

INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

**Public Transportation Systems:
Basic Principles of System Design,
Operations Planning and Real-Time
Control**

Carlos F. Daganzo

**Course Notes
UCB-ITS-CN-2010-1**



October 2010

Institute of Transportation Studies
University of California at Berkeley

Public Transportation Systems: Basic Principles of System Design, Operations Planning and Real-Time Control

Carlos F. Daganzo

COURSE NOTES
UCB-ITS-CN-2010-1

October 2010

Preface

This document is based on a set of lecture notes prepared in 2007-2010 for the U.C. Berkeley graduate course “CE259-Public Transportation Systems”--a course targeted to first year graduate students with diverse academic backgrounds.

The document is different from other books on public transportation systems because it is informal, has a narrower focus and looks at things in a different way. Its focus is the planning, management and operation of public transportation systems. Important topics such as financing, governance strategies and urban transportation policy are not covered because they are not specific to transit systems, and because other books and courses already treat them in depth. The document is also different because it deemphasizes facts in favor of ideas. Facts that constantly change and can be found elsewhere, such as transit usage statistics and transit system characteristics, are not covered.

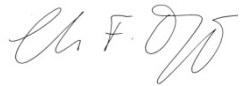
The document’s way of looking at things, and its structure, is similar to the author’s previous book “*Logistics systems analysis*” (Springer, 4th edition, 2005) from which many basic ideas are borrowed. (Transit systems, after all, are logistics systems for the movement of people.) Both documents espouse a two-step planning approach that uses idealized models to explore the largest possible solution space of potential plans. The logical organization is also similar: in both documents systems are examined in order of increased complexity so that generic insights evident in simple systems can be put to use as knowledge “building blocks” for the study of more complex systems.

The document is organized in 8 modules: 5 on planning (general; shuttle systems; corridors; two-dimensional systems; and unconventional transit); 2 on management (vehicles; and employees); and 1 on operations (how to keep buses on schedule). The planning modules examine those aspects of the system that are usually visible to the public, such as routing and scheduling. The management and operations modules analyze the more mundane aspects required for the system to work as designed. Two more modules are in the works: management of special events (e.g., evacuations; Olympics); and operations in traffic.

Although the document includes new ideas, which could be of use to academics and professionals, its main aim is as a teaching aid. Thus, a companion document including 7 homework exercises and 3 mini-laboratory projects directly related to the lectures is also made available. It can be obtained by visiting the Institute of Transportation Studies web site and

looking for a publication entitled: “*Public Transportation Systems: Mini-Projects and Homework Exercises*”. Versions of these exercises and mini-projects were used in the 2009 and 2010 installments of CE259: a 14-week course with two 1-hour lectures and one 1-hr discussion session per week. Sample solutions to the mini-projects and exercises can be obtained by university professors by writing to the ITS publications office and requesting a third document entitled: “*Public Transportation Systems: Solution Sets*”.

The various modules were originally compiled by PhD students Eric Gonzales, Josh Pilachowski and Vikash Gayah, directly from the lectures. Subsequently, my colleague Prof. Mike Cassidy used them in an installment of CE259 and offered many comments. This published version has been edited and reflects the input of all these individuals. Their help is gratefully acknowledged. The errors, of course, are mine. The financial support of the Volvo Research and Educational Foundations is also gratefully acknowledged.

A handwritten signature in cursive script, appearing to read 'C. F. Daganzo'.

Carlos F. Daganzo
September, 2010
Berkeley, California

CONTENTS

Preface	<i>i</i>
Module 1: Planning—General Ideas	1-1
• Course substance and organization	1-1
• Transit Planning	1-2
○ Definitions	1-2
○ How to account for politics	1-3
○ How to account for demand	1-6
○ The shortsightedness tragedy	1-6
○ Planning and design approaches	1-7
• Appendix: Class Syllabus	1-10
Module 2: Planning—Shuttle Systems	2-1
• Overview	2-1
• Shuttle Systems	2-2
○ Individual Transportation	2-2
▪ Time-independent Demand	2-2
▪ Time-Dependent Demand – Evening (Queuing)	2-3
▪ Time-Dependent Demand – Morning (Vickrey)	2-4
○ Collective Transportation	2-7
▪ Time-Independent Demand	2-7
▪ Time-Dependent Demand	2-8
○ Comparison between Individual and Collective Transportation	2-10
• Appendix A: Vickrey’s Model of the Morning Commute	2-12
Module 3: Planning—Corridors	3-1
• Idealized Analysis	3-2
○ Limits to The Door-to-Door Speed of Transit	3-2
○ The Effect of Access Speed: Usefulness of Hierarchies	3-5
• Realistic Analysis (spatio-temporal)	3-8
○ Assumptions and Qualitative Issues	3-8
○ Quantitative formulation	3-11
○ Graphical Interpretation	3-12
○ Dealing with Multiple Standards	3-13
○ No transfers	3-14
○ Transfers and Hierarchies	3-17
○ Insights	3-22
○ Standards-Revisited	3-24
○ Space- and Time-Dependent Services	3-26
▪ Average Rate Analysis	3-26
▪ Service Guarantee Analysis	3-28

Module 4: Planning—Two Dimensional Systems	4-1
• Idealized Case (New 2-D Issues)	4-1
○ Systems without Transfers	4-2
○ The Role of Transfers in 2-D Systems	4-4
• Realistic Case (No Hierarchy)	4-9
○ Logistic Cost Function (LCF) Components	4-9
○ Solution for Generic Insights	4-10
○ Modifications in Practical Applications	4-12
○ General Ideas for Design	4-14
• Realistic Case (Hierarchies--Qualitative Discussion)	4-16
• Time Dependence and Adaptation	4-17
• Capacity Constraints	4-19
• Comparing Collective and Individual Transportation	4-20
 Module 5: Planning—Flexible Transit	 5-1
• Ways of delivering flexibility	5-1
○ Individual Public Transportation	5-1
○ Collective Transportation	5-2
• Taxis	5-2
• Dial-a-Ride (DAR)	5-6
• Public Car-Sharing	5-10
• Appendix: Determination of Expected Distance to a Taxi	5-13
 Module 6: Management—Vehicle Fleets	 6-1
• Introduction	6-2
• Schedule Covering 1 Bus Route	6-3
○ Fleet Size: Graphical Analysis	6-4
○ Fleet Size: Numerical Analysis	6-6
○ Terminus Location	6-7
○ Bus Run Determination	6-8
• Schedule Covering N Bus Routes	6-9
○ Single Terminus Close to a Depot	6-9
○ Dispersed Termini and Deadheading Heuristics	6-10
• Discussion: Effect of Deadheading	6-12
• Appendix: The Vehicle Routing Problem and Meta-Heuristic Solution Methods	6-13

Module 7: Management—Staffing	7-1
• Recap	7-1
• Staffing a Single Run	7-2
○ Effect of Overtime	7-3
○ Effect of Multiple Worker Types	7-4
• Staffing Multiple Runs	7-5
○ Run-Cutting	7-5
○ Covering	7-6
○ Simplified estimation of cost	7-6
• Choosing Worker-Types	7-8
• Dealing with Absenteeism	7-9
• What is Still Left to be Done	7-11
 Module 8: Reliable Transit Operations	 8-1
• Reliability	8-1
• Systems of Systems	8-1
○ Example 1: a stable single agent	8-2
○ Example 2: an unstable single agent	8-4
○ Example 3: two agents	8-5
• Uncontrolled Bus Motion	8-6
• Conventional Schedule Control	8-8
○ Optimizing the Slack	8-9
• Dynamic (Adaptive) Control	8-11
○ Forward looking Method	8-11
○ Two Way Looking Method (Cooperative)	8-14

Module 1: Planning—General Ideas

(Originally compiled by Eric Gonzales and Josh Pilachowski, January 2008)

(Last updated 9-22-2010)

Outline

- General course info (admin)
- Course substance and organization
- Transit Planning
 - Definitions
 - How to account for politics
 - How to account for demand
 - The shortsightedness tragedy
 - Planning and design approaches

Course Substance and Organization

Goal of the Course

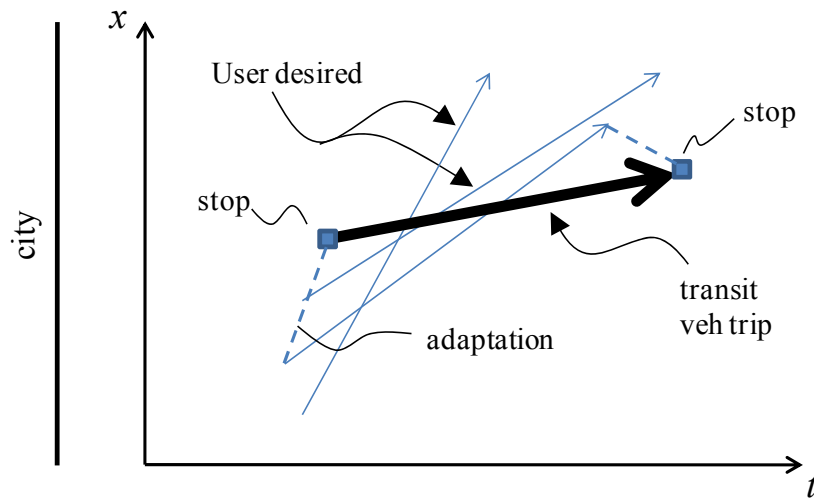
- What transit can and can't do realistically
- How to do it (large/small scale)
- How to make it happen practically (focus on engineering)

Brief Explanation of Syllabus (see Appendix)

- The planning part of the course explores what is possible and how to do it with building blocks of increasing realism and complexity; it shows the limits of transit systems and gives you the tools to develop systematic plans.
- The management and operations part explores the “plumbing” of transit systems. This includes management items that are hidden from the user's view such as fleet sizing/deployment and staffing plans, as well as more visible operational items such as adaptive schedule control and traffic management.
- Planning ideas will be reinforced with two lab projects and five homework exercises. Management/operations ideas will be reinforced with one lab project and two exercises.

Imagine public transit in a linear city. Many people travel between different origins and destinations at different times (thin arrows in the time-space diagram below). Note how people have to adapt their travel in space to the location of stops and in time to the scheduled service in order to use transit (thick arrow), and how this adaptation could be reduced by providing more transit services (more thick arrows). Unfortunately, the thick arrows cost money; and this

competition between supply costs versus demand adaptation turns out always to be at the heart of transit planning. It will be a central theme in this course.



Transit Planning

Definitions

- Guideway – fixed parts of a transportation system, modeled as links and nodes (infrastructure)
- Network – set of links and nodes, uni- or multi-modal
- Path – a sequence of links and nodes
- Origin/Destination – beginning and end of a path through a network
- Terminal – node where users can change modes
- Planning – art of developing long term/large scale schemes for the future
- Mobility – the distance people can reach in a given time (e.g. VKT/VHT)
- Accessibility – the opportunities people can reach in a given time (depends on land use)

We can improve accessibility by improving mobility and/or by changing the distribution of opportunities. But if the opportunities are fixed in space, then a change in mobility is equivalent to a change in accessibility.

As shown in the previous figure, there is a trade-off inherent in public transportation because users give up flexibility (suffering a “level of service” penalty) for economy. To strike this balance between level of service (LOS) and supply cost in networks for individual modes (e.g. highway, bike-lanes, and sidewalks), planners can only change the infrastructure. But in collective transportation, planners also have control over the vehicles’ routes and schedules.

Public Transportation Systems: Planning—General Ideas

The goal of planning is to achieve efficiency, measured as a combination of LOS and supply costs. Costs come in different forms, such as time, T , comfort, safety, and money, $\$$, and should be reduced to some common units. The result is called a generalized cost or disutility, which can be defined both for individuals and groups, and is usually expressed as a linear combination of component costs; e.g. for one individual experiencing time T and cost $\$$ it could be:

$$\text{Generalized Cost} = \beta_T T + \beta_\$ \$$$

How to Take into Account Politics

Note that β_T and $\beta_\$$ will vary between individuals, so even though an individual may have a well-defined generalized cost, the choice of appropriate weights to represent a diverse group is always a political decision that cannot be resolved objectively.

Note too that transit systems involve costs to non-users—energy, pollution, noise, etc.—and that since people also disagree about how these should be valued, they further complicate the decision-making picture.

Clearly, we need to simplify things! (but without ignoring the effects of politics).

To this end, we will assume in this course that there is a political process that has converged to the establishment of some standards, which would apply to all the non-monetary outputs of the transit system; e.g.,

T – Door-to-door time (no more than a standard, T_0)

E – Energy consumed (no more than E_0)

M – Mobility (at least M_0)

A – Accessibility (at least A_0)

And our goal will be minimizing the cost, $\$$, of meeting the standards; i.e.,

$$\text{Mathematical Program (MP): } \min \{ \$: T \leq T_0; E \leq E_0; M \geq M_0; A \geq A_0 \dots \}$$

Note how each standard is associated with an inequality constraining the value of the performance output in question. Since these outputs are usually directly connected to 4 key measures of aggregate motion: VHT, VKT, PHT, PKT, we can often reformulate the standards in terms of passenger time (distance) and vehicle time (distance).

Alternatively, since all variables in this MP (both monetary and non-monetary), which we collectively call $\mathbf{y} = (\$, T, E, M, A)$, are functions of the system design, \mathbf{x} , (i.e., the routes and schedules used for the whole system) and the demand, $\boldsymbol{\alpha}$ (which we assume to be given), we can express the MP in terms of \mathbf{x} and $\boldsymbol{\alpha}$.

Public Transportation Systems: Planning—General Ideas

To make this formulation more concrete, let us define these relations by means of a vector-valued function \mathbf{F}_m :

$$\mathbf{y} = \mathbf{F}_m(\mathbf{x}, \alpha)$$

where,

\mathbf{y} – performance outputs for the entire system (both monetary and non-monetary)

m – mode

\mathbf{x} – design variables for the entire system

α – demand

We then look for the value of \mathbf{x} that minimizes the \$-component of \mathbf{y} while the other components satisfy the standards constraints. The result is as a best design, $\mathbf{x}^*(\alpha)$, which if implemented would yield $\mathbf{y}^*(\alpha) = \mathbf{F}_m(\mathbf{x}^*(\alpha), \alpha) = \mathbf{G}_m(\alpha)$. This function represents the best performance possible from mode m with given demand α . We will, in this course, compare the $\mathbf{G}_m(\alpha)$ for different modes.

To see all this more concretely, consider a simple transit system where all users are concentrated at two points.



In this case we have:

x – frequency of service (a single design variable: buses/hr)

α – demand (a single demand variable: pax/hr)

Define now the components of \mathbf{F}_m . We assume that each vehicle dispatch costs c_f monetary units. Thus we have:

$$\text{\$} = F_m^{\text{\$}}(x, \alpha) = c_f x / \alpha \text{ [\$ / pax]}$$

Note: we have defined $\text{\$}$ as an average cost per passenger. We could instead have defined $\text{\$}$ as the total system cost per hour. Both definitions lead to the same result since they differ by a constant factor: the demand, α . If we now assume that headways are constant but the schedule is not advertised, we have:

$$T = F_m^T(x, \alpha) = 1/x \text{ [hrs]} \text{ (out of vehicle delay assumes } \frac{1}{2} \text{ headway at origin and } \frac{1}{2} \text{ headway at the destination)}$$

And finally, if each vehicle trip consumes c_e joules of energy we also have:

$$E = c_e x / \alpha \text{ [joules / pax]}$$

Public Transportation Systems: Planning—General Ideas

If the political process had ignored energy and simply yielded a standard T_0 for T , and if we choose the monetary units so $c_f = 1$, the MP would then be:

$$\min \{ x/\alpha : 1/x \leq T_0 \}.$$

Note that the OF is minimized by the smallest x possible. Thus, the constraint must be binding, and we have:

$$x^* = 1/T_0$$

Therefore the “optimum” monetary cost per passenger would be:

$$\$^* \equiv G_m^{\$}(\alpha) = 1/(\alpha T_0)$$

We call the above the “standards approach” to finding efficient plans.

There is another approach, which we call the “Lagrangian approach.” It involves choosing some shadow prices, β , and minimizing a generalized cost with these “prices” without any constraints. Although the selection of prices cannot be made objectively, one can always find prices that will meet a set of standards (see your CE 252 notes). So the Lagrangian approach is equivalent to the standards approach. For example, we can formulate:

$$\min_x \{ \$ + \beta T \equiv x/\alpha + \beta(1/x) \}$$

The solution is:

$$x^* = \sqrt{\alpha\beta}$$

You can verify that the “standards” solution ($x^* = 1/T_0$ and $\$^* = x^*/\alpha = 1/(\alpha T_0)$) is achieved for $\beta = (1/T_0^2)(1/\alpha)$. So no matter what standard you choose, there is a price that achieves it.

In summary, there are 2 approaches to obtain low cost designs that satisfy policy aims:

1. Standards: $\min \{ \$ \text{ s.t. } T \leq T_0, E \leq E_0 \dots \}$

This minimizes the dollar cost subject to policy constraints, e.g. for trip time, energy consumption and possibly other outputs. Usually, as shown in the example, constraints become binding when solved $\rightarrow T = T_0, E = E_0$

2. Lagrangian: $\min \{ \$(x, \alpha) + \beta_T(T(x, \alpha)) + \beta_E(E(x, \alpha)) \}$

This minimizes the generalized cost, and gives the same solution as the standards method when suitable shadow prices, β_T and β_E , are chosen. The shadow prices can be found by solving the Lagrangian problem for some prices, finding the optimum T and E and then adjusting the prices until T and E meet the standards. In simple cases, such as the above example, this can be done analytically in closed form.

How to Account for Demand: Some Comments about Demand Uncertainty and Endogeneity

So far, we have assumed that the demand, α , is given, and critics could say that this is not realistic. However, if we are lucky and the design one provides happens to be optimum for the demand that materializes, then the issue is moot. Suppose we design x for a chosen level of demand, α , that is expected to materialize at some point in the future. Normally, we expect realized demand to change with time, and for a well-designed system that provides improved service this demand should be increasing. Then, the question of whether the system design is optimal in reality (given that we assumed a demand α_0) is less a question of if, but of when, since the demand α_0 will eventually be realized. Furthermore, we will learn later in the course that the cost associated with a design, x^* , that is optimal for α_0 is also near-optimal for a broad range of values of α (within a factor of 2 of α_0). Thus, if the realized demand does not change quickly with time, the system design is likely to produce near optimal costs for a long period of time.

Furthermore, we should remember that demand is difficult to predict in the long run. So, building complicated models that endogenize α in order to predict precise values is not a worthwhile activity in my opinion. Rough estimates of future demand are sufficient for design purposes. This is not to say that a vision for the future is not important; only that it does not need to be anticipated precisely. The following example illustrates what happens if one ignores the vision.

The Shortsightedness Tragedy

This example shows that when demand changes with time, then incrementally chasing optimality with short-term gain objectives in mind can lead us to a much worse state than if we design from the start with foresight and long term objectives.

Now, consider the investment decisions for a system with potential for 2 modes:

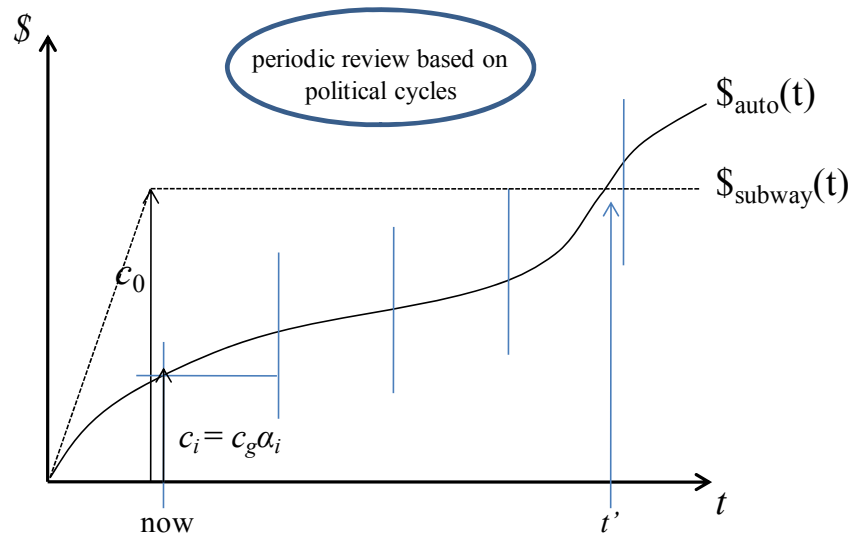
automobile – divisible capacity with cost per unit capacity, c_g

subway – indivisible and very large capacity with cost for a very large capacity, c_0

Politicians, who make decisions about how much money to invest in transportation infrastructure, tend to focus on short-run returns because of the relatively short political cycle. If elections for city leaders occur every couple of years, then politicians have incentives to look at costs only in the near future. This can be “tragic.”

