithm for approximating convex functions. *Computing 48*, 337–361.

Song, J.-S. and P. Zipkin (2003). *Supply Chain Operations: Assemble-To-Order Systems*, Volume XXX of *Handbooks in Operations Research and Management Science*, Chapter 11. North-Holland.

Tsay, A. (1999). The quantity flexibility contract and supplier-customer incentives. *Management Science 45*(10), 1339–1358.

Uno, T. (2008, January). *Semiannual TFT LCD Materials and Components Report*. Austin, TX: DisplaySearch, L.L.C.