Proceedings

Thirty-Third
annual meeting

Transportation Research Forum (TRF)

Halloween In New Orleans

October 31 -
November 2, 1991
New Orleans, LA

Editor:
Louis A. Le Blanc

DISPATCHING EFFECTIVENESS WITH ADVANCED TRAIN CONTROL SYSTEMS -- QUANTIFICATION OF THE RELATIONSHIP

Michael E. Smith
Manager, ARES Traffic Planning Systems
Burlington Northern Railroad
Overland Park, Kansas 66201

Randolph R. Resor, Director, Costing and Pricing
Pradeep K. Patel, Senior Engineer
Zeta-Tech Associates, Inc.
Cherry Hill, New Jersey 08002

Introduction: The Significance of Latency

In any controlled environment, the objective of a controller is to keep a system in the desired state. To accomplish this, the first task of the controlled is to monitor the system's state and compare it to the desired state. If the comparison reveals no difference, nothing needs to be done. If a difference is detected, then the controller must apply a corrective force that attempts to restore the system to the desired state.

If the system performs inadequately in spite of the controller's efforts, then the controller must be improved. There are three, and only three, ways to do this:

1. Increase the size of the corrective forces used
2. Increase the frequency with which corrective forces are applied
3. Increase the accuracy with which corrective forces are applied

Fundamentally, a railroad control system is subject to the same principles. The objective of the system is to keep trains running on time. If trains are not running on time, a corrective force must be applied to bring the trains back on schedule.

In current practice, train operations on North American railroads are controlled by several means. They range from verbal movement authorities issued by voice radio (for branch lines and lightly used through routes), to voice radio authorities ("track warrants" or similar terms) overlaid on an automatic block signal system (ABS), to a full train control installation ("centralized traffic control" or "train control system") in which movement authorities are given by signal indication only and, in theory, no voice radio traffic is required. In any of these systems, improving control by increasing the size of the corrective force applied is very difficult. In order to keep the cost of operation reasonably low, North American railroads generally run trains with low horsepower-to-trailing-ton ratios. Therefore, the size of the available corrective force is inherently small. If a train is beginning to run late, it is usually next to impossible to make up the time through faster running.

Other corrective actions are possible, however. Most North American railroads run on single track with sidings. Decisions on when and where to meet trains can result in large differences in line-haul running times for individual trains. Thus, a train behind schedule can usually be expedited by forcing other trains to wait for it on sidings. However, while a larger corrective force can be applied in this way, the time saved for the one train is usually exceeded by the total delay to other trains on the railroad. Therefore, it seems reasonable to assert that corrective actions available to North American train dispatchers are small.

In order to achieve on-time service, railroads must therefore rely on improving the accuracy and frequency with which corrective actions can be taken. ARES provides this improved accuracy and frequency, by providing real-time information on where trains are and what they are doing, via the digital data link. As a result, when schedules are not being met, that information can be received and processed, and new commands sent out, within two minutes. That time interval, called latency in this report, is at least six minutes with a conventional CTC system, and often much longer.

ARES also improves the accuracy with which corrective forces are applied. This is done through high-speed automatic processing of system state data in the control office. This processing will be handled mostly be an element of ARES called the Tactical Traffic Planner (TTP). The TTP will produce an optimal movement plan for train on a dispatcher's territory given a schedule and the relative value (priority) of each train.

In an analysis of the benefits of the Advanced Railroad Electronics System (ARES), carried out in 1988, the largest benefit of ARES was identified as its potential for increasing management control of rail operations (Smith and Resor, 1990). This increased control could be used to produce improvements in line capacity, service quality, and equipment utilization (Smith and Resor, 1991). However, a large part of this benefit was found to depend upon the use of computer-aided dispatching (CAD). CAD, which is characterized by computer algorithms that maximize efficiency of operations, can be used with existing train control systems, as well as with ARES and other advanced train control systems. Since the ARES equipment is costly, an obvious question arose: could some substantial part of the anticipated benefit be realized by simply pairing a CAD algorithm with existing train control tools? This analysis attempts to answer that question.