Suppose that demand for transportation in a city is growing over time and is expected to continue growing long into the future (this tends to be the case in nearly all cities around the developing

world). Suppose too that the goal is supplying (at all times) enough capacity to meet demand. The politicians must decide whether to invest a large amount of money, c_0 , in digging tunnels and laying track for a subway that will have enormous capacity to handle demand for decades into the future or to incrementally expand road infrastructure to handle the demand α_i expected over the next political cycle, i . This would cost $c_i = c_g \alpha_i$ monetary units and will be the decision made if $c_i < c_0$ (assuming cost is the main political issue.) The result of this “periodic review” decision making is shown by this figure:



If the decision rule for investing in infrastructure is to choose the lowest cost over the next political cycle and demand increases gradually, “automobile” will always win because with gradual increases in demand: $c_i < c_0$. In the long run, however, the cost of investment in automobile infrastructure is unbounded. Had decisions been made with a view to the long run

($t > t'$), the subway (i.e. the less costly investment) would have been chosen.

Another point pertaining to “the future demand vision” is that systems often create their own demand; and this should be recognized (even exploited) when developing design targets. Planning actions that have long-term consequences should be made with a long-term horizon and long term vision.

Planning and Design Approaches

Comparative Analyses – This is planning by looking at what similar cities have done and trying to copy it. Although this is useful, “safe” and often done, it can exclude opportunities to come up with innovative solutions that may only be appropriate for the case of concern. (We will not do this in this course; we will instead create designs from scratch, systematically.)

Public Transportation Systems: Planning—General Ideas

Step-wise Approach – This is how systematic planning must be done -- problems are too big to be explored in one shot. We first plan generally for the big picture; then fill in the design/engineering step.

In order to conduct broad planning for the large scale, it is useful to simplify the analyses. Decision variables, such as number of buses, number of stops, and number of bus routes are integer values in reality, but we will treat them as divisible (continuous) variables. This greatly simplifies matters, for example turning integer programming problems into linear programs, so that complex problems can be solved much more easily. This will work if the simplification does not introduce large errors.

	Decision	Methods
1. Planning	Large/Long scale	Simplified/Broad
2. Design		Detailed/Specific

Example

Consider a simple mathematical (integer) program, e.g. for maximizing personal mobility subject to a budget constraint:

$$\begin{aligned} \max \{ z = 22x + 18y \} \\ \text{s.t. } 2.1x + 1.9y \leq 2 \\ x, y \in \mathbf{Z} \text{ (integer valued)} \end{aligned}$$

This is so simple that the solution can be obtained graphically (try it); the solution is:

$$x^* = 0, y^* = 1, z^* = 18.$$

Now, if we start with the planning approach and simplify the problem by treating x and y as continuous variables. We are now solving a linear program which has the (optimistic) solution:

$$x^* = 0.952, y^* = 0, z^* = 20.95,$$

(The solution is optimistic because it is the optimum for a problem with fewer constraints.) To obtain a feasible solution the LP solution can be rounded up or down. Because of the constraint, we must round down and we obtain:

$$x^* = 0, y^* = 0, z^* = 0.$$

Public Transportation Systems: Planning—General Ideas

This solution will be pessimistic since it is feasible, but not necessarily optimal. In fact, this is much worse than the optimum solution! So, the simplifying assumptions of the step-wise approach do not work so well for this small scale problem.

Now, if we do the same problem on a much larger scale (e.g. for a budget that would cover a whole city) we would solve instead the mathematical program,

$$\begin{array}{ll}\max & \{ z = 22x + 18y \} \\ \text{s.t.} & 2.1x + 1.9y \leq 200 \\ & x, y \in \mathbf{Z} \text{ (integer valued)}\end{array}$$

Starting with a planning step, assuming the variables can take non-integer values (linear program), the (optimistic) solution is

$$x^* = 95.2, y^* = 0, z^* = 2095.$$

Rounding to the nearest integer value (the design step) gives a pessimistic final objective function value:

$$x^* = 95, y^* = 0, z^* = 2090$$

Now the pessimistic value associated with the integer solution we obtained with the step-wise approach is very close to the optimistic value, and therefore should be even closer to the real optimum that could have been obtained. So, simplifying the problem for large-scale planning purposes, as we will do in this course, is not detrimental to the results of the analysis.

Appendix: Class Syllabus (spring 2010)

The schedule below lists the topics covered in 1-hr lecture periods in the spring semester (2010) and how they were coordinated with the homework exercises and the mini-project activities. Not listed, a 1-hr weekly discussion session was also scheduled to cover the homework exercises and the mini-projects.

Period	Date	Lecture subject	Problems	Mini-project
1	1/19	Introduction: general ideas, politics		
2	1/21	Introduction: standards, demand uncertainty		
3	1/26	Planning: shuttle systems, fixed demand	1 (EOQ)	
4	1/28	Planning: shuttle systems, adaptive demand	1	
5	2/5	Planning: modal comparisons, idealized corridors	2 (Vickrey)	
6	2/4	Planning: idealized corridor hierarchies	2	
7	2/9	Planning: corridors (detailed analysis, standards)		
8	2/11	Planning: corridors (standards vs. generalized costs)		
9	2/16	Planning: inhomogeneous corridors	3 (spacing only CA)	1
10	2/18	Planning: idealized grid systems (issues)	3	1
11	2/23	Planning: realistic grid systems (no hierarchy)		1
12	2/25	Planning: grid systems (practical issues)		1
13	3/2	Planning: hybrid systems (modal comparisons)	4 (modal competition)	2
14	3/4	Planning: hierarchical systems, adaptation	4	2
15	3/9	Planning: paratransit (general concepts; taxis)	5 (hierarchy design)	2
16	3/11	Planning: paratransit (dial-a-ride)	5	2
17	3/16	Planning: paratransit (car-sharing)		2
18	3/18	Management: vehicle fleets (1 route)		2
SPRING BREAK				

Public Transportation Systems: Planning—General Ideas

Period	Date	Lecture subject	Problems	Mini-project
19	3/30	Management: vehicle fleets (n routes)	6 (feeder DAR)	
20	4/1	Management: methodology (meta-heuristics)	6	
21	4/6	Management: staffing (1 run)		3
22	4/8	Management: staffing (n runs)		3
23	4/13	Operations: vehicle movement (theory, systems of systems)		3
24	4/15	Operations: vehicle movement (pairing)		3
25	4/20	Operations: vehicle movement (pairing avoidance)	7 (bus pairing)	
26	4/22	Operations: right-of-way (issues, nodes)	7	
27	4/27	Operations: right-of-way (links, systems)		
28	4/29	Operations: special events (capacity management)		

Module 2: Planning--Shuttle Systems

(Originally compiled by Eric Gonzales and Josh Pilachowski, February, 2008)

(Last updated 9-22-2010)

Outline

- Overview
- Shuttle Systems
 - Individual Transportation
 - Time-independent Demand
 - Time-Dependent – Evening (Queuing), Morning (Vickrey)
 - Collective Transportation
 - Time-Independent
 - Time-Dependent
 - Comparisons and Competition

Overview

Recall from Module 1 that public transportation can be thought of as a system that consolidates individual trips in time and space to exploit economies of scale that result from collective travel. Since this course is about developing insights as well as recipes, we will analyze simple systems starting with point-to-point shuttles, then expand to transit in corridors, and finally build up to the more realistic case of organizing public transportation in 2 dimensions.

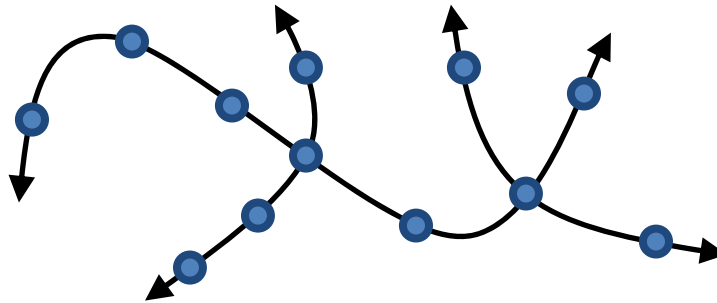
1. Shuttle Systems – Assume the population is already consolidated at two points (an origin and destination) so that there is no spatial consolidation of trips. Collective transportation, in this case, will involve temporal consolidation as individuals adjust their departure times to match the scheduled departure of transit vehicles from the shared origin to the shared destination.



2. Corridors – Assume now that the population is spread along a corridor so that all travel is made in 1 dimension along which transit service is provided. Here, collective transportation must involve spatio-temporal consolidation as individuals must travel to discrete stations where they can board transit vehicles which depart at discrete times.



3. Cities – Finally we consider the more realistic case of a population spread across 2 dimensions. Now transit services must be aligned in a route structure to cover the 2-D space, and this routing adds circuitry to travel as transit systems carry individuals out of the way of their shortest path in order to consolidate trips spatially.



Shuttle Systems

We start by analyzing point-to-point shuttle systems. For comparison purposes we will do this for both, individual and collective transportation modes. In both cases we look first at the time-independent case where we assume steady state conditions (supply and demand are constant over time). This is the way many economic models treat transportation. We then look at the (more interesting) time-dependent case. Individual modes, like private automobiles, incur significant guideway costs in proportion to the capacity provided, which cannot be easily adapted to a time-dependent demand. Public transit modes without extensive guideways will be shown to be more flexible, because a significant part of their costs come from vehicle operations.

Individual Transportation Modes

Time-Independent Demand

In order for individuals to travel in private vehicles (such as automobiles) without much delay, some amount of capacity, μ (pax/hr), must be provided to serve the demand, λ (pax/hr). For private modes, there is a roughly constant infrastructure cost, c_g , per unit of capacity provided. There is also a cost per vehicle trip, c_f , that each driver perceives as a fixed cost of making a trip by private car. Assuming as an approximation that there is no delay whatsoever when the capacity exceeds demand ($\mu \geq \lambda$), the cost per passenger of a private vehicle system is

$$\$ = \frac{c_g \mu}{\lambda} + c_f, \quad \text{for } \mu \geq \lambda.$$

In order to minimize this cost, we would always choose to provide the least possible capacity, which means $\mu = \lambda$. Therefore the minimum cost per passenger is given by

$$\$/ = c_g + c_f$$

which is independent of demand, so there are no economies of scale in our idealization of private transportation; i.e., the *total* cost accrues at rate $\lambda\$/$. Doubling the number of drivers on the road would double the total cost of transportation when just enough capacity is provided to serve demand. We now look at the time-dependent case, both for the evening and morning rush hours, which are different.

Time-Dependent Demand –The Evening Commute with Known Demand (Queuing Analysis)

Until now, we have assumed that demand is time-independent so that as long as capacity matches demand there is no delay, but in reality travel demand rises and falls over the course of a day. Below is a cumulative plot of demand showing the difference between the daily average demand, $\bar{\lambda}$, and the maximum demand in the peak of rush hour, λ_m . We assume that the demand curve is given and (for simplicity only) that the day has a single rush instead of two. Note that $\lambda_m \geq \bar{\lambda}$, and that in a time-independent system where the demand rate does not fluctuate over the course of the day, λ_m would equal $\bar{\lambda}$.

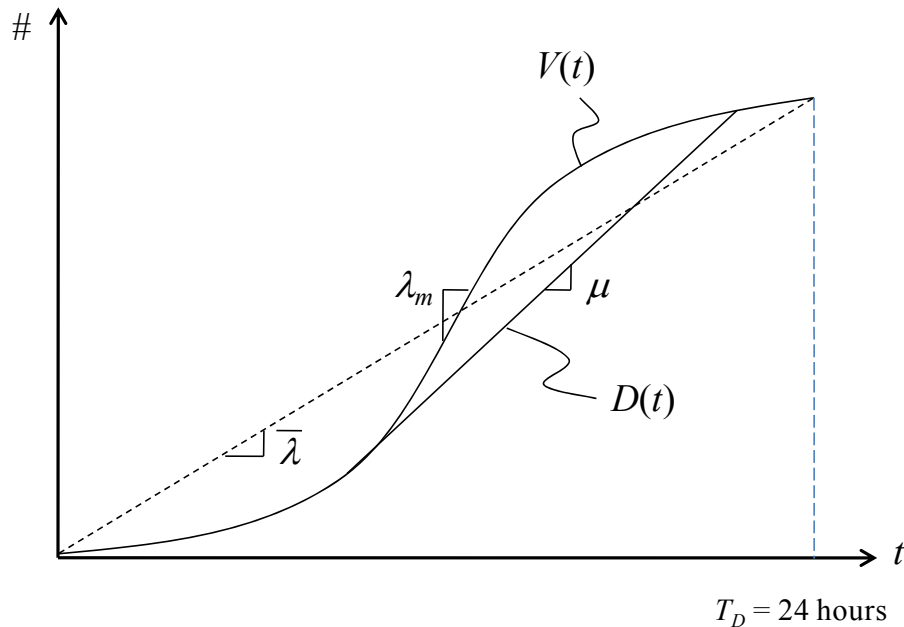


Figure 1.

The minimum monetary cost of providing service subject to a travel delay standard, T_0 , can take a range of values depending on the standard and the capacity it requires. This range can be

easily identified. A lower bound for the cost is obtained by relaxing the standard and simply assuming, $T < \infty$. This relaxed standard is achieved by providing just enough capacity to meet the average daily demand ($\mu = \bar{\lambda}$) such that there are no unserved vehicles carrying over from day to day. This yields a lower bound equal to the monetary cost of the time-independent case: $c_g + c_f$. An upper bound for the cost is obtained by tightening the standard to $T_0 = 0$. This standard is achieved by providing sufficient capacity so that there is never congestion: $\mu = \lambda_m$. The upper bound is therefore as shown below:

$$c_g + c_f \leq \min\{ \$: T \leq T_0 \} \leq c_g \left(\frac{\lambda_m}{\bar{\lambda}} \right) + c_f$$

Note that these bounds apply whether we interpret T as the average delay experienced by drivers, or as the maximum delay experienced in the worst case. The choice of which standard to use is a political decision. But these bounds show that a rush hour can only make costs worse than in the time-dependent case because the cost of serving uniform demand is the lower bound of this expression. So, we still do not see economies of scale.

Aside (showing how to calculate the actual values T^* and $\*): If desired, one can also estimate T^* and $\* (not just the bounds) by using a cumulative plot diagram and/or a spreadsheet. For example, if T and T_0 are averages across drivers, we would evaluate the total time delay, $T_T(\mu)$, for a given capacity, μ , as the area between the arrival curve described by $V(t)$ and the departure curve, $D(t)$, determined by the capacity, μ . The average time delay per driver, $T(\mu)$, is thus given by

$$T(\mu) = \frac{T_T(\mu)}{\bar{\lambda}}.$$

Note from the picture that the area between $V(t)$ and $D(t)$, and therefore $T(\mu)$ declines with μ ; and since the monetary cost of private transportation always increases with capacity, $\$(\mu) \equiv c_g \mu / \bar{\lambda}$, the constraint of our mathematical program must be binding. Thus,

$$T(\mu^*) = T_0$$

which yields μ^* (and $\*).

Time-Dependent Demand –The Morning Commute (Vickrey Model with Endogenous Demand)

In our idealization of the morning commute the times at which people leave their homes and would arrive at our mythical bottleneck are not given. Instead, the demand is driven by work appointments characterized by a cumulative curve of desired departure times through the bottleneck, which we call the wish curve, $W(t)$. If the slope of the wish curve, s , is less than the capacity of the bottleneck, μ , all drivers can pass through the bottleneck exactly when they would

like; then there would be no delay. Curves $V(t)$, $D(t)$ and $W(t)$ would match. However, if the s exceeds capacity, some drivers would have to depart the bottleneck earlier or later than their wished time and the three curves could not match.

To see what could happen as drivers adjust their home departure times (over days) in response to their delays, we suppose that each driver values time in queue at a rate β (\$/hr), time arriving early at rate $e\beta$ and time late at a rate $L\beta$. The constants e and L are dimensionless and such that:

$$e \leq 1 \leq L$$

According to Vickrey (1969), if s exceeds μ and drivers minimize their generalized costs including delay, earliness, and lateness, an equilibrium curve of arrival times to the bottleneck arises in which the order of arrivals to the bottleneck is the same as the order of wished departures.

The equilibrium principle is that no driver should be able to decrease its generalized cost by changing their arrival time. In Vickrey's equilibrium, shown in Fig. 2, there is a critical driver, numbered N_c in the sequence of arrivals and departures, who experiences no earliness or lateness and whose entire cost is time in queue. (Note how the departure curve $D(t)$ crosses $W(t)$ for the ordinate of this driver.) All drivers who arrive before N_c will depart the bottleneck before their wished departure time. We will define N_e as the count of such drivers. All drivers who arrive after N_c will depart the bottleneck after their desired departure time. We will define N_L as the count of such drivers. If there are a total of N_R drivers then the following is true:

$$N_e + N_L = N_R$$

You can convince yourselves that the queuing diagram for the equilibrium is uniquely defined if you are given T , N_e and N_L . It can be shown (see Appendix) that:

$$T = \frac{N_R L e}{\mu(L + e)}; \quad N_e = \frac{L N_R}{L + e}; \quad \text{and} \quad N_L = \frac{e N_R}{L + e}.$$

It also turns out that if $s \gg \mu$, the generalized level of service cost (including both queuing delay and unpunctuality cost) is nearly the same for all commuters, approximately βT . When $L \gg e$, this generalized cost is $\beta N_R / \mu$.

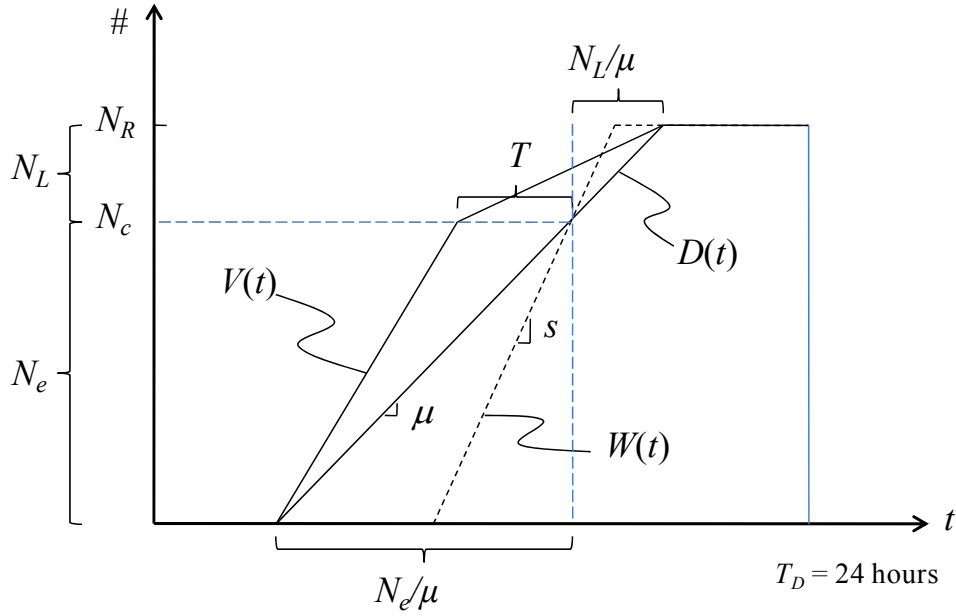


Figure 2.

The total cost of congestion in this morning commute is the sum of total queuing delay (the area between $V(t)$ and $D(t)$), the total earliness penalty (e times the area between $D(t)$ and $W(t)$ where $D(t) > W(t)$), and the total lateness penalty (L times the area between $W(t)$ and $D(t)$ where $D(t) < W(t)$). This calculation can be most easily done based on the geometry of the figure.

A little reflection shows that if we choose a bottleneck capacity that minimizes the out-of-pocket cost per person \$ required to cover the cost of said capacity subject to a time standard (say for the critical commuter), we obtain the same bounds as in the evening rush:¹

$$c_g + c_f \leq \min\{\$: T \leq T_0\} \leq c_g \left(\frac{s}{\bar{\lambda}} \right) + c_f,$$

where $\bar{\lambda} = N_R / T_D$.

So, in the morning rush we continue to be worse-off than in the time-independent case; and economies of scale still do not appear.

¹ This is true because the practical range of μ is $[\bar{\lambda}, s]$ and $\$ = c_f + c_g \mu / \bar{\lambda}$.

Collective Transportation

We now repeat this analysis for public transit and find that the results are quite different (and encouraging).

Time-independent Demand

Consider now a shuttle service provided on an existing guideway from a common origin to a common destination, where the frequency of service is the decision variable that the transit agency can determine.



We assume that shuttle vehicles (e.g., trains) are large enough to carry any number of passengers that may show up and define:

H – headway between vehicle dispatches [hours]

x – frequency of vehicle dispatch [number of vehicles per hour] = $\frac{1}{H}$

c_f – cost per vehicle dispatch of providing shuttle service [dollars per vehicle]

λ – demand [number of passengers per hour]

So, the monetary cost per passenger, \$, of providing shuttle service is given by the cost per hour of dispatching the transit vehicles divided by the total number of passengers using the system.

$$\text{\$} = \frac{c_f x}{\lambda}$$

The out-of-vehicle delay experienced by passengers in the system (ignoring the time in motion between the origin and destination, which is the same for every traveler) is always proportional to the headway of service. For example, if people know the headways but not the schedule and they have specific appointments at the destination (as in the morning commute), they will leave home with at least one headway of slack, which they will spend either at the origin or at the destination. Combined, their total delay would be H . If people do not have specific appointments (as happens for many people in the evening commute) their delay would be $\frac{1}{2}H$ on average. Thus, for the worst-case situation (with appointments) the average delay T is:

$$T = \frac{1}{x}$$

So if we apply a standard T_0 (as we did for individual modes) we have to solve:

$$\min \left\{ \$ \equiv \frac{c_f x}{\lambda} : \frac{1}{x} \leq T_0 \right\}$$

and since the constraint is binding, we find:

$$\$^* = \frac{c_f}{\lambda T_0}$$

Note: There are economies of scale in providing collective transportation because the monetary cost, $\* , decreases with the demand! This is the promise of public transportation vis a vis individual transportation. In reality the contrast is not so pronounced because as we shall see there exist compensating complications, but the promise is real. The reason is that with more demand more individuals can consolidate their travel onto each vehicle without changing the number of vehicle runs; and this lowers the cost of providing transportation per person. We now show that economies still arise if we allow the demand to vary with time.

Time-Dependent Demand

The analysis above assumes that the demand is uniformly spread throughout the course of the day, but in reality the demand for travel is concentrated into rush hours. Let us now evaluate the cost of providing collective transportation for this case, assuming that the passenger arrivals are given.²

Consider now a simplified case of a day with two demand periods: a peak demand, λ_p , for a period of T_p hours of the day, and an off-peak demand, λ_o , for the remaining $T_D - T_p$ hours. The cumulative plot of Fig. 3 shows this demand profile and that N_p passengers travel in the peak, leaving $N_D - N_p$ passengers for the off-peak hours.

² This assumption can now be used for both the evening and morning commutes (with and without appointments) because with our large-vehicles, passengers do not have to compete for limited system capacity.

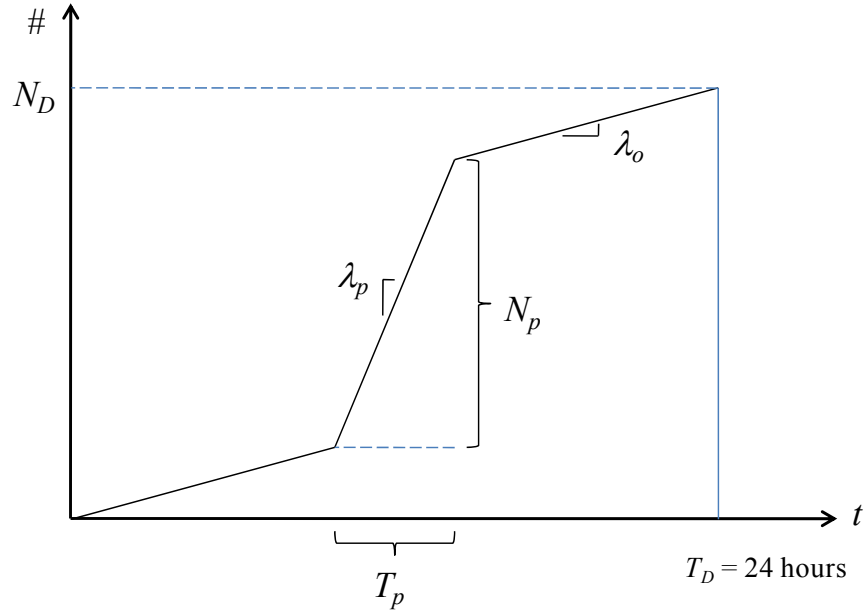


Figure 3.

To design a transit system for this demand, we can break up the day into two regimes and choose a peak period headway, H_p , and an off-peak headway, H_o , to minimize the cost in providing transit service over the course of the whole day. This can be done by minimizing the total generalized cost by the Lagrangian approach with the two decision variables, H_p and H_o :

$$\min\{Z = \beta(\text{Total amount of waiting time}) + c_f(\text{Number of bus dispatches})\}$$

$$\min\left\{Z = \beta\left(H_p N_p + H_o (N_D - N_p)\right) + c_f\left(\frac{T_p}{H_p} + \frac{T_D - T_p}{H_o}\right)\right\}$$

The headways that minimize the generalized cost are

$$H_p^* = \sqrt{\frac{c_f T_p}{\beta N_p}} = \sqrt{\frac{c_f}{\beta \lambda_p}}$$

$$H_o^* = \sqrt{\frac{c_f (T_D - T_p)}{\beta (N_D - N_p)}} = \sqrt{\frac{c_f}{\beta \lambda_o}}.$$

Using these optimal headways gives a minimum total generalized cost of

$$Z^* = 2\sqrt{\beta c_f} \left(\sqrt{T_p N_p} + \sqrt{(T_D - T_p)(N_D - N_p)} \right).$$

Public Transportation Systems: Planning—Shuttle Systems

Note that for a given ratio N_p/N_D this total generalized cost is proportional to $\sqrt{N_D}$, so the generalized cost of collective transportation *per person* is proportional to $1/\sqrt{N_D}$; i.e., it decreases with increasing ridership, N_D , and therefore with the average daily demand $\lambda = N_D/T_D$. So even with time-dependent demand, public transit displays economies of scale.

Technical aside: Note that the optimum cost does not change much if the demand is spread evenly across the whole day. Suppose, for example, that the coefficient $2\sqrt{\beta c_f} = 1$ and 30% of the trips are made in 4 of the 24 hours in a day (i.e., there is quite a bit of peaking). If we use a dummy value $N_D = 10$ in the formula, we find that the total generalized cost for this time-dependent case is

$$1(\sqrt{4 \times 3} + \sqrt{(24 - 4)(10 - 3)}) = 15.30.$$

Using the same logic we see that if the $N_D = 10$ trips had been spread uniformly across the entire 24 hrs, the generalized cost would have been: $(24 \times 10)^{1/2} = 15.49$.

Note the very small difference, and that peaking actually reduces the cost to society, which was not the case for individual modes! You can also convince yourself that the *relative* difference between these two costs is independent of N_D . The relative difference is so small because we can adapt the provision of transit service to match demand. The small and favorable relative error suggests that *to plan collective transportation systems with dominant vehicle costs* (as in our examples) *one can assume a time-independent demand as a simplification*. Infrastructure costs, on the other hand, must be provided in a time-invariant (non-adaptable) way, so the same cannot be said when guideway costs are important, as happens for transportation by individual modes and some collective kinds (e.g., subways).

Comparison between Individual and Collective Transportation Modes

In many cases, individual modes are used in parallel with public transit lines, and an equilibrium is reached in which some trips are made by individual modes and the rest by transit. If a traveler's decision of which mode to take is based only on the level of service (LOS) cost (i.e. the delay time), the equilibrium will be reached when the level of service costs are the same for both choices.

We have seen from Vickrey's model that the generalized cost of delay for automobile commuters is approximately $\beta N_R/\mu$, when $L \gg e$ and $s \gg \mu$. Note that this cost increases proportionally with the number of individuals using the roadway, N_R , and decreases as capacity, μ , is expanded.

For collective transportation, by contrast, the level of service cost is always proportional to the service headway, H , and is independent of the number of individuals using the transit system. It is βH if everyone has appointments. Assuming the vehicles are sufficiently large, this makes

sense because the time cost of riding a transit shuttle depends only on how long a rider must wait for the vehicle, not on how many other people are sharing the vehicle.

So the following diagram plotting general cost vs. number of users helps explain what happens when the two modes provide competing shuttle services for a population of N_R travelers and we have to decide where to allocate funds for increased capacity. The increasing lines correspond to “automobile” and the horizontal lines to “public transit”.

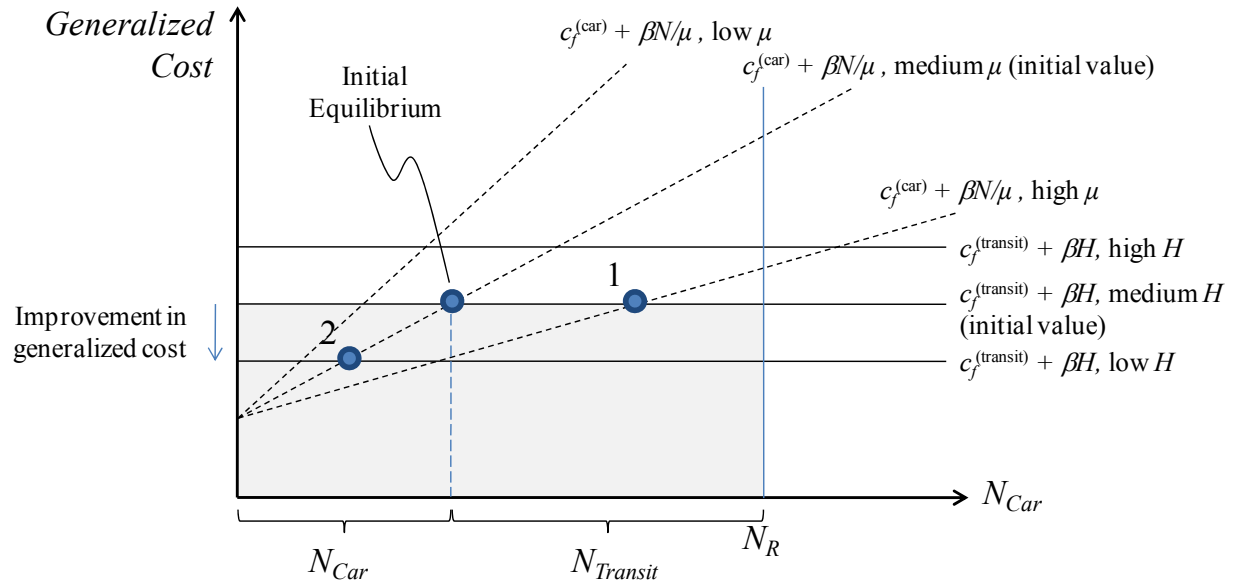


Figure 4.

Assume now that the automobile and public transit systems are initially described by the two curves labeled “medium” in the figure. If people choose shuttle service based on generalized cost, then the intersection of these two curves is the initial equilibrium. The total generalized cost is then the sum of the total cost for all modes (which is the same for all trips, regardless of mode), depicted by the shaded area: $N_R(c_f^{(transit)} + \beta H)$.

Now, suppose some public funds become available and we can choose whether to invest in public transit or individual modes. We can choose to improve the headway for transit service, H , (option 2 in the figure) or the roadway capacity, μ , (option 1); so... where should we spend the money?

An investment in automobile infrastructure lowers the cost of driving which will cause a shift in mode share to more drivers (point 1). The user cost (shaded area), however, remains unchanged because drivers fill the new road capacity until the time delay is equivalent to the time cost of taking transit.

Investing in public transit, however, lowers the user cost for transit riders by reducing the headway, and this creates a mode share shift towards transit (point 2). In this case the improvement benefits both transit riders and drivers (by taking drivers off the road). Therefore, in this idealized example everyone benefits from investing more funds in collective transportation, even those people who never set foot on a transit vehicle.

Related Reading

Vickrey, W.S. (1969). “Congestion theory and transportation investment.” *The American Economic Review*, **59**(2) 251–260.

Appendix A: Vickrey Model of the Morning Commute

We look for an equilibrium where the critical driver is indifferent to any arrival time, and the first and last drivers to the bottleneck experience no delay. Thus, given a fixed slope, μ , of $D(t)$, we can find this equilibrium (see Figure 2) by setting the delay experienced by the critical driver, T , equal to the earliness cost experienced by arriving first or the lateness cost experienced by arriving last:

$$T = \frac{N_e e}{\mu} \quad \text{and} \quad T = \frac{N_L L}{\mu}.$$

With these two equalities and the relation $N_e + N_L = N_R$ we can solve for T , $N_e + N_L$, with the result of the text:

$$T = N_R \left(\frac{\frac{1}{\mu}}{\frac{1}{L} + \frac{1}{e}} \right) = \frac{N_R L e}{\mu(L + e)} \quad ; \quad N_e = \frac{L N_R}{L + e} \quad \text{and} \quad N_L = \frac{e N_R}{L + e}$$

So this shows that the critical driver would not have an incentive to change its arrival position. But for the curves of Figure 2 to be in equilibrium, other drivers—whether their wished times are before or after the critical time—would also have to lack an incentive to change their arrival positions. A good way to verify this is in two steps:

- (a) Draw an “indifference curve” for a generic non-critical driver (with a given wish time) showing for each possible arrival position from 0 to N_R the time at which the driver would have to join the virtual queue when arriving in this position to achieve the generalized cost currently experienced. (Note that each arrival position has a given earliness or lateness for this driver.)

(b) Noting that the latest time at which the queue can be joined for any position is given by $V(t)$; and that $V(t)$ is never to the right of the indifference curve; i.e., the indifference times are not feasible and the driver cannot improve his or her position.

Step (a) requires some care. The following references can perhaps help. They are not required reading, but they contain more detail and additional applications.

Related Reading

Daganzo, C.F. (1985). “The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck.” *Transportation Science*. **19**(1) 29–37.

Daganzo C.F. and Garcia, R.C. (2000). “A Pareto improving strategy for the time-dependent morning commute problem.” *Transportation Science*. **34**(3) 1–9.

Module 3: Planning—Corridors

(Originally compiled by Eric Gonzales and Josh Pilachowski, February, 2008)

(Last updated 9-22-2010)

Outline

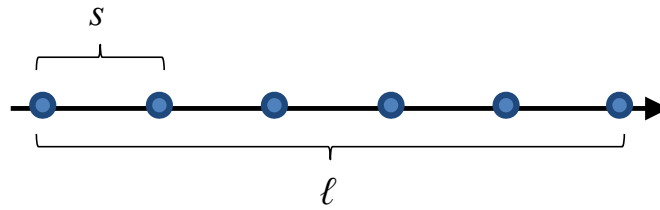
- Idealized Analysis
 - Limits to The Door-to-Door Speed of Transit
 - The Effect of Access Speed: Usefulness of Hierarchies
- Realistic Analysis (spatio-temporal)
 - Assumptions and Qualitative Issues
 - Quantitative formulation
 - Graphical Interpretation
 - Dealing with Multiple Standards
 - No transfers
 - Transfers and Hierarchies
 - Insights
 - Standards-Revisited
 - Space- and Time-Dependent Services
 - Average Rate Analysis
 - Service Guarantee Analysis

In the previous module we looked at the special case where all trips originate at one point and end at another point. Now, we consider demand spread along a corridor, so trips must be consolidated both in time and in space. The design of transit service in a corridor requires choosing a stop spacing, S , and service headway, H .

We will first focus exclusively on S in order to isolate the effect of spatially distributed demand from that of its temporal distribution, which we saw in Module 2. Whereas temporal consolidation involved a trade-off between out-of-vehicle (waiting) time and vehicle operating cost, which had huge economies of scale as demand increased, we will now see that in the spatial case the trade-off is between out-of-vehicle (access) time and in-vehicle time, and that this trade-off is less favorable to public transit: it imposes a severe limit on door-to-door speed even if we make the most favorable assumptions possible for collective transportation.

Idealized Analysis***Limits to Door-to-Door Speed***

Consider a very long transit corridor serving customers that travel from left to right. Customer origins are continuously distributed anywhere along the corridor and their trips can take any length up to a maximum ℓ . The stops are separated by distances, $s \leq \ell$. We are interested in the tightest door-to-door travel time guarantee that can be extended to all customers.



Now we will make a number of optimistic (although unrealistic) assumptions in order to identify this guarantee while accounting for the fact that passengers must access the transit stop and then ride vehicles which make periodic stops to pick up and drop of passengers. This bound will be independent of demand and many other parameters, so it is very general.

- Assume vehicles are dispatched so frequently that once a passenger arrives at a stop, he or she does not wait at all for the next vehicle; i.e., $H = 0$.
- Assume the doors of the vehicle open and close instantly, and passengers take no time to get in or out of the vehicles.
- Finally assume that there is no upper bound to the speed that can be achieved by a transit vehicle while traveling between stops, so that $v_{max} = \infty$.

Although we would agree that these conditions would favor operation extremely, the transit system will still be limited by:

- A maximum acceleration above which passengers will feel physical discomfort from the force ($a_0 \approx 1 \text{ m/s}^2$).
- The average walking speed at which passengers travel to access their nearest transit stop ($v_a \approx 1 \text{ m/s}$).

There are two components of travel time in this case: access time, t_a , and riding time, t_r . In the worst case, the access time results from a passenger walking half of a stop spacing from the origin and another half stop spacing to the destination. So:

$$t_a = \frac{s}{v_a}$$

Riding time is the consequence of the commercial speed of transit (the average speed of the vehicle v_v) which is affected by the stop spacing. If there is no maximum speed, then the transit vehicle will accelerate as it departs a stop until it is half way between stops. Then the vehicle will decelerate to make the next stop (see figure below). Under these conditions, the riding time t_s for a trip between stops can be decomposed into two equal parts of length: $s/2 = \frac{1}{2}a_0(t_s/2)^2$. From this we find:

$$t_s = 2\sqrt{\frac{s}{a_0}},$$

and the riding time t_r for a trip of length $\ell \gg s$ will be approximately ℓ/s times longer; i.e.:

$$t_r \approx \frac{2\ell}{\sqrt{sa_0}}.$$

Note that the commercial speed is therefore:

$$\frac{\ell}{t_r} \approx \frac{\sqrt{sa_0}}{2}.$$

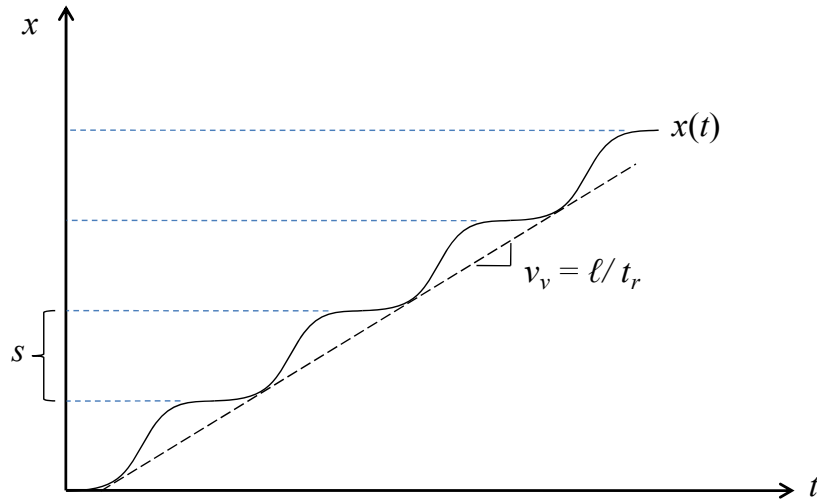


Figure 5.