The analysis presented here was carried out in 1990 for Burlington Northern Railroad. Burlington Northern's routes contain examples of all these control systems. The most important difference between them, for the purposes of this analysis, is their latency. This is the time lag between occurrence of a train delay or deviation from schedule and transmission of revised movement instructions based on that information. The latency of a control system can be estimated in several ways. In this analysis, latency was assumed to be half the average running time between location reporting ("OS") points on the railroad (on average, that is how long it takes to become aware of a train's position). However, in track warrant territory latency is assumed to be half the average time between issuance of movement authorities by dispatchers. This is about 35 minutes on BN (examination of train sheets indicates that track warrants are issued about every 70 minutes). Latency for lanes with traffic control systems varies according to the length of signal blocks.

To build and test a model in which latency of control systems could be varied, data was required that reflected the results of dispatching the railroad under various combinations of accuracy, size, and frequency of corrective forces. Comparative data for existing operations and for operations under ARES were also needed. For existing operations, actual train movement data on lines with a variety of signal systems was collected. For the ARES operation, simulations were carried out using the Schedule Analyzer (SCAN), an optimization model developed at the University of Pennsylvania (and described in other papers by the authors). The optimizations assumed the headways and latency that would be typical of ARES.

For the dependent variable in the analysis, train travel time was used as a measure of the efficiency of line-haul operations (the lower the travel time, all other things being equal, the more efficient the operation of the railroad). Train travel time on each studied route depended on:

1. The physical limitations of train and route
2. Speed limits
3. Delays unrelated to traffic (mechanical, signal, etc.)
4. Volume, type, and timing of other traffic on the line

For each train, the minimum travel time over the route (absent constraint number four above) was calculated using the Train Performance Simulator, or TPS (Smith and Resor, 1990). This is called the Unobstructed Travel Time (UTT). For each train, there is also an Actual Travel Time (ATT), defined as the actual observed running time for each train in the data collected for this analysis. The dependent variable, called dispatching effectiveness or η, is simply the sum of UTT divided by the sum of ATT.

The independent variables are more difficult to define. What is called latency here is the inverse of the frequency of application of a "restoring force" to return trains to schedule. Latency is the amount of time required for a response to occur once a schedule deviation has occurred and corrective action has been taken. A description of how this variable is measured in each case is provided in the next section.

The second independent variable was originally intended to measure the dispatcher's ability to make the best decision in each case. However, it proved impossible to directly measure dispatcher ability or competence. A surrogate measure was developed; this was essentially a measure of the difficulty of the problems facing a dispatcher. It seemed obvious that, as the volume of traffic increased, the dispatcher's job became more complex. This would likely compromise the dispatcher's ability to make the best decisions. Therefore, a traffic density variable, train minutes per route mile, was used to measure dispatcher ability.

Finally, to measure the amount of "restoring force", the corrective action that could be taken to return trains to schedule, horsepower per ton was initially used. This proved unsatisfactory because horsepower per ton varies relatively little, even across a railroad as large as Burlington Northern. Also, locations and numbers of meet points are more important than train performance in defining the corrective actions available to the dispatcher. Therefore, track miles per route mile (reflecting the number of length of sidings) was used as a proxy for restoring force.

After definition of the variables, the shape of the function relating these variables to effectiveness was chosen. Intuitively, an exponential form seemed appropriate. The value of dispatching effectiveness is likely to be asymptotic with respect to the independent variables. For example, as traffic density increases, transit times will increase exponentially with an infinite value occurring at the line's jam density. (Chen and Harker, 1989) Therefore, a log-log form was chosen for the regression (Chatterjee and Price, 1977).

A cross-sectional logarithmic regression has been carried out to relate these three variables to dispatching effectiveness. Actual train movement data for sixteen lanes has been used in the regression, along with optimized movement data resulting from the application of a traffic planning algorithm to the actual movement data. In the optimized cases, changes in the total running time for all trains are a result of both the application of a traffic planning algorithm and a reduction in the latency time. The coefficients derived in the regression analysis quantify the relative importance of these factors, as well as physical track capacity, in determining overall dispatching effectiveness.