We assume that people walk to the nearest station. Then, you can verify that for any spacing s you choose, there always is an unlucky passenger who would have to walk a distance s and then

ride for a distance $s\lceil \ell/s \rceil$.¹ As a result, the total door-to-door time for this worst-case passenger is: $t = t_a + t_r = s/v_a + 2s\lceil \ell/s \rceil/(sa_0)^{1/2}$. This function increases with s except and declines only when s is a sub-multiple of ℓ . At these points it takes on the form:

$$t = \frac{s}{v_a} + \frac{2\ell}{\sqrt{sa_0}}.$$

So we look for the minimum of this expression, and as (a very good) approximation we ignore the fact that s should be a sub-multiple of ℓ . There is a trade-off here for choosing the stop spacing s . On the one hand, a longer stop spacing increases the distance passengers must walk to access the mode, so the access time increases with s . However, a greater space between stops allows vehicles to accelerate to higher speeds so that riding time decreases with s . Therefore, an optimal stop spacing, s^* , can be chosen to minimize the door-to-door travel time. The result of this optimization is:

$$s^* = \left(\frac{v_a^2 \ell^2}{a_0} \right)^{1/3}; \quad t^*(a_0, v_a, \ell) = 3 \left(\frac{\ell^2}{v_a a_0} \right)^{1/3}$$

Of course, this result is valid only if $s^* \leq \ell$, as we assumed; i.e., only if $\ell \geq v_a^2/a_0$. Fortunately, since realistic values of v_a^2/a_0 are comparable with 1 m, this requirement is comfortably satisfied for the trip lengths that interest us. Since the unluckiest passenger has a trip length close to ℓ we can approximate the speed of this passenger by:

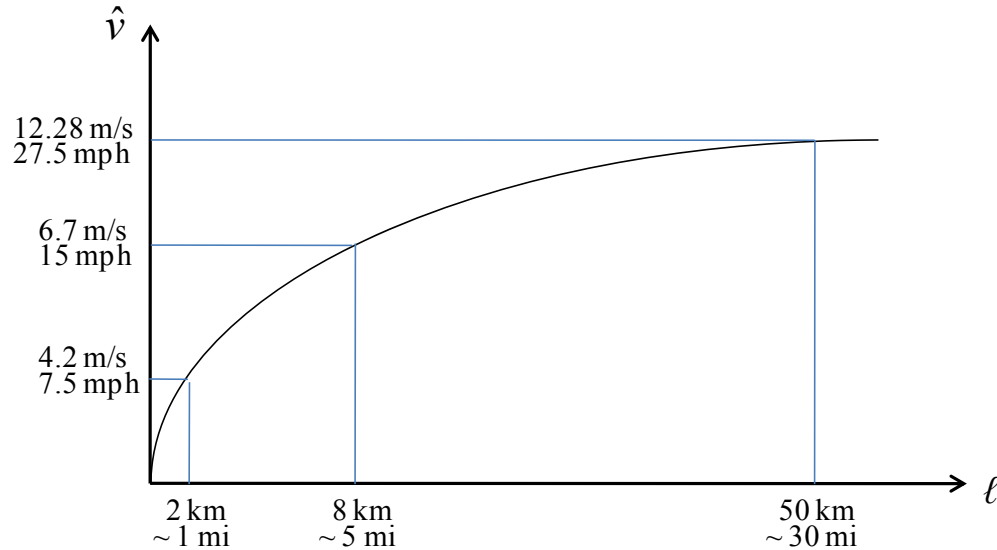
$$\hat{v} \approx \frac{\ell}{t^*} = \frac{1}{3} (\ell v_a a_0)^{1/3},$$

This expression can also be interpreted as the door-to-door speed that can be guaranteed to all passengers with trips of length close to ℓ .

Let us plug in some numbers to see how this upper bound of door-to-door speed changes with the length of trips made. If passengers walk with speed $v_a = 1$ m/s and the maximum allowable

¹ To see this, draw a picture with an unlucky trip as follows: (i) an origin displaced by an infinitesimal amount ε toward the left of a mid-point between stations, and (ii) a trip length, $y = \ell$ if $s = \ell$; or else, $y = s\lceil \ell/s \rceil + 2\varepsilon$ if $s < \ell$. (This is an admissible choice, since for sufficiently small ε the trip length is valid: $y < \ell$.) Now note that in both cases the trip length is a multiple of s , so both the origin and the destination are near a mid-point and access distance is s . Note too that both cases involve severe backtracking with total in-vehicle distance $s\lceil \ell/s \rceil \geq \ell$. You can also convince yourselves that $s\lceil \ell/s \rceil$ is also an upper bound to the in-vehicle distance traveled by any passenger; and that therefore, our unlucky passenger is actually the unluckiest.

acceleration is $a_0 = 1 \text{ m/s}^2$, the figure below shows the fastest door-to-door speeds that can be guaranteed.



This result is very slow, even with all the favorable assumptions we have made for transit (including $v_{max} = \infty$). Why? We are minimizing total travel time including the access time (i.e. maximizing door-to-door travel speed) which relies on passengers walking to the stops. Since people walk very slowly, the stops must be spaced closely enough to limit the time passengers spend accessing transit. This spacing, along with the limit of acceleration, prevents the vehicles from achieving high speeds. With individual transport modes the results are better.² Is there a way of improving collective transportation so it can be more competitive? The answer, as we shall see next day, is yes.

(Hint: the door-to-door speed of public transit depends on the access speed; and if we could increase this speed by some means, the door-to-door speed would increase.) We will explore this issue next, and how to exploit it. We will also study how to plan real corridor systems without the simplifying assumptions we have made – fully recognizing spatiotemporal effects.

The Effect of Access Speed: Usefulness of Hierarchies

For the moment we continue with our idealized and favorable scenario for public transit service. So far, our goal has been to understand how transit door-to-door service speed depends on ℓ . We

² If we made similar favorable assumptions for individual transportation modes on uncongested guideways, their commercial speed would be close to the mode's maximum speed for all ℓ ; i.e., much better than for public transit. The reason is that by being individual these modes do not require much of an access displacement: a great virtue.

Public Transportation Systems: Planning—Corridors

made a couple of assumptions, shown below, in order to obtain an optimistic but very simple upper bound of door-to-door time. The demand, λ , does not matter for this bound.

$$H \cong 0$$

$$t_s = 0$$

$$v_{\max} = \infty$$

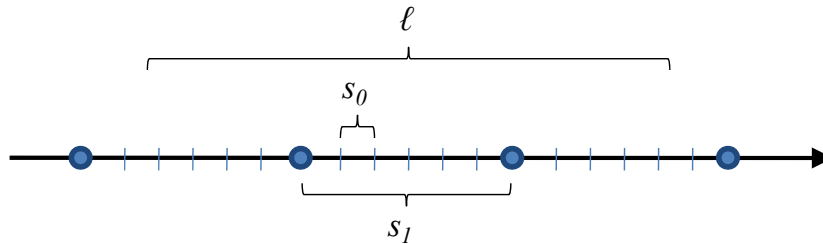
Recall that the door-to-door travel time for the unluckiest passenger was shown to be:

$$t = \frac{s}{v_a} + \frac{2\ell}{\sqrt{sa_0}}.$$

By minimizing this expression with respect to s we obtained the following approximate formulae for the door-to-door travel time and speed of the unluckiest passenger with trip length ℓ :

$$t(\ell) = 3 \left(\frac{\ell^2}{v_a a_0} \right)^{\frac{1}{3}} \quad \text{and} \quad \hat{v} = \frac{1}{3} (\ell v_a a_0)^{\frac{1}{3}}.$$

Note how if we could increase the speed of access the situation would improve. We can do this by using another transit service to provide access!



Let's reexamine our logic assuming this is done. By providing a local transit service with stop spacing, s_0 , to access an express service with stop spacing, s_I , the access speed would now be:

$$v_a = \hat{v} \left(\frac{s_I}{2} \right) = \frac{1}{3} \left(\frac{s_I}{2} v_w a_0 \right)^{\frac{1}{3}}$$

where v_w is the speed of walking. The derivation of this would actually be slightly different so we do not double-count access time, so for simplicity we will assume some small transfer time

equal to $\frac{s_0}{2v_w}$. This will allow us to continue using the same equation. The improved door-to-door travel time is then:

$$t_l(s_1) = \frac{2\ell}{\sqrt{a_0}} s_1^{\frac{1}{2}} + \left[3 \times 2^{\frac{1}{3}} (v_w a_0)^{-\frac{1}{3}} \right] s_1^{\frac{2}{3}}$$

You can verify that:

$$s_0^* < s^* < s_1^*.$$

$t_l(s_1)$ will be the best travel time for a fixed s_l , assuming that you have optimized s_0 already.

Note: you can notice that this equation is in the form:

$$z = Ax^n + Bx^{-m}; \text{ with } n, m > 0$$

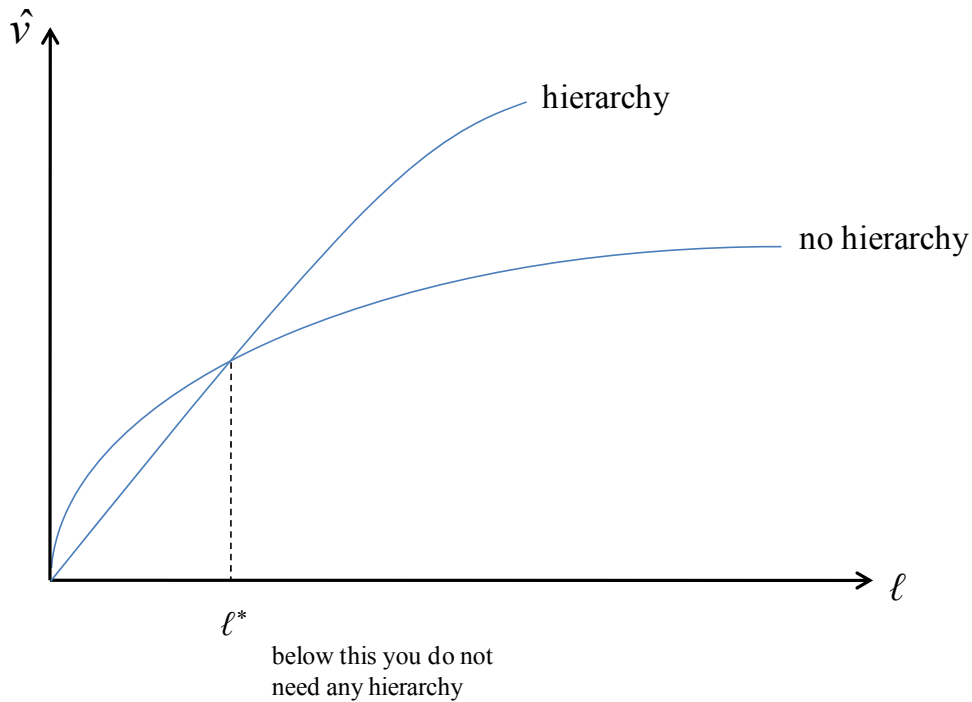
which we will be analyzing in more detail in Homework #2. You will find that the optimum solution $x^* = \left(\frac{Bm}{An} \right)^{\frac{1}{m+n}}$ is insensitive to different values of A and B .

After we optimize t_l with respect to s_l we find the result to be:

$$t_l \approx 5.3 \left(\frac{\ell^4}{a_0^3 v_w} \right)^{\frac{1}{7}} = 5.3 a_0^{-\frac{3}{7}} v_w^{-\frac{1}{7}} \ell^{\frac{4}{7}}$$

This equation shows that t_l is of order $\ell^{\frac{4}{7}}$ and $\hat{v} \propto \frac{\ell}{t_l}$ is of order $\ell^{\frac{3}{7}}$ and of order $v_w^{\frac{1}{7}}$. By

plotting \hat{v} with respect to ℓ with and without a hierarchy we can see for which trip lengths it is optimum to provide a local service.



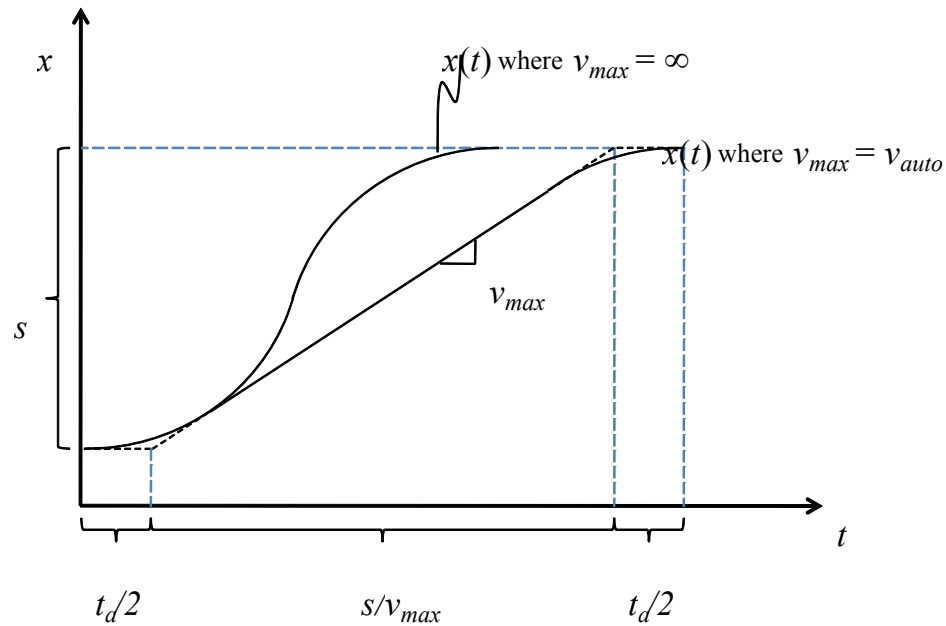
Realistic Analysis with Spatio-Temporal Effects

We have so far made a number of favorable and unrealistic assumptions about our transit system in order to derive generic insights about the effects of the spatial dispersion of passengers along a corridor. So with these insights in mind we now turn our attention to the development of specific plans introducing more realism. The analysis will include both, the spatial and temporal effects of dispersed demand; combining the ideas we have so far seen with those of Module 2. We shall see that in addition to ℓ , two other important variables affect a corridor system's structure: the trip generation rate, λ , and the "user's value of time" β .

Assumptions and Qualitative Issues

Here are the improvements to realism we now consider:

- 1) Remove the assumption that $v_{max} = \infty$; for example define $v_{max} = v_{auto}$ (for buses)



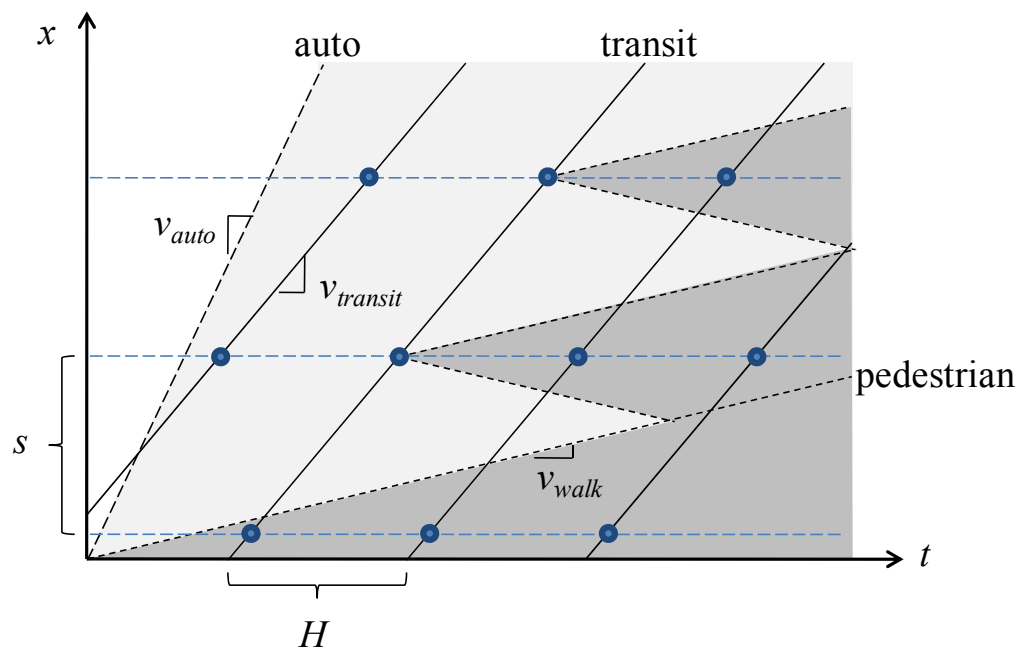
- 2) Remove the assumption that $t_s = 0$. If we approximate the trajectory of the bus with piecewise linear segments of v_{max} and stop time then we can define t_s as the dwell time at a stop plus the loss time due to acceleration and deceleration. The total travel time will then be:

$$t = \frac{dist}{v_{max}} + (\#stops)t_s$$

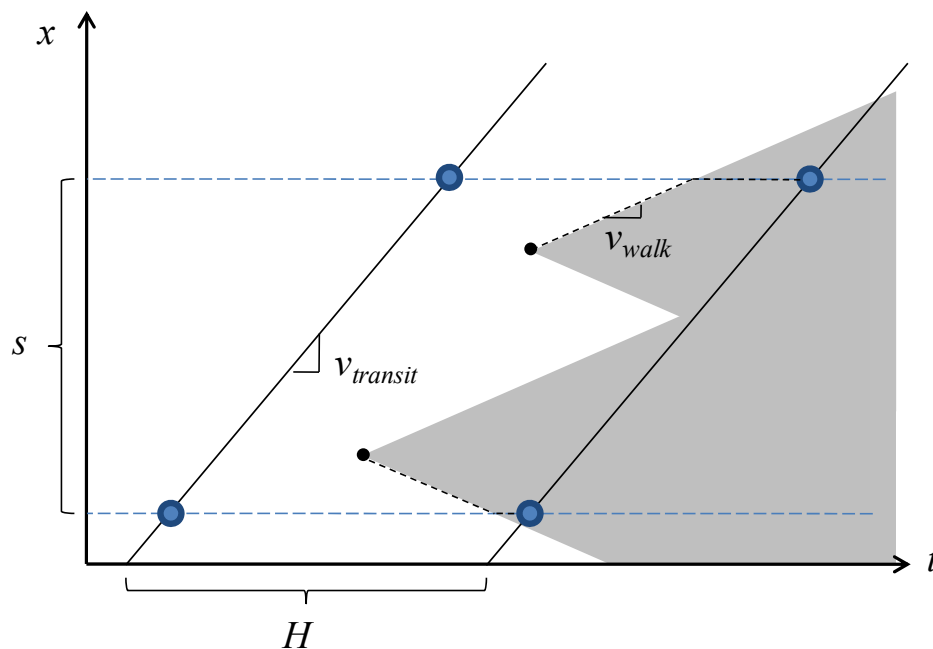
- 3) Remove the assumption that $H = 0$

Before starting quantitative analysis, let us compare the spatio-temporal accessibility provided by different modes with a plot showing the area that a person can reach in a given time depending on their mode of transportation.

Public Transportation Systems: Planning—Corridors



We can look at the area covered by a single stop spacing and headway. Notice how a person, depending on their origin in space and time, will choose a bus stop based on their accessibility:



Quantitative Formulation

Let's try to design a realistic corridor without any hierarchy. We propose choosing the H^* and s^* that minimize the cost of service given some door-to-door travel time standard. For example:

$$\begin{aligned} \min \{ \text{cost of service} \} \\ \text{s.t. } t(\ell) \leq T_0 \end{aligned}$$

We assume for now that we focus on a single “ ℓ ”; e.g. the longest trips people make. To do this, we need formulae for the cost of service and the constraint in terms of our decision variables:

$$\begin{aligned} \text{Cost of service} &= \frac{c_s}{\lambda s H} + \frac{c_d s}{\lambda s H} \\ t(\ell) &= \frac{\ell}{v_{\max}} + t_s \frac{\ell}{s} + \frac{s}{v_a} + H \end{aligned}$$

Note: λ is the average demand density in the corridor (trips/time·dist) and $\lambda s H$ is the number of customers associated with one stop and one vehicle. The constants c_s and c_d are unit costs for a bus stop and a bus-mile. How would you derive these?