The primary focus of this analysis is the quantification of the importance of latency as a determinant of the effectiveness of train control systems. However, there are other factors involved. One is the physical capacity of each route (as measured by track miles per train mile). A second is the total traffic volume. A third is the competence of the human dispatcher. The variable train-minutes per route mile addresses these two factors, by capturing both volume and speed of traffic.

Additional sidings or double track, or more competent dispatchers, can increase line capacity. In fact, the most effective -- and also the most expensive -- way to increase line capacity is to add track miles. For a fixed volume of traffic, additional track will increase the effectiveness of dispatching. Changes in traffic volume, absent any change in the type of control system, will also increase or decrease dispatching effectiveness. However, the quality and timeliness of the information provided to dispatchers is also of major importance. With inadequate information, even the most skillful dispatcher will be ineffective.

If the latency of a train control system can be reduced, the timeliness and accuracy of information reaching dispatchers can be improved, and dispatching effectiveness can be increased. Improved effectiveness equals increased line capacity. The purpose of this analysis is to quantify the benefits of reduced latency in terms of improved dispatching effectiveness and increased line capacity.

Description of Methodology

To quantify the relationship between latency, restoring force, and dispatching effectiveness, an equation was constructed using the previously discussed analysis:

$$\eta = y^a * v^b / d^c$$

where, η = dispatching effectiveness

y = latency (feedback rate on train location)

v = traffic volume and flow (train-minutes per route mile)

d = capacity (track miles per route mile)

and a, b, and c are the exponents determined from analysis of train movement data. The exponents a, b and c > 0.

It follows from this equation that $\eta \alpha 1/d$, $\eta \alpha y$, and $\eta \alpha v$. The overall effectiveness, as previously defined, of dispatching on any lane is expressed as:

$$\eta = \sum_i t^i_{CASE} / \sum_i t^i_{TPS}$$

where, t = running time for train i,

TPS signifies the minimum feasible running time (as determined by the Train Performance Simulator), and

CASE signifies the actual train performance on the lane.

The Schedule Analyzer (SCAN) model was used in the ARES benefits analysis to determine the improvements over actual dispatching practice that could be achieved by meet/pass planning. SCAN was developed at the University of Pennsylvania as a tool to determine the feasibility of proposed schedules (Harker, 1989). During the ARES benefits analysis, SCAN was extensively modified to allow for simulation of railroad operations under ARES control, both with and without the constraint of existing block signals. In this analysis, SCAN was applied to actual train movement data to produce a "best" dispatching plan for each lane, given train priorities and topology. The difference between the TPS minimum running times and times taken in actual operation indicates the efficiency of dispatching in each case.

When SCAN was applied to actual train movement data, the result was a more efficient dispatching plan (lower total running time for all trains). This was due in part to a reduction in latency and in part to the use of an optimization algorithm. Smaller latency values permit closer spacing of trains. In the SCAN optimization minimum headways between following trains were set at five minutes. This separation was based on a 1.5 minute latency assumed for ARES, and a 3.5 minute separation to permit a following train to stop safely in the event of a derailment or other accident to the leading train. The magnitude of the improvement calculated by SCAN was a function of traffic density and route topology as well as the characteristics of the trains (e.g. horsepower per ton).

A cross-sectional logarithmic regression was used to determine values for the exponents a, b, and c. The regression was carried out over a total of thirty-two observations: sixteen with actual train performance and actual latency (as determined) and sixteen for SCAN optimized performance and an assumed SCAN latency value. Dispatching effectiveness in each case was expressed as the difference between the sum of all train running times and the sum of the minimum feasible times (TPS times) for all trains in each lane. Table 1 shows actual latencies for the sixteen lanes, determined from signal block length, timetables, and control system type.

Another sixteen observations were developed for an assumed ARES implementation. In these, restoring force remained the same, while latency was reduced to 1.5 minutes and minimum train headway was reduced to five minutes in all lanes. The 5 minute minimum headway allows sufficient time and distance for a train to stop if a preceding train derails or otherwise comes to a rapid and unanticipated halt. Assuming an average travel speed of 40 m.p.h and a 1.5 minute latency for ARES, there will be 3.5 minutes available to stop a train, or about 2.5 miles of distance at 40 m.p.h. Emergency stopping distances for trains are generally about one mile from initiation of braking.

TABLE 1. LATENCY BY LANE

Lane	Latency (min)	Lane	Latency (min)
Galesburg - Creston	14.17	Glendive - Dickinson	8.54
Savanna - La Crosse	9.86	McCook - Denver	6.26
Creston - Pacific Jct	5.68	Denver - Sterling	6.28
Havre - Whitefish	6.86	Alliance - Edgemont	32.46
Whitefish - Yardley	4.85	Staples - Dilworth	67.77
Spokane - Pasco	9.00	Dilworth - Jamestown	12.00
Ft. Scott - Tulsa	7.18	Dilworth - Minot	9.22
Madill - Irving	25.00	Superior - Northtown	9.76

For each lane, the SCAN model was run to determine the improvement in total travel time for all trains, taking into account train performance capabilities, route topology, and the need for trains to meet on single track. A cross-sectional log/log regression (since the relationships were assumed to be non-linear) of these values against n for all lanes, optimized and base, was then carried out to produce coefficients for y (latency), v (traffic density) and d (capacity). There was, as expected, substantially more scatter in the base case data than in the optimized data. A plot of the standardized residuals is attached as Figure 1.

Statistical results of the regression were good. Table 2 shows the regression results and statistics. The R^2 of 0.69 indicated that a large part of the observed variance

TABLE 2. REGRESSION RESULTS AND STATISTICS

Regression Statistics		Regression Results			
Statistic	Value	Element	Latency	Train-Min. / Route MIle	Track-Mi. / Route-Mile
Deg. of Freedom	29	Coefficient	-0.09939	-0.06053	0.4328
Std. Error of Est.	0.03849	Std. Error	0.01510	0.01242	0.0950
R Squared	0.6940	T-Statistic	-6.58	-4.87	4.55

was explained by the three variables in the regression. The remaining variance may in part be due to variations in dispatcher competence. All three variables tested positive for significance at the 99% confidence level.

Alternative model formulations were tried. In one case, a dummy variable was used to capture the variance of dispatcher competence in the base case. In another, the difference between the optimum dispatching plan and the minimum TPS times was inserted as a variable and regressed against a manufactured number termed "dispatcher effectiveness". In a third, a dummy variable was used for double track. The model described above produced results superior to all these formulations.

Table 3 shows the actual effectiveness for each lane, using the regression-derived coefficients, against the predicted point on the regression line for each lane. Also shown are the latency, train minutes per route mile, and track miles per route mile used in the regression.

Figure 2 graphs latency against percentage of total benefits, holding other values constant, for an average of the improvement in effectiveness across all lanes. Calculating percentage benefits in each case was not a trivial exercise. First, total benefit had to be defined. This was defined based on the improvement in dispatching effectiveness observed in the SCAN simulations as opposed to the field observations. For example, referring to Table 3, the field dispatching effectiveness for Havre to Whitefish was 0.84, while the SCAN effectiveness was 0.96. This is an improvement of 0.12 out of 0.84, or 14%. The relevant question is, how fast will this 14% disappear if latency starts to grow?

We can use the equation we have generated to answer that question. We can take the values used in each SCAN case and insert a higher latency value to see what happens to dispatching effectiveness. However, we must be careful to take account of the effect of this procedure on the values of traffic volume (v) in train minutes per track mile. A lower effectiveness yields more train minutes, therefore a higher v.