To solve the problem we can write the Lagrangian as below. Can you associate the four terms with specific passenger activities?

$$\begin{aligned} &\$_{\text{moving}} + WD + \$_{\text{stop}} + IVD + AD + LH \\ z &= \left(\frac{c_d}{\lambda H} + \beta H \right) + \frac{c_s}{\lambda s H} + \left(t_s \frac{\ell}{s} + \frac{s}{v_a} \right) \beta + \frac{\ell \beta}{v_{\max}} \end{aligned}$$

which (ignoring the “ c_s ” term) has the solution:

$$H^* \cong \left(\frac{c_d}{\lambda \beta} \right)^{\frac{1}{2}}; \quad s^* \cong (v_a t_s \ell)^{\frac{1}{2}}$$

giving us:

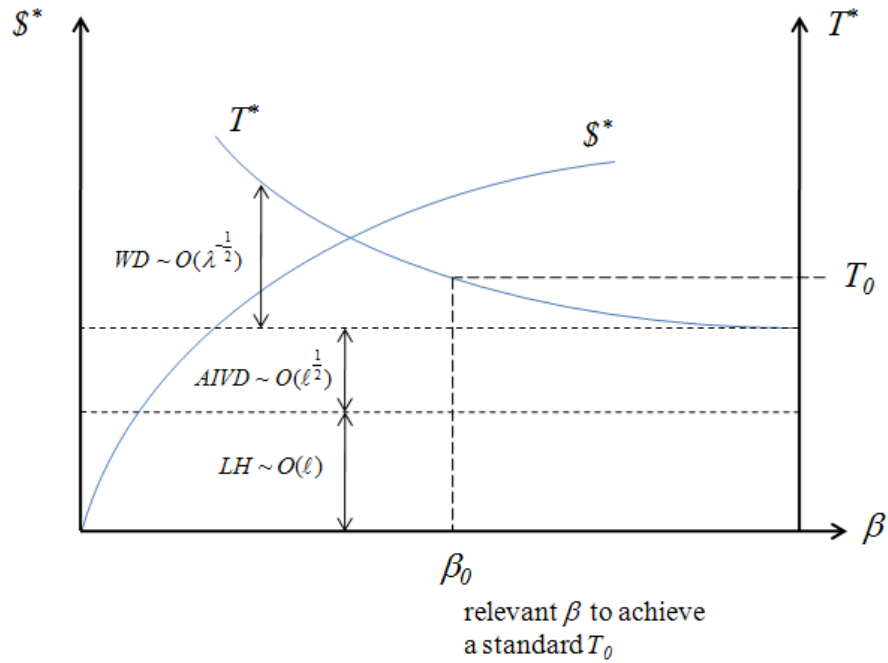
$$\$^* = \left(\frac{\beta c_d}{\lambda} \right)^{\frac{1}{2}} + \left\{ \begin{array}{l} 0 \\ c_s \left(\frac{\beta}{\lambda c_d} \right)^{\frac{1}{2}} (v_a t_s \ell)^{-\frac{1}{2}} \end{array} \right\} \begin{array}{l} \text{lower bound} \\ \text{upper bound} \end{array}$$

$$T^* = \left(\frac{c_d}{\lambda \beta} \right)^{\frac{1}{2}} + 2 \left(\frac{t_s \ell}{v_a} \right)^{\frac{1}{2}} + \frac{\ell}{v_{\max}}$$

Note: the UB solution is obtained by sticking H^* and s^* into the neglected term and adding the result to $\* .

Graphical interpretation:

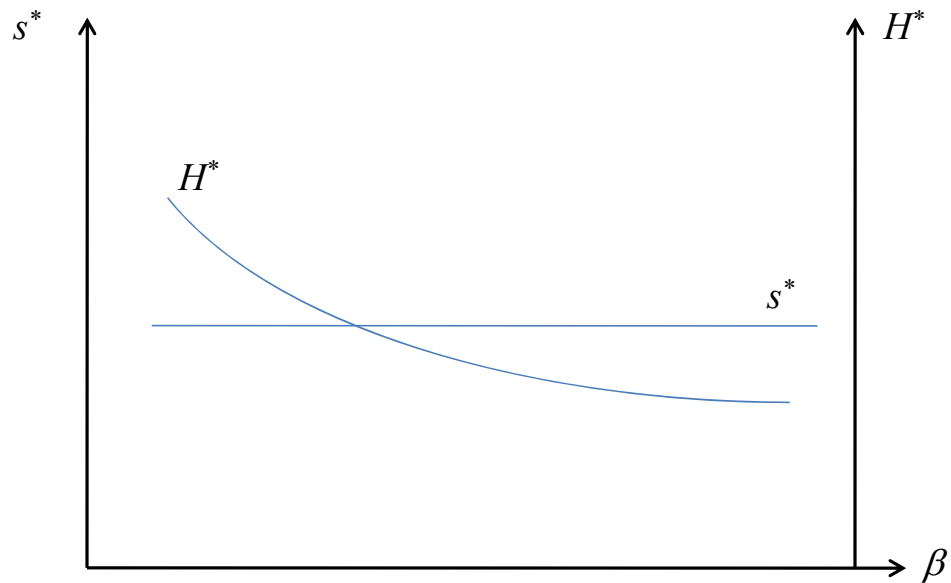
This picture shows how the solution depends on λ , ℓ , and β .



Where WD represents waiting delay, $AIVD$ represents access and in-vehicle-delay, and LH represents line haul time.

Note: “ β ” is a proxy for the wealth of a city and the diagram illustrates the kind of system that cities of wealth might use to satisfy a demand characterized by λ and ℓ .

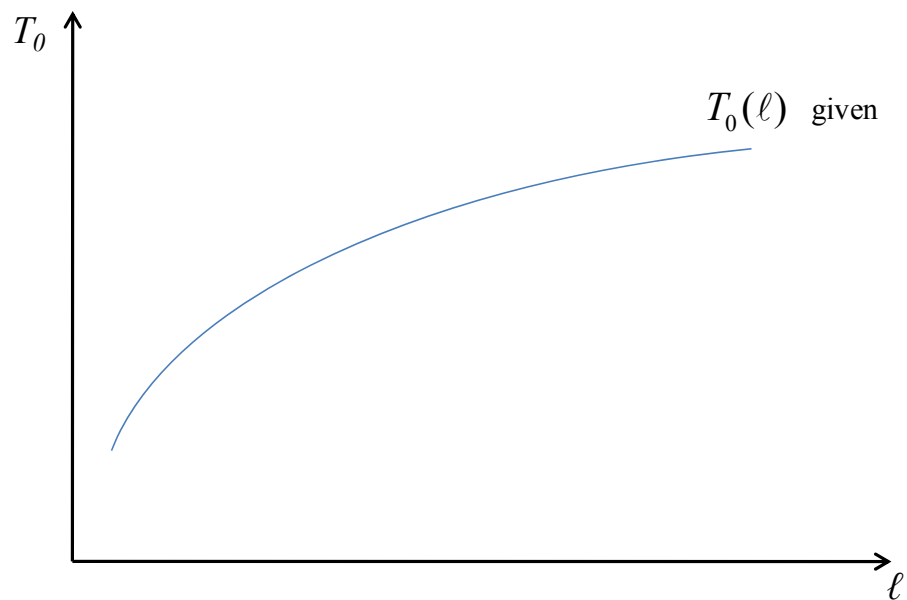
Public Transportation Systems: Planning—Corridors



Dealing with Multiple Standards

A more realistic situation would require adherence to level of service for more than a single trip length. Let's examine the situation where our constraint is:

$$t(\ell) \leq T_0(\ell); \forall \ell$$



We end up with a minimization problem that looks like:

$$\min_{s,H} \{ \text{agency cost}(s, H) \} \quad (1)$$

$$\text{s.t. } T(s, H, \ell) \leq T_0(\ell); \forall \ell \quad (2)$$

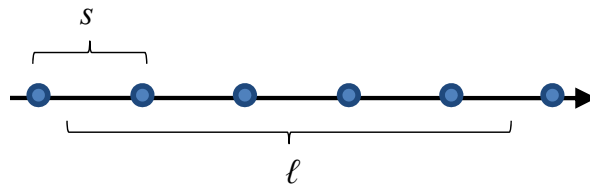
Note: There will always be at least one binding constraint when the problem is minimized. We will call this (unknown) binding trip length ℓ_c . If we knew it and we knew this length provided the only binding constraint (a reasonable assumption), we could formulate the problem as a single-constraint problem and solve it:

$$\begin{aligned} & \min_{s,H} \{ \text{\$}(s, H) \} \\ & \text{s.t. } T_0(\ell_c) = T(s, H, \ell_c) = 0 \end{aligned}$$

This would be an easy task because it can be done with the Lagrangian method we have just seen. Note that the remaining constraints would be satisfied as strict inequalities. If we don't know the critical length, this property of the optimal solution of the single-constraint problem can be used to see if a test value for ℓ is the correct one. So to solve the problem we can solve the single-constraint Lagrangian problem for different ℓ until we find one that exhibits this property.

No Transfers

For our specific corridor formulae and assuming no transfers, this procedure can be simplified even more and the result is intuitive. This is now explained.



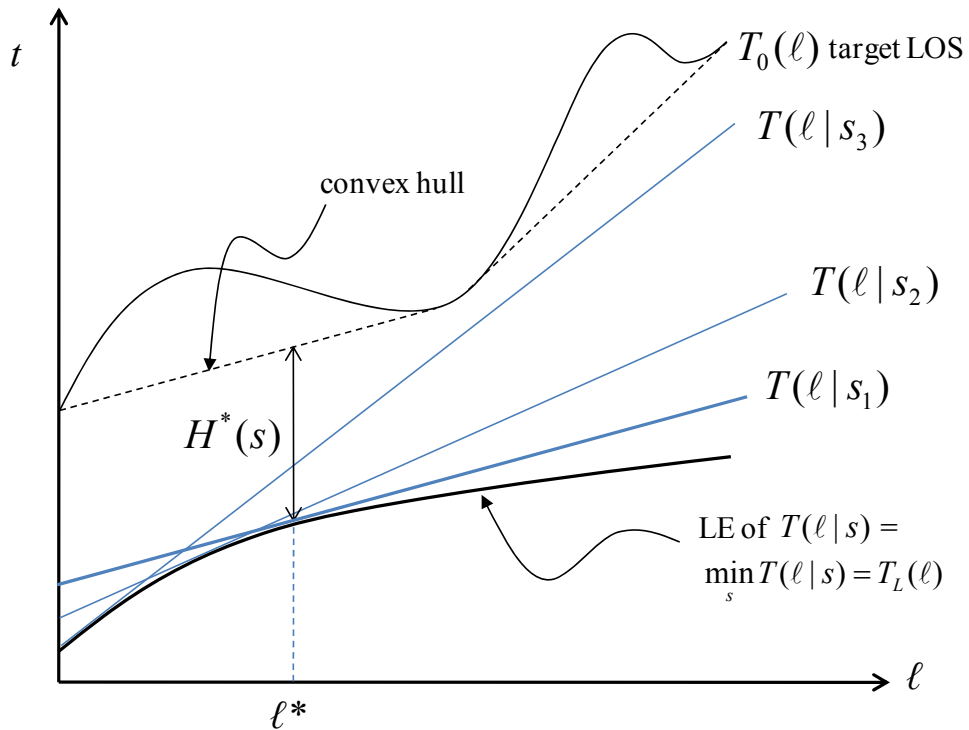
Given our assumptions, the mathematical program corresponding to (1) and (2) is:

$$\begin{aligned} & \min_{s,H} \left\{ \text{\$} = \frac{c_d}{\lambda H} \right\} \\ & \text{s.t. } H + \frac{s}{v_a} + \frac{\ell}{s} t_s + \frac{\ell}{v_{\max}} \leq T_0(\ell), \forall \ell \end{aligned}$$

Notice that the cost function omits the component related to making a stop ($c_s/\lambda sH$) because this value is small, and the objective function here gives a lower bound. Notice too that the constraint separates into a part that depends only on the choice of headway, H ; we call this the waiting delay (WD). The rest of the constraint depends only on the choice of stop spacing, s ; this will be called the access and in-vehicle time (AIVT $\equiv T(\ell|s)$).

When we plot the expression $T(\ell|s)$ for a fixed s the result is a straight line: the vertical intercept is the fixed maximum access time (s/v_a), and the slope the vehicle's average pace ($1/v_{max} + t_s/s$).

The minimum vertical distance between the travel time standard, $T_0(\ell)$, and an AIVT line for a given s , $T(\ell|s)$, represents the fixed amount of waiting delay that can be added to every trip and still keep the travel time with the constraint. Note that this minimum vertical distance is the maximum vertical displacement of our AIVT line until it becomes tangent from below to the $T_0(\ell)$ curve. This vertical displacement is the maximum headway, H , that can be chosen for a given s and still meet the standard, thus minimizing the cost of providing transit service. Now, the AIVT line can be changed by our choice of s , so let's choose the s that gives us the maximum displacement so we can choose the greatest possible H and therefore achieve the lowest possible operating cost. This is the sought result.



This optimization can be done in one shot by considering the lower envelope (LE) of travel time across all choices of s .

$$\text{Lower Envelope of } T(\ell | s) = \min_s \{T(\ell | s)\} = T_L(\ell)$$

To this end, note that when an AIVT line is displaced it cannot possibly touch $T_\theta(\ell)$ in an upward bulge; so we only need to look for points of tangency on the convex hull (CH) of $T_\theta(\ell)$.³ So, we propose the following: slide $T_L(\ell)$ up until it touches (and is tangent to) the convex hull of the time standard $T_\theta(\ell)$.⁴ Then, the displacement is the optimum headway H^* , and the tangent to the envelope at the point of contact ($\ell = \ell^*$) is the optimum AIVT line (with $s = s^*$).⁵

Applying this result,

$$T_L(\ell) = \frac{\ell}{v_{\max}} + 2\sqrt{\frac{\ell t_s}{v_a}}$$

$$s^* = \sqrt{\ell^* t_s v_a}$$

To summarize, we have split the optimization into two parts: (i) a spatial step to find a stop spacing, s , that minimizes the access and in-vehicle time and (ii) a temporal step to find the headway, H , to minimize the cost of meeting the service constraint.

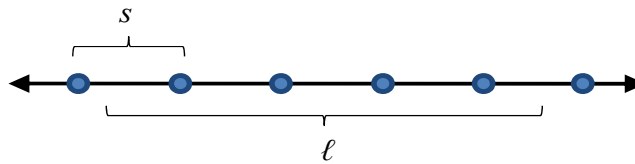
This is approximate and works neatly because we left out the cost of the stopping. So the analysis above gives us a lower bound of cost. If the stopping cost were left in the analysis, the mathematical program can still be solved with brute force in a spreadsheet, but this gives us very little insight. If we solve the simplified formulation and then plug the resulting $T_L(\ell)$ and s^* into the cost function, we will get an upper bound for the cost. No further analysis is necessary when the lower bound and upper bound are close.

What if buses run in both directions along a corridor?

³ The CH is the highest convex curve that can be drawn without exceeding $T_\theta(\ell)$.

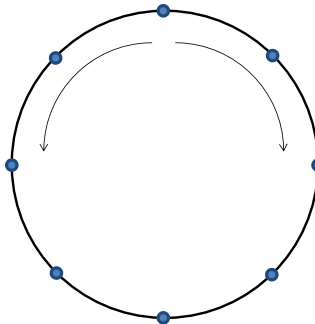
⁴ Note that this point of tangency does not have to be on $T_\theta(\ell)$, as occurs on the figure.

⁵ Why is this true? (i) You see from the geometry of the picture that the displacement of the optimum AIVT line (which is straight) to first contact with $T_\theta(\ell)$, i.e. the optimum headway $H(s^*)$ for the $s = s^*$, is always equal to the displacement of the LE to first contact with the CH; thus, the displacement we propose is the optimum headway for s^* . And (ii) s^* is the optimum spacing because no other AIVT line can be displaced by a greater amount.



The stop spacing will remain unchanged, because s is chosen only to minimize travel time, and the demand plays no role in the travel time expression. The cost of operating service will double, however, because twice as many buses are needed to serve the same demand per unit length.

Exercise: Consider transit service in a loop demand uniformly distributed between all points.

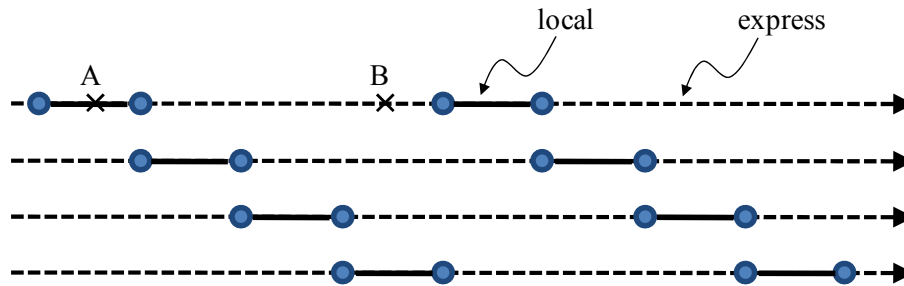


Would we want to serve trips with bi-directional transit routes or is it better to reduce headways by putting all vehicles in service in the same direction? You should be able to convince yourself that if the route has 4 buses or more, it is always better to operate bi-directional service. (Hint: If you had only one bus, it should be obvious that it is most time efficient to operate service in one direction. Likewise, if you had an infinite number of buses, it should be obvious that buses should be deployed in both directions to serve the demand. Where is the tipping point where it becomes more efficient to operate buses in the both directions?)

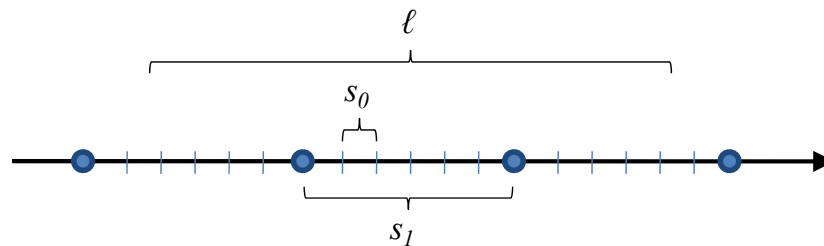
Transfers and Hierarchies

Now, what if we introduce transfers to an express service operating in parallel to the local service with frequent stops. There are couple ways this service could be structured. So far, we have been looking at translationally symmetric route patterns, but this need not be the case. We could run offset local-express services as shown below.

Public Transportation Systems: Planning—Corridors

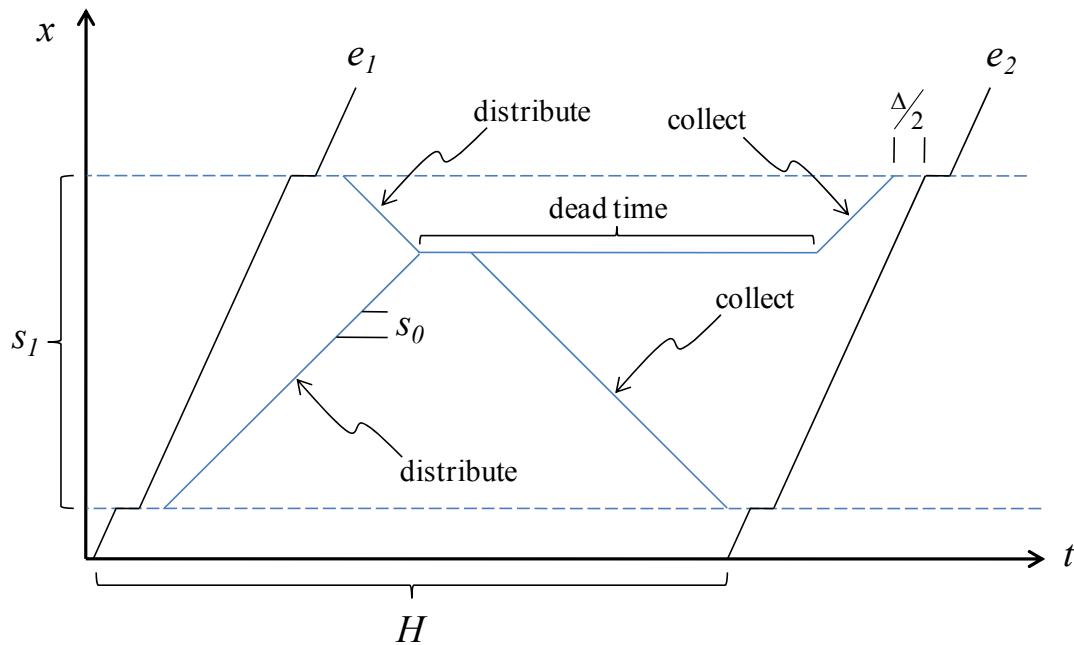


The disadvantages of such a network design outweigh the benefits for cases where the demand is spread out because for trips between points such as A and B we would require multiple transfers. But if all the trips have a common destination (e.g., for feeder systems that collect passengers from many destinations and deliver them to a single hub) the strategy has merit. For spread-out (many-to-many) service it makes sense to consider a local bus service that is paralleled by an express service where passengers can transfer from one service to the other at designated transfer stops.



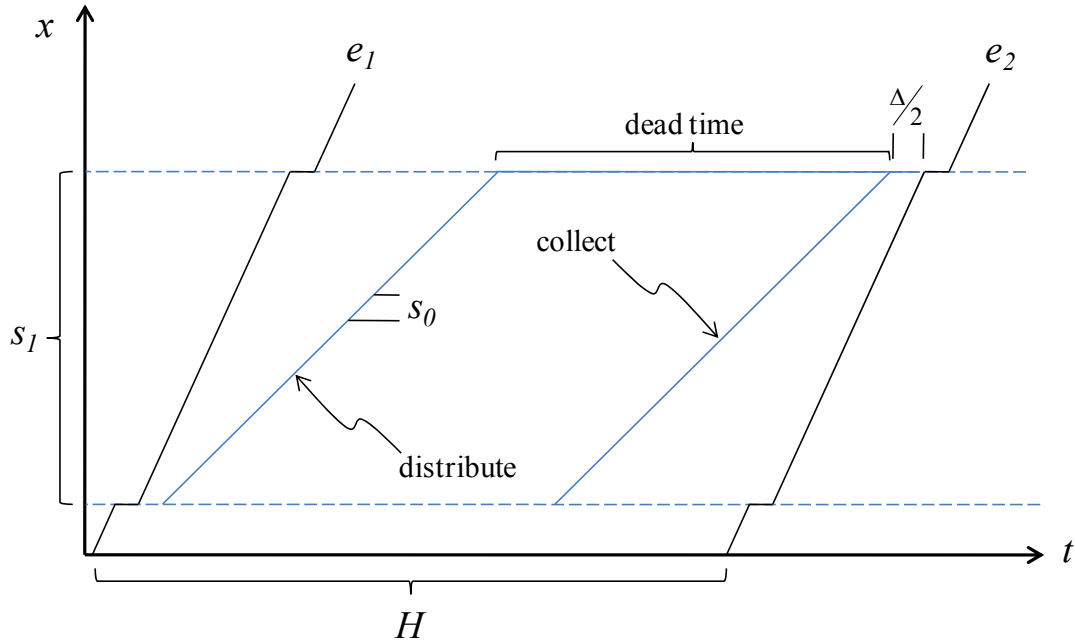
Assume that the headways are synchronized with the same H for local and express services, but the local buses stop with spacing, s_0 , and the express buses make less frequent stops with spacing s_1 . Even this structure of service can be operated in different ways.

Strategy 1: Express buses are scheduled at consistent headways, and the local feeders are dispatched in to depart in both directions along the corridor every time an express bus reaches a transfer station. At some point between transfer stations, the local buses wait and then begin a return trip, bringing passengers to the transfer station just in time for the arrival of the next express bus.



Strategy 2: Express buses are again dispatched at a scheduled headway. Instead of running feeder buses in both directions, a bus is dispatched from the transfer station after the arrival of an express bus, and a second feeder is dispatched in the same direction to collect passengers and drop them off at the downstream transfer station in time to catch the next arriving express bus.⁶

⁶ If service is not synchronized there is no need for “dead-times” and buses can both collect and deliver passengers. The two bus systems can even have different headways, H_0 and H_1 . Could you draw a picture such as those above?



Both of these operational strategies tessellate across time and space and require two local bus dispatches for each express bus dispatch. Therefore they require the same number of vehicle kilometers of service, and a lower bound to the cost of providing service based on vehicle-km is

$$\$_ = \frac{3c_d}{\lambda H}$$

for both timed-transfer strategies. (Convince yourselves that the coefficient would be “2” for unsynchronized service with $H_0 = H_1 = H$). To be complete we must account for bus-hrs while stopping. Then, the cost in a system with timed transfers is

$$\$(s_0, s_1, H) = \frac{1}{\lambda s_1 H} \left\{ 3c_d s_1 + \left(1 + 2 \frac{s_1}{s_0} \right) t_s c_t + 2(\text{dead time})c_t \right\}$$

The unsynchronized case with $H_0 = H_1 = H$ would have a very similar form except for some of the coefficients: “3” would be “2”, the next “2” would be “1” and the final “2” would be “0”. Test yourselves and see if you can derive the unsynchronized expression for $H_0 \neq H_1$.

The door-to-door travel time T is composed of the following components:

H = waiting delay

$$\frac{s_0}{v_w} = \text{access time}$$

v_0 = average speed of local vehicle including stops but not dead time

$$\frac{s_1}{v_0} = \text{local in-vehicle travel time}$$

v_1 = average speed of express vehicle including stops but not dead time ($v_1 > v_0$)

$$\frac{\ell}{v_1} = \text{express in-vehicle travel time}$$

Δ = transfer time

where the vehicle pace $\frac{1}{v_i} = \left(\frac{1}{v_{\max}} + \frac{t_s}{s_i} \right)$

So the door-to-door travel time is given by⁷

$$T(s_0, s_1, H) = H + \Delta + \frac{s_0}{v_w} + s_1 \left(\frac{t_s}{s_0} + \frac{1}{v_{\max}} \right) + \ell \left(\frac{t_s}{s_1} + \frac{1}{v_{\max}} \right); \ell > s_1 > s_0$$

and we can optimize the system with a mathematical program of the familiar form:

$$\begin{aligned} \min_{s_0, s_1, H} &= S(s_0, s_1, H) \\ \text{s.t. } &T(s_0, s_1, H) \leq T_0(\ell), \forall \ell \end{aligned}$$

The lower bound of the cost is now $3c_d/\lambda H$, and the door-to-door time, $T(s_0, s_1, H) = H + T(\ell | \bar{s})$. The maximum possible H can be determined by the same method described for a system with only local service, although here we determine a lower envelope of travel time in 2 parameters, s_0 and s_1 .

$$T_L(\ell) = \min_{\bar{s}} \{T(\ell | \bar{s})\}$$

Example: Considering s_1 for the time being as a constant, find the optimal s_0^* .

$$s_0^* = \sqrt{s_1 t_s v_w}$$

⁷ The only changes for the unsynchronized cases involve the coefficient of H (or of H_0 and H_1 , if $H_0 \neq H_1$).

$$T^*(\ell | s_1) = \Delta + \frac{\ell}{v_{\max}} + 2\sqrt{\frac{s_1 t_s}{v_w}} + \frac{s_1}{v_{\max}} + \frac{\ell t_s}{s_1}$$

The s_1/v_{\max} term is typically much less than $2\sqrt{t_s/v_w}$ so we can ignore s_1/v_{\max} and get an approximate solution.

$$s_1^* \cong (t_s \ell^2 v_w)^{\frac{1}{3}}$$

$$T_L(\ell) = \Delta + \frac{\ell}{v_{\max}} + 3\left(\frac{t_s^2 \ell}{v_w}\right)^{\frac{1}{3}} + \frac{1}{v_{\max}}(t_s \ell^2 v_w)^{\frac{1}{3}}$$

Insights (Comparisons across Countries)

Imagine combining the cost and time into a Lagrangian expression of generalized cost when we value time at a rate of β dollars per unit time. If we neglect the (small) effects of dead times and transfer times, the result is:

$$z = \frac{3c_d}{\lambda H} + \beta H + \beta \left[\frac{s_0}{v_w} + \frac{s_1}{s_0} t_s + \frac{s_1}{v_{\max}} + \ell t_s \frac{1}{s_1} + \frac{\ell}{v_{\max}} \right]$$

Three of the parameters that appear in this expression (λ , β and ℓ) can vary by orders of magnitude across cities and countries, and the others vary much less. Therefore, (λ , β and ℓ) can be thought of as the main drivers of system structure or design. Now, if we divide through the above expression by β so that the generalized cost (GC) is always expressed in units of time, then $\lambda\beta$ always appear together so $z^*(\lambda, \ell, \beta)/\beta$ is really a function of only two drivers of design: ($\lambda\beta$ and ℓ). This generalized cost in units of time is the total time required to make a trip including the time people must spend working to afford system. We can think of the $\lambda\beta$ driver as the “wage generation rate per unit time and distance” because λ is the trip generation rate and β the value of time associated with each trip generated, which should be similar to the wage rate.

It is nice to use intrinsic units that are independent of a currency or country. We can express wages β in any equivalent units we want. For example we could use units of c_t (where c_t is the operating cost per unit time of running a bus), using β/c_t as our wage metric. Note that this ratio is the number of buses that can be continuously operated with the wages of one person. (In rich countries the ratio can be close to 1 and in poor countries much, much less.) Thus, we can think of $\lambda\beta/c_t$ as the “bus generation rate”.

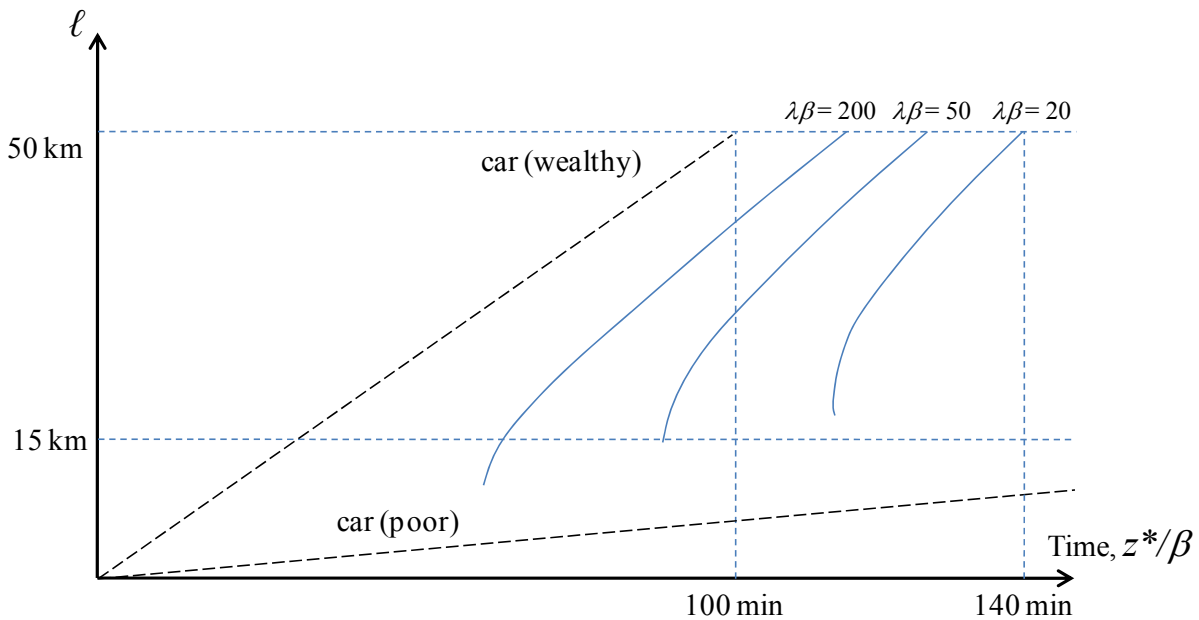
Whether one uses intrinsic units or not, the fact that demand and wealth can be combined into a single driver means that low-density wealthy neighborhoods in developed countries and poor

dense neighborhoods in developing countries (with the same bus-generation rates) should have approximately the same system structure. And they should also share the time-based GC. (This happens because as we have seen the time-based GC depends only on the combined value of $\lambda\beta$.) Isn't it nice that we can say this even before optimizing the system?

Example: Plugging some numbers into this model helps illustrate the difference between transit competitiveness in wealthy versus poor countries. Using extrinsic units of hrs, km, \$:

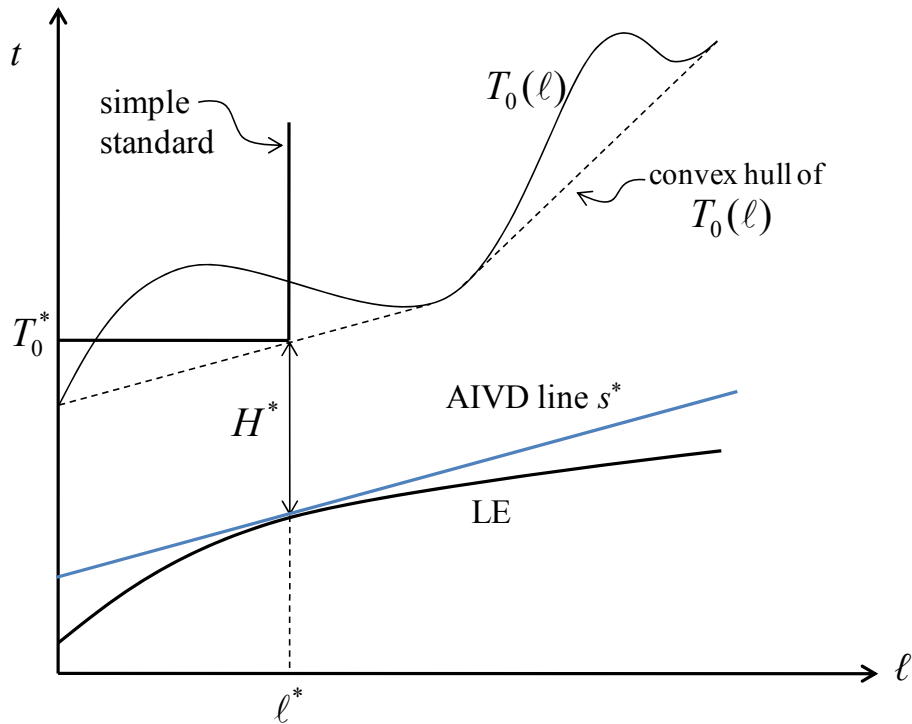
- $v_w \cong 3 \text{ km/hr}$
- $v_{\max} \cong 36 \text{ km/hr}$
- $t_s \cong 5 \times 10^{-3} \text{ hr}$
- $\beta \sim 1 \rightarrow 20 \text{ \$/hr}$
- $c_d \cong 1 \text{ \$/km}$
- $c_s \cong 10^{-1} \text{ \$/stop}$
- $c_t \cong 20 \text{ \$/hr}$
- $\ell \sim 2 \rightarrow 40 \text{ km}$
- $\lambda \sim 1, 2.5, 10, 20, 50, 200 \text{ trips/km}^2$

The values with the greatest range of values (marked with •) are our drivers of design. The figure below shows how the generalized cost (in units of time) relates to the length of a trip for transit serving neighborhoods of different values of $\lambda\beta$ and the cost of making the trip by car in a wealthy or poor country. More accessibility is associated with greater trip length for a generalized cost.



Standards–Revisited (Two Additional Points)

The first point is that every length-based standard can be reduced to a “simple standard”. Recall from the earlier discussion how, for a defined “political” standard $T_0(\ell)$ for door-to-door trip time, we were able to find the critical length of trip and critical headway to satisfy that standard with the graphical construction below.



Note that if we replace $T_0(\ell)$ with the simple standard shown with its corner at point (ℓ^*, T_0^*) we arrive at the same solution! This simple standard can be interpreted such that all trips shorter than a certain length (ℓ^*) must be completed within a certain time (T_0^*) and longer trips can be ignored.

The simplification is useful because it involves just two parameters (ℓ^* and T_0^*). Therefore, by exploring the structure of optimum transit systems for all possible values of these two parameters one would have explored all possible optimum solutions. Note too from the figure that ℓ^* must be the binding length and therefore we can treat it as the only (equality) constraint. As a result, there is a 1:1 relationship between (ℓ^*, T_0^*) and (ℓ^*, β) , and we see that we can alternatively explore the space of all solutions by plotting the Lagrangian solution for all values of (ℓ^*, β) , as we had suggested earlier.

The second point is that there is a neater way of eliminating the socioeconomic drivers (λ and β) from the formulation of the problem, simply by working with the total system costs per day, rather than the unit cost per passenger carried. In the standards formulation we wrote formulas for $\$(s, H)$ and $T(s, H)$ with units per passenger. But if instead we had (equivalently) used $\$_T(s, H) \equiv \lambda \(s, H) , with units of cost per unit time and length, then you can see from the earlier notes that the parameter λ would not appear in any of our formulas for $\$_T(s, H)$. In fact, the mathematical program:

$$\begin{aligned} \min \quad & \$_T \\ \text{s.t.} \quad & T \leq T_0(\ell); \forall \ell \end{aligned}$$

would not include either of our socioeconomic drivers (λ or β) in its formulation! This allows you to find the optimum yearly cost and the system structure by defining a standard and nothing else. The socioeconomic variables enter the picture only when a city chooses the standards it can afford. The average cost per passenger carried expressed in units of local wages, which is $\$/\beta \equiv \$_T/(\lambda\beta)$, should be an important factor in any such decision.

Example: (optional problem for students to solve to understand these two ideas)

Show that the equivalent simple standard to the linear standard $T_0(\ell) = T_0 + P_0\ell$ for the lower bound formulation of the case with no transfers is:

$$\ell^* = \frac{\left(\frac{t_s}{v_a} \right)}{\left(P_0 - \frac{1}{v_{\max}} \right)^2} \text{ if } P_0 > \frac{1}{v_{\max}}$$

and that:

$$\left. \begin{aligned} T_0^* &= T_0 + \frac{P_0 \left(\frac{t_s}{v_a} \right)}{\left(P_0 - \frac{1}{v_{\max}} \right)^2} \\ S_0^* &= \frac{c_d}{T_0 - \frac{\left(\frac{t_s}{v_a} \right)}{\left(P_0 - \frac{1}{v_{\max}} \right)}} \end{aligned} \right\} \text{ if } T_0 > \frac{\left(\frac{t_s}{v_a} \right)}{P_0 - \frac{1}{v_{\max}}}$$

Note how the solution does not involve λ or β . Then, use the Lagrangian approach to show that the shadow price that would achieve the above is:

$$(\lambda\beta)^* = \frac{c_d}{\left(T_0 - \frac{t_s}{v_a \left(P_0 - \frac{1}{v_{\max}} \right)} \right)^2}$$

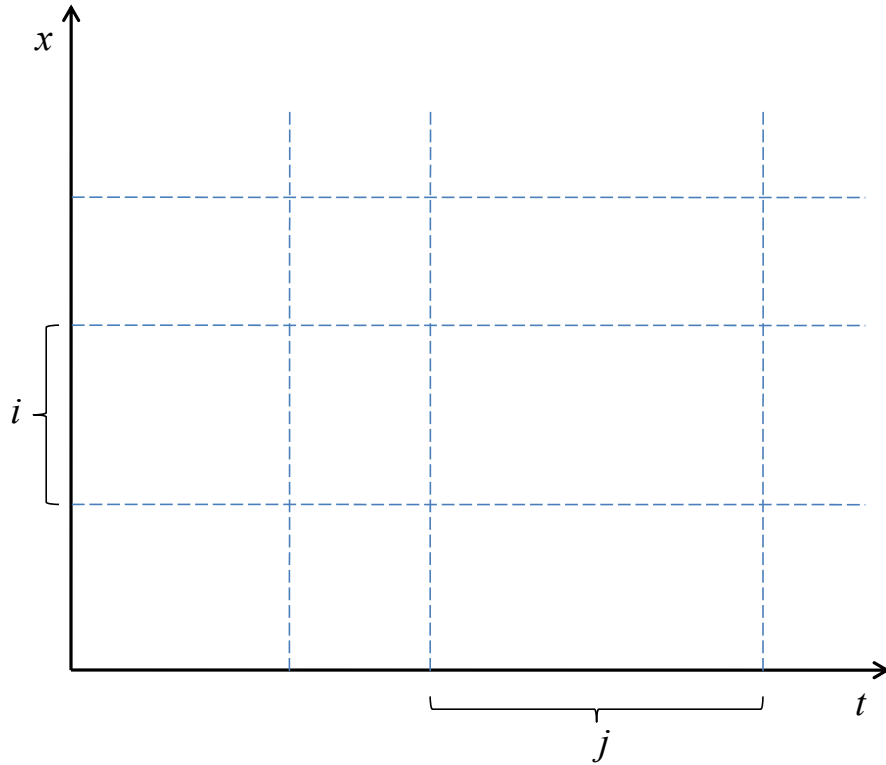
To repeat: The importance of this is that standards are connected to total costs, and you don't need anything else to determine this cost.