TABLE 3. VARIABLE VALUES AND REGRESSION RESULTS FOR EACH LANE

Lane	Standard Residual	Independent Variable Values			Dispatching Effectiveness	
		Latency	Trn.-min /Trk.-mi.	Trk.-mi. /Rte.-mi.	Observed	Predicted
FIELD OBSERVATIONS:						
Galesburg - Creston	0.95	14.17	29.89	2	0.92	0.84
Savanna - La Crosse	0.36	9.86	16.72	1.48	0.82	0.80
Creston - Pacific Jct	-0.91	5.68	36.09	1.52	0.75	0.81
Havre - Whitefish	0.24	6.86	19.53	1.49	0.84	0.82
Whitefish - Yardley	-0.53	4.85	33.18	1.22	0.72	0.75
Spokane - Pasco	0.30	9.00	27.31	1.19	0.73	0.71
Ft. Scott - Tulsa	-0.42	7.18	23.77	1.17	0.75	0.73
Madill - Irving	-0.62	25.00	42.95	1.12	0.58	0.61
Glendive - Dickinson	1.75	8.54	21.16	1.17	0.84	0.72
McCook - Denver	-1.26	6.26	11.06	1.15	0.68	0.77
Denver - Sterling	0.32	6.28	17.83	1.18	0.77	0.75
Alliance-Edgemont	-2.05	32.45	66.25	1.55	0.55	0.66
Staples - Dilworth	1.70	67.77	59.42	2	0.81	0.69
Dilworth - Jamestown	-0.30	12.00	46.78	1.25	0.66	0.68
Dilworth - Minot	1.34	9.22	9.15	1.28	0.88	0.78
Superior - Northtown	-2.26	9.76	8.16	1.2	0.62	0.76
SCAN-SIMULATED OBSERVATIONS:						
Galesburg - Creston	0.73	1.5	27.62	2	0.99	1.06
Savanna - La Crosse	0.28	1.5	14.56	1.48	0.94	0.97
Creston - Pacific Jct	-0.08	1.5	28.99	1.52	0.93	0.94
Havre - Whitefish	0.01	1.5	16.99	1.49	0.96	0.96
Whitefish - Yardley	0.42	1.5	26.80	1.22	0.89	0.86
Spokane - Pasco	0.25	1.5	22.71	1.19	0.88	0.86
Ft. Scott - Tulsa	0.90	1.5	19.24	1.17	0.93	0.86
Madill - Irving	-0.35	1.5	31.10	1.12	0.79	0.82
Glendive - Dickinson	0.55	1.5	19.71	1.17	0.90	0.86
McCook - Denver	0.27	1.5	8.23	1.15	0.92	0.90
Denver - Sterling	0.32	1.5	14.62	1.18	0.94	0.88
Alliance-Edgemont	-0.02	1.5	39.53	1.55	0.93	0.93
Staples - Dilworth	-0.49	1.5	48.78	2	0.98	1.02
Dilworth - Jamestown	0.91	1.5	33.49	1.25	0.93	0.86
Dilworth - Minot	-0.47	1.5	8.95	1.28	0.90	0.94
Superior - Northtown	-1.79	1.5	6.4	1.2	0.79	0.93

The new v can be used to generate a new effectiveness and the process repeated until stability is reached. We now have a new effectiveness value to use for each postulated latency.

Performance of this iterative process at each latency level yielded a dispatching effectiveness of 75% at a ten-minute latency. This was very close to the average base case (non-ARES) effectiveness of 74%. At a latency of 1.5 minutes, effectiveness was 91%. Taking these two values as extremes, with 74% representing zero benefit from reduced latency (base case) and 91% representing the maximum benefit (since latency cannot be further reduced), intermediate latency values were assigned percentages of total benefits. As can be seen from the graph, half of the benefits of ARES (in terms of dispatching effectiveness) are lost when latency increases only from 1.5 minutes to four minutes. Even a latency of four minutes may only be achievable with ARES-type technology.

Figures 3 through 8 show the relationships between latency and effectiveness for particular lanes. For consistency across all lanes, the longest latency shown is six minutes. Depending on route topology and traffic levels, some benefits can obtain even at longer latencies than this. Take, for example, Figure 8 which shows the Madill to Irving lane. Part of this lane is now "dark" (unsignaled) territory, with a very long latency. Also, traffic volume is relatively light. In this case, much of the benefit of computer-aided dispatching can be realized even at relatively long latencies, since current effectiveness is rather low. By contrast, the Alliance to Edgemont lane is partially double track, is controlled by centralized traffic control, but is very heavily trafficked. Although base case latency is relatively low, the volume of traffic results in a relatively low effectiveness level. Therefore, reduction of latency to 1.5 minutes can produce substantial benefits, and the benefits curve is very steep. In the middle is a lane like Savanna to La Crosse (Figure 3). This lane is relatively heavily trafficked, mostly double track. Here the benefits curve is shallower, since existing capacity is high relative to traffic volume.