Space- and Time-Dependent Service

Assuming we have a corridor, we want to see how performance is affected by changing the design variables in space and time. Of our two decision variables, s and H , spacing is a physical aspect of the route, and so is only a function of space: $s(x)$, while headway will remain constant as buses travel the route, and so is only a function of time: $H(t)$. The (t, x) area of concern can be partitioned into space (i) and time (j) slices as shown below and we can find the cost of delivering service for s_i and H_j . We will do this first for average-case analysis (which you should know) and then for the service guarantee (standards) approach.

Average Case Analysis

For average case analysis, demand plays an important role, so we start by defining an OD matrix of trip selection rates. The OD matrix can be represented as $\lambda_{i'j}$, where i is the origin, i' the destination and j the time (the units of λ would be pax/time·dist²). We shall find that it is not necessary to use the entire OD matrix, only the relevant parts for which we want standards.



If we ignore the cost of stops, the total cost of service is:

$$\$_T = \sum_j \frac{c_d L}{H_j} T_j$$

Note: it does not depend on the OD matrix.

The generalized cost of waiting delay, where λ_j is the total number of trips generated per unit time along the complete corridor during time slice j (units of pax/time), is:

$$\beta \sum_j H_j (\lambda_j T_j)$$

Similarly, the generalized cost of inbound access is:

$$\beta \sum_i \frac{s_i}{4} \frac{1}{v_a} (\lambda_i L_i)$$

where λ_i is the total number of trips generated per unit distance with destinations for the whole corridor during the course of a day (units = pax/dist). Since the cost of egress should be the

same, we can multiply this equation by 2 to account for the total access cost. Finally, if we let A_i be the number of people crossing a screen-line in region i during the course of a day (units=pax/hr), we can express the generalized cost of stops as:

$$\beta \sum_i \left(\frac{L_i}{s_i} \right) t_s \Lambda_i$$

As you can see, we don't need to know the whole OD matrix, only the summary information embodied in $\{\lambda_i, \lambda_j, \text{ and } A_i\}$.

Also note that the optimization is very simple. The first two equations are functions of H_j and not s_i and can be optimized alone and separately for each time period. Likewise, the last two equations are functions of s_i and not H_j and can be optimized alone and separately for each location.

Service Guarantee Analysis

Instead of optimizing for the average case with a choice of β , we can choose a set of time standards $T_0(i, i', j)$ for selected origin and destination pairs and times of day. Then, there is no need to know the demand to estimate the optimum cost. It would be the job of policy-makers to decide on a reasonable standard. The objective function is the same as above, and the standards would simply introduce constraints of the form:

$$T_0(i, i', j) \geq AT_i + AT_{i'} + IVTT_{ii'} + H_j$$

for relevant sets of (i, i', j) . Note that the four terms of the RHS have simple subscripts. This MP can often be solved by introducing shadow prices and decomposing the Lagrangian into parts that can be optimized separately. If this does not work we can resort to a numerical solution.

Further Readings

The following readings may be useful to reinforce the concepts you have learned in this module.

Clarens, G. and Hurdle, V. (1975) "An operating strategy for a commuter bus system", *Transportation Science* **9**, 1-20. (Average-case analysis of non-hierarchical many-to-one 2-D systems with inhomogeneous demand.)

Wirasinghe, C.S., Hurdle, V.F. and Newell, G.F. (1977) "Optimal parameters for a coordinated rail and bus transit system" *Transportation Science* **11**, 359-74. (Average-case analysis of a 2-mode hierarchy serving 1-D, many-to-one demand.)

Module 4: Planning—Two-Dimensional Systems

(Originally compiled by Eric Gonzales and Josh Pilachowski, March, 2008)

(Last updated 9-22-2010)

Outline

- Idealized Case (New 2-D Issues)
 - Systems without Transfers
 - The Role of Transfers in 2-D Systems
- Realistic Case (No Hierarchy)
 - Logistic Cost Function (LCF) Components
 - Solution for Generic Insights
 - Modifications in Practical Applications
 - General Ideas for Design
- Realistic Case (Hierarchies--Qualitative Discussion)
- Time Dependence and Adaptation
- Capacity Constraints
- Comparing Collective and Individual Transportation

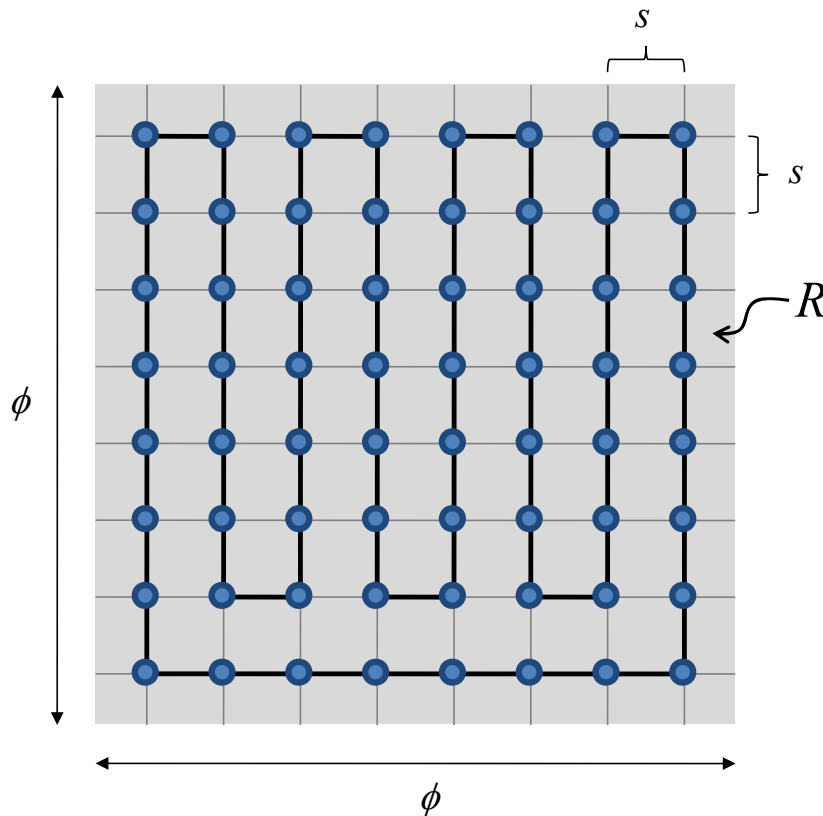
Remember from previous modules the types of systems we have analyzed. Shuttle systems had one decision variable, H , and could only be optimized temporally. Corridors had two decision variables, H and s , and could be optimized temporally and spatially. These design decisions defined all the passengers travel choices; i.e., when and where to board a transit vehicle. Think now about a two-dimensional system and the new travel choices available to passengers. This should illuminate the extra issues that must now enter into the analysis. They include considerations of total route length and layout, the role of transfers and travel circuitry. As before we start with an idealized analysis that isolates the new issues and then proceed with a more realistic treatment that combines them all.

Idealized Case

We will perform the idealized analysis in a similar manner as the corridor analysis. We consider a system with a single line with no transfers allowed and bi-directional service. We assume $H=0$ and $t_s=0$. For the two-dimensional system we will also assume that $a_0=\infty$, which removes all penalty for stopping meaning that $v=v_{max}$ at all times. We make this assumption because if we had allowed $a_0=\infty$ in the shuttle and corridor analysis then the door-to-door speed would be v_{max} . Yet, this turns out not to be true in the two-dimensional case. So, this set of assumptions allows us to isolate the new effect introduced by the second spatial dimension. Let us see...

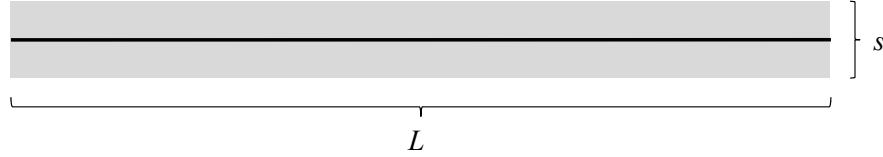
Systems without Transfers

Consider a square city with sides ϕ , area $R=\phi^2$ and an infinitely dense grid of streets; see figure below. No matter how long a transit line is, it cannot cover all points. Therefore, we anticipate that coverage and access become important issues in 2-D, and that our new decision variable will be route length and placement. To minimize worst-case access time in 2-D we should place stops on a (square) grid, with spacing s to be determined. The worst-case access time would then be $2s/v_w$ since there is an access distance of s at both the origin and destination. What then about travel time? Note that since stops don't matter it will be the maximum distance a person spends in a vehicle, divided by v_{max} . And since service is bi-directional, the maximum distance is $\frac{1}{2}$ of the length of the line, which we denote L . Thus, $IVTT=L/2v_{max}$, and we minimize $IVTT$ by choosing the shortest route to cover our lattice of stops. The problem of shortest-path routing for pre-existing points is a famous and complex problem known as the Traveling Salesman Problem. Fortunately for us, the solution for a two-dimensional lattice structure with an even number of points, such as the one shown above, is easy and efficient since there always is a path where the distance between any two consecutive stops along the route is s (you can convince yourself of this.)



Public Transportation Systems: Planning—Two Dimensional Systems

If you now imagine cutting the grid between parallel route lines then the area can be imagined as a corridor with length L and width s



where L is the total length of the route. Thus, the area can be expressed as:

$$Ls = R,$$

and the in-vehicle travel time for the worst case person would be:

$$IVTT = \frac{L}{2v_{\max}} = \frac{R}{2sv_{\max}}.$$

The door-to-door travel time guarantee is then:

$$t = \frac{2s}{v_w} + \frac{R}{2sv_{\max}}$$

Note: This is an EOQ expression with respect to the lattice spacing. When optimized the solutions is:

$$t^* = 2 \left(\frac{R}{v_w v_{\max}} \right)^{\frac{1}{2}} = 2 \frac{\phi}{\sqrt{v_w v_{\max}}}$$

This gives a door-to-door travel speed for the worst-case person:

$$\hat{v} \approx \frac{\phi}{t^*} = \frac{1}{2} \sqrt{v_w v_{\max}}$$

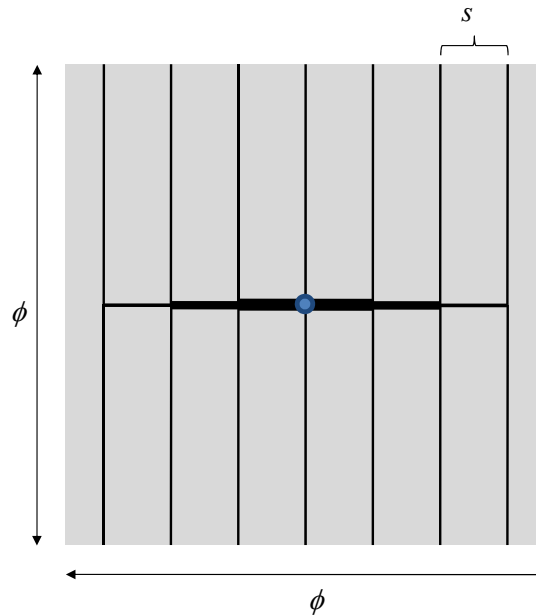
If we assume values of $v_w=3\text{kph}$ and $v_{\max}=36\text{kph}$, then the resulting door-to-door speed is $\hat{v} \approx 5.5\text{kph}$, which is not much faster than walking speed. The underperformance arises because to achieve low access time the route needs to be very winding. And buses in a windy route entrap passengers unfortunate to go a long distance. So, how can we improve the system? If we allow for transfers then passengers are no longer entrapped, and all we have to do is look for routings that give good coverage while providing good travel options to passengers that can transfer. So what are these routings? To get an understanding of this issue, we look at some idealized systems with one transfer.

Public Transportation Systems: Planning—Two Dimensional Systems

The Role of Transfers in 2-D Systems

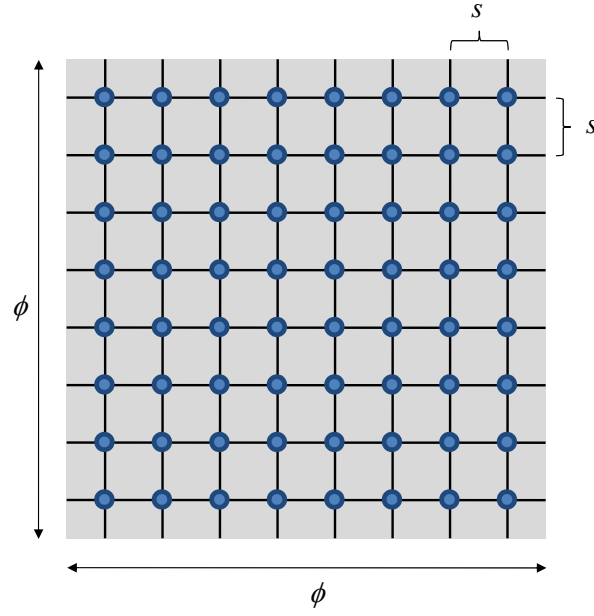
Two extreme possibilities are considered here. A hub and spoke system (H) with only one transfer point; and a grid system that allows for transfers at every stop. See the illustrations below. Note that for the same route spacing, the grid system requires more route-kilometers; so it should be more expensive to cover with vehicles.

Another disadvantage of the grid system relative to the hub is that coordination is more difficult. An advantage is that users can always choose a direct route without backtracking. We now compare the performance of these two systems (and of the no-transfer, single-line system (*O*)) for different values of L . This is reasonable because if one holds H and the commercial speed of vehicles invariant across scenarios, then L is the most important driver of cost.



A Hub and Spoke System (H)

Public Transportation Systems: Planning—Two Dimensional Systems



A square grid system (G)

We now change notation and use L to denote the kilometers of undirected service provided. A little bit of reflection shows that the total lengths of service for the three cases are:¹

$$L_0 = \frac{2\phi^2}{s}$$

$$L_H = \frac{3\phi^2}{s}$$

$$L_G = \frac{4\phi^2}{s}$$

For the same L , the three services provide different coverage, as represented by s :

$$s_O = \frac{2\phi^2}{L}; s_H = \frac{3\phi^2}{L}; s_G = \frac{4\phi^2}{L}$$

¹ To get these simple expressions, it is assumed that it takes 1 spacing to turn the buses at the end of each route.

Public Transportation Systems: Planning—Two Dimensional Systems

These values represent the sideways spacing between lines achieved by the three system types. Thus, the worst-case sideways access times are: $2\phi^2 / Lv_w$, $3\phi^2 / Lv_w$, and $4\phi^2 / Lv_w$. If we ignore the longitudinal access times (which should be the same for the three systems) and focus on cross-town trips (of length $\ell \approx \phi$), the worst-case door-to-door travel times are then:

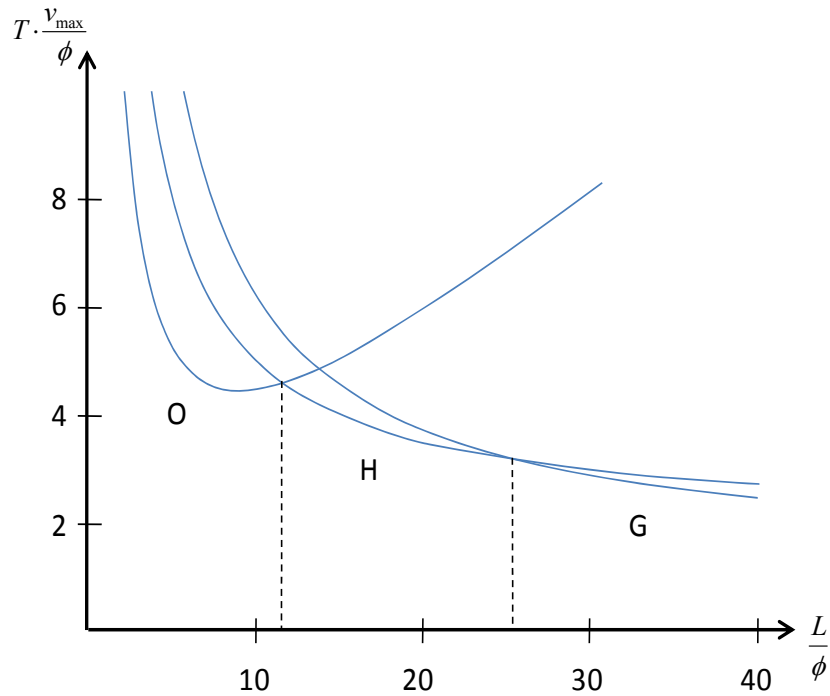
$$\begin{aligned} T_0 &= \frac{2\phi^2}{Lv_w} + \frac{L}{4v_{max}} \\ T_H &= \frac{3\phi^2}{Lv_w} + \frac{2\phi}{v_{max}} \\ T_G &= \frac{4\phi^2}{Lv_w} + \frac{\phi}{v_{max}} \end{aligned}$$

If we now choose $\frac{v_{max}}{v_w} \sim 10$ then we can compare the three cases based on the dimensionless variable $\frac{L}{\phi}$. The formulae become:

$$T = \frac{\phi}{v_{max}} \times \begin{cases} 20\left(\frac{\phi}{L}\right) + \frac{1}{4}\left(\frac{L}{\phi}\right) & (O) \\ 30\left(\frac{\phi}{L}\right) + 2 & (H) \\ 40\left(\frac{\phi}{L}\right) + 1 & (G) \end{cases}$$

These expressions can be expressed graphically as follows:

Public Transportation Systems: Planning—Two Dimensional Systems

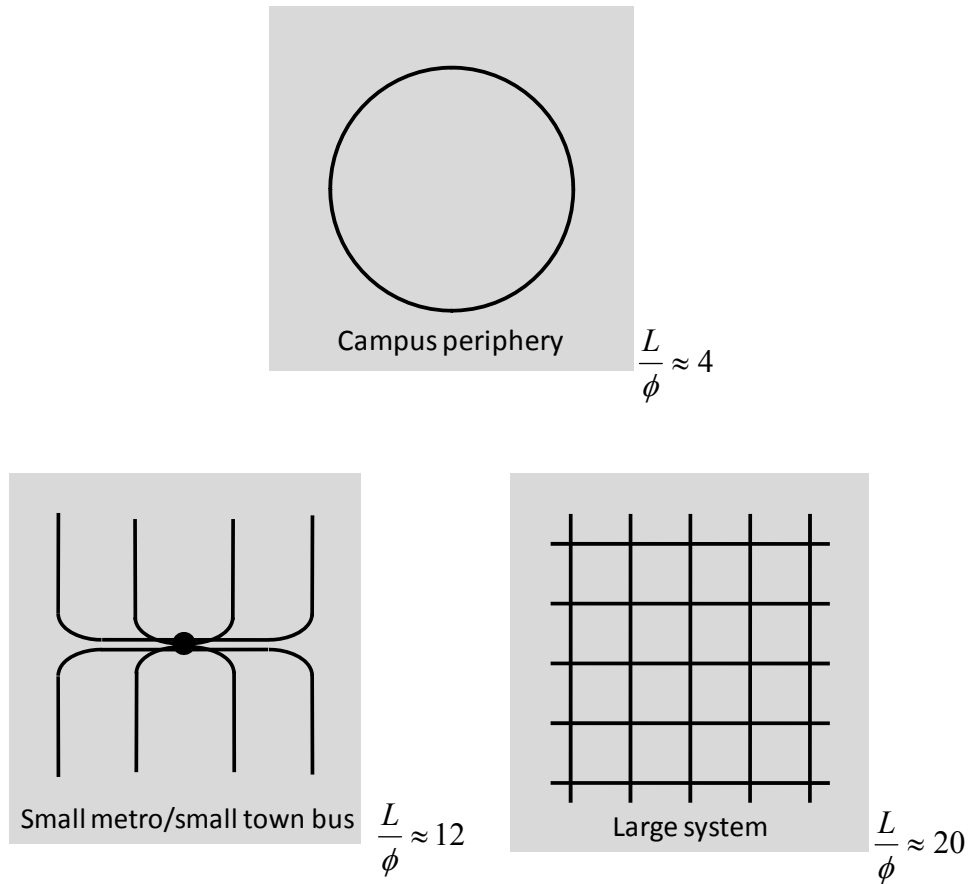


This illustrates that “short systems” with few stops, whose total length is not much greater than the perimeter of their service region do not require transfers. The figure also illustrates that long systems with many stops do benefit, and that in these cases the longer the system the greater the benefit.