Conclusions

The purpose of the analysis presented here was to determine quantitatively whether most or all of the benefits of computer-aided dispatching could be realized with conventional train-control technology and conventional dispatcher interfaces. The largest benefits of ARES were previously identified as having to do with "precision dispatching". If this precision, or a large part of it, could in fact be provided without an investment in ARES hardware, BN would have little reason to make the capital outlays for full ARES.

Findings of the analysis are as follows:

1. Latency, train minutes per route mile, and track miles per route mile are good predictors of dispatching effectiveness. Other variables were tried, and were found not to enhance the statistical validity of the regression.

2. Use of a dispatching optimization algorithm produces a large reduction in train minutes per mile on every lane, when it is used with a technology which can reduce latency to low levels. When latency is long, especially when traffic volumes are heavy, dispatching optimization alone produces only minor benefits.

3. The low latency values typical of ARES produce a substantial increase in dispatching effectiveness, as shown in Figure 2. This increase in effectiveness equates to an increase in line capacity.

4. The large coefficient for track-miles per route mile indicates the importance of the physical features of a railroad line (number of sidings, miles of double track) in determining dispatcher effectiveness and line capacity. Per unit, additions to track

mileage will increase capacity much more than reductions in latency. However, unit costs are far higher for additional track miles than for investments in reducing the latency of control systems.

5. The low latency value used here for ARES is only achievable with ARES or a similar system. It is simply impossible for human dispatchers, with existing technology, to update themselves on train position and take corrective action as frequently as once every 90 seconds. In fact, in track warrant territory the issuance of warrants was found (from examination of dispatchers' sheets) to occur only about once every 70 minutes.

REFERENCES

Chaterjee, S. and Price, B. Regression Analysis by Example, New York, 1977.

Chen, B. and Harker, P., "Two-Moments Estimation of the Delay on Single Track Railway Lines with Scheduled Traffic," Decision Sciences Working Paper, University of Pennsylvania, Philadelphia, 1988.

Harker, P., "The Use of Advanced Train Control Systems in Scheduling and Operating Railroads: Models, Algorithms, and Applications," Transportation Research Record 1263, Washington, D. C., 1989.

Smith, M. E., Resor, R. R., Patel, P. K., and Kondapalli, S., "Benefits of Meet / Pass Planning and Energy Management Subsystems of the Advanced Railroad Electronics System," Journal of the Transportation Research Forum, Vol. XXX, No. 2, 1990.

Smith, M. E., and Resor, R. R., "The Value of High Quality Service -- How Should the ARES-Equipped Railroad Operate," Presented at the first international conference on advanced train control systems (Denver, 1991). Proceedings to be published in Transportation Research Record.

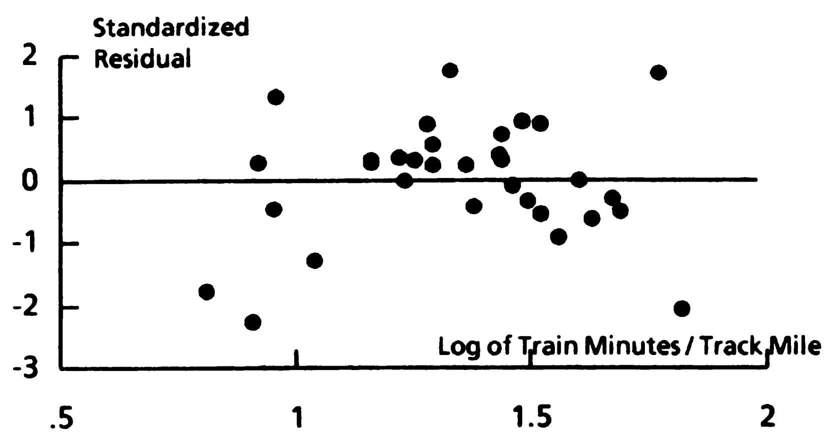

FIGURE 1. STANDARDIZED RESIDUAL vs. LOG. TRAIN MINUTES PER TRACK MILE

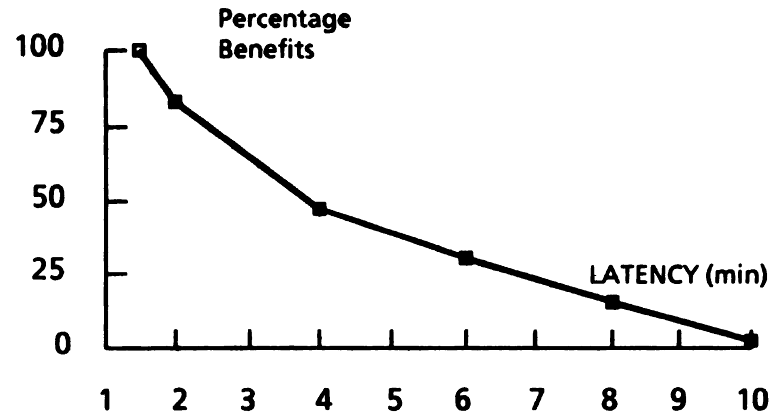

FIGURE 2. LATENCY vs. PERCENTAGE BENEFITS

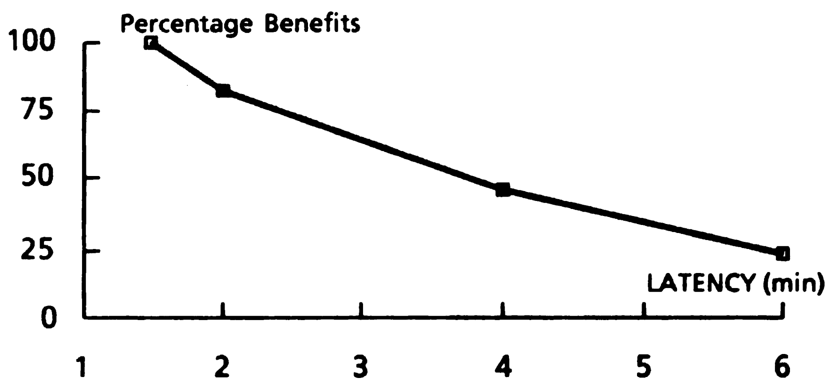

FIGURE 3. LATENCY vs. PERCENTAGE BENEFITS (SAVANA TO LACROSSE)

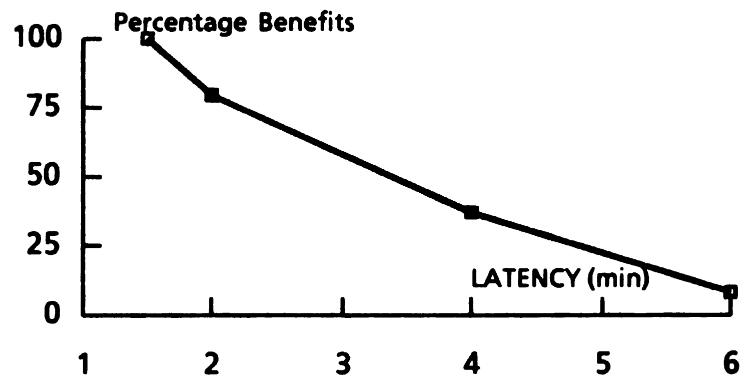

FIGURE 4. LATENCY vs. PERCENTAGE BENEFITS (ALLIANCE TO EDGEMONT)

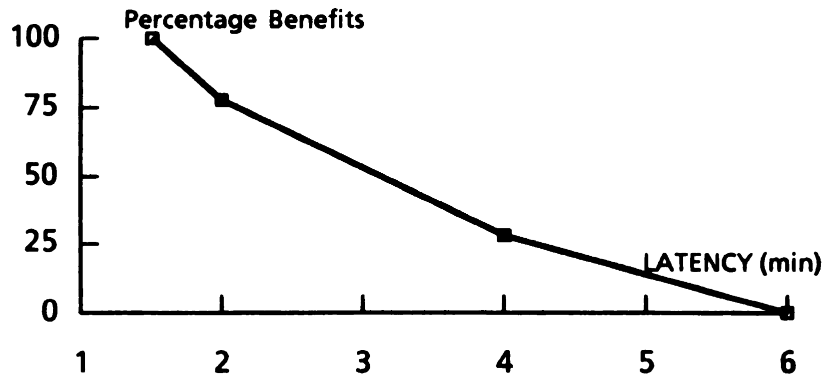

FIGURE 5. LATENCY vs. PERCENTAGE BENEFITS (CRESTON TO PACIFIC JUNCTION)

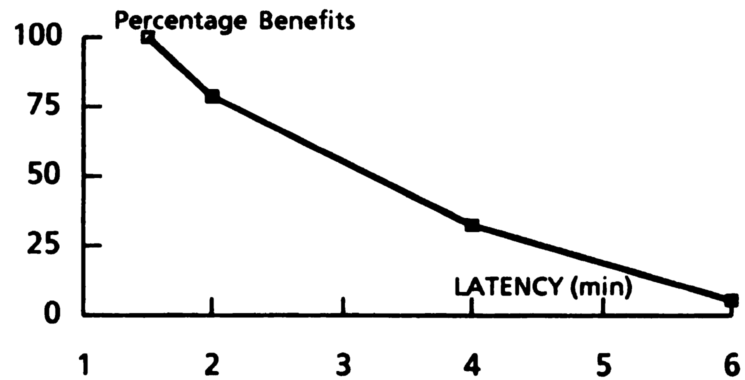

FIGURE 6. LATENCY vs. PERCENTAGE BENEFITS (DENVER TO STERLING)

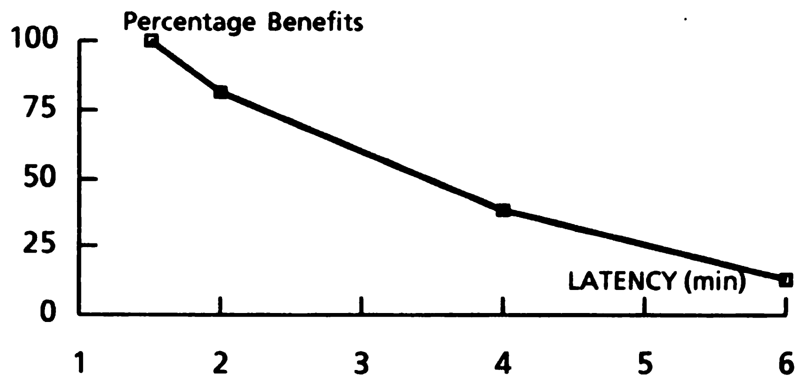

FIGURE 7. LATENCY vs. PERCENTAGE BENEFITS (FT. SCOTT TO TULSA)

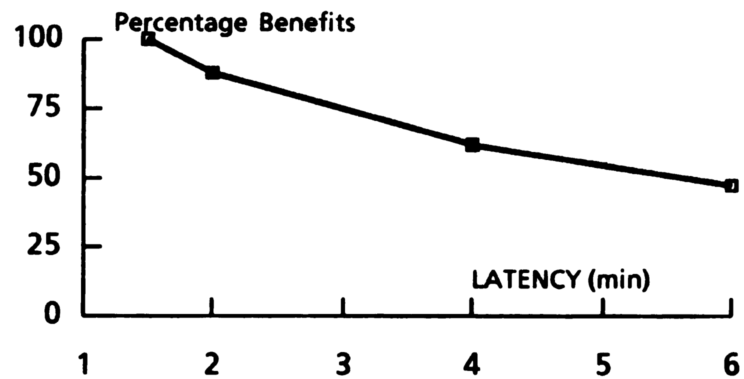

FIGURE 8. LATENCY vs. PERCENTAGE BENEFITS (MADILL TO IRVING)