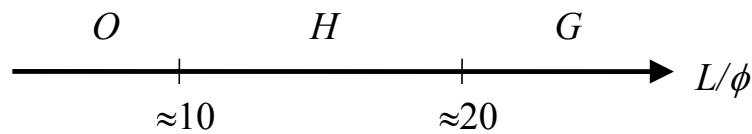
Just so you get a feel for the meaning of $\frac{L}{\phi}$ we look at four common routing examples and their

$\frac{L}{\phi}$ values:

Public Transportation Systems: Planning—Two Dimensional Systems



We find that the optimal routing layout depends on the value of $\frac{L}{\phi}$ and, if we add a 25% access penalty to the grid system to reflect the added cost of an uncoordinated transfer, we find that the critical points are as follows:



This explains why systems in real-life often have the structures shown in the above figures. Also, note that when we allow one transfer, then $\hat{v} = v_{\max}$ as $L \rightarrow \infty$ for the grid system. So, transfers really do help with performance in 2-D.

Realistic Case – No Hierarchies

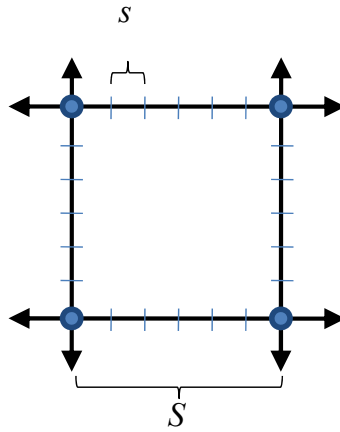
We could do a realistic analysis for each case we introduced earlier, however in the interest of time we will be concentrating on the grid case since it is the most useful for larger networks. Since we are dealing with worst case analysis we will also only concentrate on square grids. A rectangular grid would introduce directionality and add unneeded complexity.

First we will introduce stop spacing, s , within route spacing, S , such that $s < S$. We will also use S as a decision variable instead of L . We need to make assumptions about how people travel. In this case, we will assume that people only make one transfer and they choose their origin and destination stops in order to minimize their access distance.

We will then develop formulas for agency cost and passenger time (access + waiting + in-vehicle travel time)

Logistic Cost Function (LCF) Components

Recall that the transit service in 2 dimensions can be described by 3 decision variables: stop spacing s , line spacing S , and service headway H .



The total cost for such a system is cost of driving and stopping a bus multiplied by the number of buses operating per unit area.

$$\begin{aligned} \$_T &= \frac{4}{SH} \left(c_d + \frac{c_s}{s} \right) \quad \text{in units of } \frac{\text{cost}}{\text{time} \cdot \text{dist}^2} \\ \$ &= \frac{4}{\lambda SH} \left(c_d + \frac{c_s}{s} \right) \quad \text{in units of } \frac{\text{cost}}{\text{pax}} \end{aligned}$$

Notice this is very similar to the case for a corridor, the only difference being a factor $4/S$, expressing the fact that cost depends on the number of lines.

The travel time is composed of access time (AT), waiting time (WT), and in-vehicle travel time ($IVTT$) just as we saw for corridors. For the worst case passenger whose trip starts and ends as far as possible from transit service (the middle of the square),

$$AT = \frac{1}{v_w} \left[\left(\frac{s}{2} + \frac{S}{2} \right) + \left(\frac{s}{2} + \frac{S}{2} \right) \right] = \frac{s + S}{v_w}$$

$$WT = H + \Delta \text{ or } 2H + \Delta \text{ or } 3H + \Delta$$

where Δ represents time required to make a transfer, such as walking time from one stop to another. The number of headways included in WT depends on the assumptions we make about the synchronization of schedules (H if services are perfectly synchronized so that passengers only wait at the first stop where they board; $2H$ if services are not coordinated and passengers have to wait when they transfer, or else if service is coordinated but passengers have appointments at the destination; etc...).

$$IVTT = \ell_0 \left(\frac{1}{v_{\max}} + \frac{t_s}{s} \right)$$

where the longest possible trip length is $\ell_0 \approx 2\phi$. So, the worst case time for a 2-D system is given by the sum,

$$T = 2H + \Delta + \frac{S + s}{v_w} + \ell_0 \left(\frac{1}{v_{\max}} + \frac{t_s}{s} \right)$$

Notice again that this is very similar to the time associated with transit service in a corridor. The difference is the waiting time, $2H + \Delta$, and an additional component of access time, S/v_w .

Solution for Generic Insights

If we consider the lower bound of cost, assuming that the cost of stopping is small, the standards approach is described by the following mathematical program:

$$\min \frac{4c_d}{SH} \tag{6.1}$$

$$\text{s.t. } \left(2H + \frac{S}{v_w} \right) + T(\ell | s) \leq T_0(\ell) \tag{6.2}$$

where $T(\ell | s) = \Delta + \frac{s}{v_w} + \ell \left(\frac{1}{v_{\max}} + \frac{t_s}{s} \right)$.

The constraint will be an equality at optimality because for any $T(\ell|s)$ the cost is minimized by choosing the highest values of S and H . Therefore, the lower envelope method (explained in Module 3) can be used to solve for s^* , and with $T_L(\ell)$ we can determine ℓ^* . The mathematical program can thus be obtained with pencil and paper.

Alternatively, we can use the Lagrangian approach, expressing the generalized cost in dollars per person as

$$z_L = \frac{4c_d}{\lambda SH} + \beta \left(2H + \Delta + \frac{S+s}{v_w} + \frac{\ell}{v_{\max}} + \frac{\ell t_s}{s} \right) \quad (6.3)$$

which decomposes so that the stop spacing, s , is isolated. Solving for s^* and substituting,

$$s^* = \sqrt{\ell t_s v_w} \quad (6.4)$$

$$z_L = \frac{4c_d}{\lambda SH} + \beta \left(2H + \Delta + \frac{S}{v_w} + 2\sqrt{\frac{\ell t_s}{v_w}} + \frac{\ell}{v_{\max}} \right)$$

The optimal headway, H^* , and line spacing, S^* , can be solved in closed form.

$$H^* = \sqrt{\frac{2c_d}{\lambda S \beta}} \quad (6.5)$$

$$z_L^*(S) = 4\sqrt{\frac{2\beta c_d}{\lambda S}} + \beta \frac{S}{v_w} + \beta \left(\Delta + \frac{\ell}{v_{\max}} + 2\sqrt{\frac{\ell t_s}{v_w}} \right)$$

$$S^* = \left(\frac{8c_d v_w^2}{\lambda \beta} \right)^{\frac{1}{3}} \quad (6.6)$$

$$z_L^* = 6 \left(\frac{c_d \beta^2}{\lambda v_w} \right)^{\frac{1}{3}} + \beta \left(\Delta + \frac{\ell}{v_{\max}} + 2\sqrt{\frac{\ell t_s}{v_w}} \right)$$

We now compare this cost to the generalized cost for corridors, assuming the same values as in module 3:

$$v_w \cong 3 \text{ km/hr}$$

$$v_{\max} \cong 36 \text{ km/hr}$$

Public Transportation Systems: Planning—Two Dimensional Systems

$$t_s \cong 5 \times 10^{-3} \text{ hr}$$

$$c_d \cong 1 \text{ \$/km}$$

In 2-D, the generalized cost is

$$z_L^* = 4.2 \left(\frac{\beta^2}{\lambda} \right)^{\frac{1}{3}} + \beta \left(0.08\sqrt{\ell} + \frac{\ell}{36} \right)$$

and in universal units of time:

$$\frac{z_L^*}{\beta} = 4.2 \left(\frac{1}{\lambda\beta} \right)^{\frac{1}{3}} + 0.08\sqrt{\ell} + \frac{\ell}{36}$$

compared to a generalized cost in a corridor of

$$z_L^* = 2 \left(\frac{\beta}{\lambda} \right)^{\frac{1}{2}} + \beta \left(0.08\sqrt{\ell} + \frac{\ell}{36} \right)$$
$$\frac{z_L^*}{\beta} = 2 \left(\frac{1}{\lambda\beta} \right)^{\frac{1}{2}} + 0.08\sqrt{\ell} + \frac{\ell}{36}$$

Note that the universal generalized cost per person declines with demand, λ , and wealth, β , more slowly in the 2D case than in the 1D case. In other words, the second dimension somewhat dilutes the economies of scale in collective transportation. Note too that the effect of distance is the same in both cases.

Remember, however, that λ is expressed in demand per area in the 2-D case, and demand per distance in the corridor case, so these expressions cannot be compared for the “same” λ .

For the hypothetical case of long trips in a relatively poor city ($\ell_0 = 40$ km, $\beta = \$1$ / hour, and $\lambda = 10^3$ pax / hr·km²), the generalized cost $z_L/\beta = 2.1$ hours which decomposes to 0.17 hours of work, 0.9 hours of delay (access, waiting, and in-vehicle stopping), and 1.11 hours of travel time (like in a car).

Modifications for Practical Applications

- 1) Some lines may require fixed infrastructure (BRT, rail, etc.), so the cost of construction, bond finance, etc. should be amortized over the life of the infrastructure. Convince yourselves that for an infrastructure cost r_s \$/hr-stop, this contributes

$$\frac{r_s}{\lambda s S}$$

to the objective function.

- 2) Stops may be skipped if the demand is low. In this case we work with expectations.

$$E(\text{time stopped per unit length}) = E(\# \text{ pax boarding/alighting moves per distance}) t_m + E(\# \text{ stops per distance}) t_s$$

where t_m is the marginal time for one passenger move and t_s is the marginal time for a vehicle stop. The expectations are now given. First, note that

$$E(\# \text{ of pax moves per stop}) = 2\lambda H s S$$

Therefore, since there are $1/s$ stops per km;

$$E(\# \text{ pax moves per distance}) = \frac{1}{s} E(\# \text{ pax moves per stop}) = 2\lambda H S,$$

and

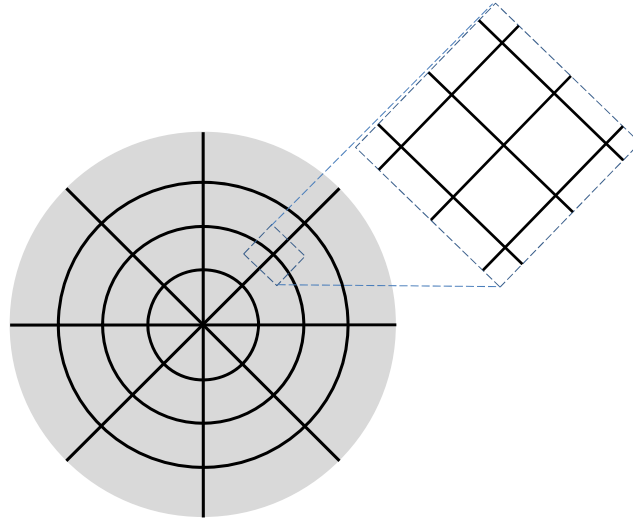
$$E(\# \text{ stops}) = \frac{1}{s} \Pr\{\text{stopping}\},$$

where

$$\Pr\{\text{stopping}\} = (1 - e^{-2\lambda H s S})$$

if the demand for stops follows a Poisson process with the given mean.

- 3) Cities have centers, so we may want to orient our grid towards the center. Notice that if we zoom in on a part of a ring-radial network it looks like a grid. Nothing prevents us from making a constant density of service in a ring-radial network by adding radial lines as we move out from the city center.

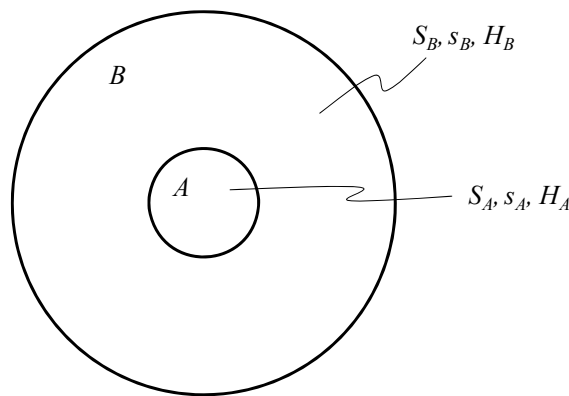


We can also use this strategy if we want to have the flexibility to have different densities of service and headways in different parts of the city as shown in the figure below. To do this systematically, we can set different standards for trips in different parts of the city. For example,

$$T_0^{(A)}(\ell) \text{ for all trips in A}$$

$$T_0^{(B)}(\ell) \text{ for all trips in B (or, even better, } B \cup A)$$

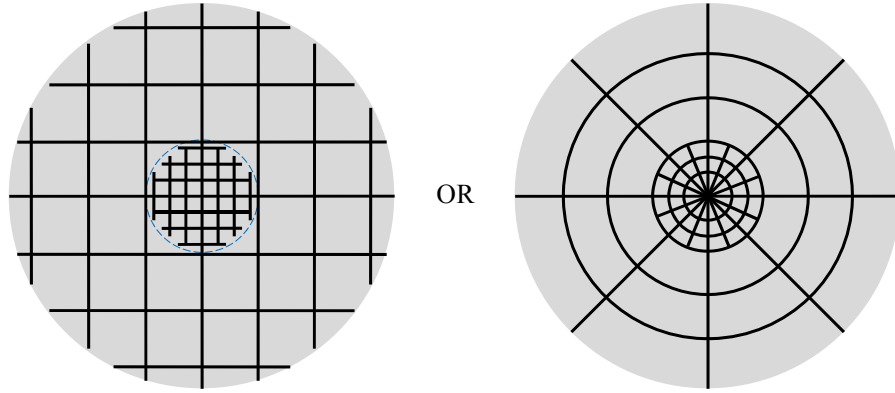
$$T_0^{(AB)}(\ell) \text{ for all trips between A and B}$$



General Ideas for Design

- 1) Think of a family of design concepts, qualitatively – e.g. grid system, ring-radial network, etc.
- 2) Identify members of the family by list of decision variables – e.g. stop spacing s , line spacing S , and headway H .

- 3) Estimate the cost and translate the specific concept into a detailed plan – e.g.



Considering all regions ($r = A, \dots$) and time periods of the day ($j = 1, 2, \dots$) solve the following mathematical program for the decision variables: $\{\dots, (s_r, S_r), \dots\}$ and $\{\dots, H_{rj}, \dots\}$:

$$\begin{aligned} \min \$_T &= \sum_{\substack{r=A, \dots \\ j=1, 2, \dots}} \$_T^{(r)}(S_r, H_{rj}) \\ \text{s.t. } T_A(s_A, S_A, H_{Aj}) &\leq T_{0Aj}, j = \text{rush, off-peak, night} \\ T_B(s_A, S_A, s_B, S_B, H_{Bj}) &\leq T_{0Bj}, j = \text{rush, off-peak, night} \\ T_{AB}(s_A, S_A, s_B, S_B, H_{Aj}, H_{Bj}) &\leq T_{0ABj}, j = \text{rush, off-peak, night} \end{aligned}$$

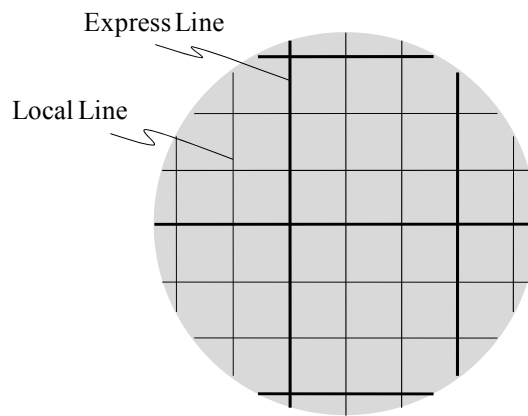
You may want to include separate constraints for access time or waiting time, depending on what the city wants, but *you should always use your judgement*. Anything is possible, but the more complicated the problem, the more difficult it is to solve the problem exactly.

Lagrangian decomposition can help us solve this mathematical program. It may be possible to simplify the problem and eliminate many of the decision variables. So, we can use shadow prices to simplify these complicated mathematical problems by assigning a different β to each of the constraints. Increasing the values of β will reduce the left side of the constraint when the Lagrangian is optimized, so we start with an estimated value of β and then increase it until the constraints are met. If we have a closed form for the optimal decision variable values in terms of β , it is easy to adjust the solution by changing the shadow price.

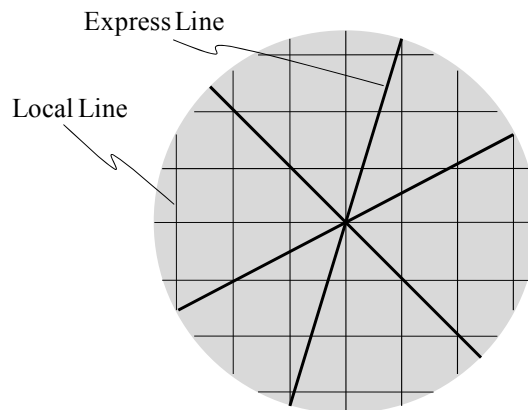
Note: This approach can be used to solve (6.1,2) by working instead with (6.3). All you should have to do is plug in (6.4,5,6) into (6.2) and find the β that solves (6.2) as an equality.

2-D Systems: Realistic Case (Hierarchies)

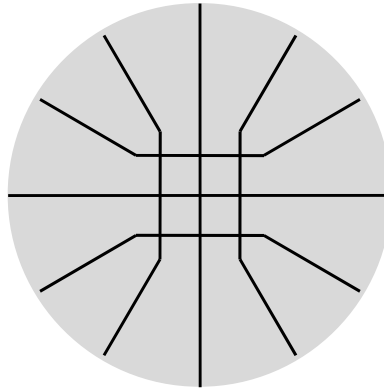
Until now, we have looked only at local systems in 2-D. However, we could introduce a hierarchy with the same method as for corridors (see module 3). There are now decision variables for the stop spacing and line spacing of both the local and express services.



The introduction of hierarchies also gives us the flexibility to design non-isotropic systems. For example, a grid may serve as a basis for local buses guaranteeing a length-based but uniform standard for short-medium trips. A radial express service may be overlaid to provide better service for inter-zonal travel (e.g. Chicago). Such an express network may be described in as few as 3 additional decision variables: # of radial lines, # of ring lines, service headway.



Perhaps one system can be designed to act a radial network outside of the city center and look more like a grid in the city center (e.g. Washington DC, London). The possibilities are many, but in all cases the goal is to reduce these concepts to as few descriptors as possible which will describe the shape and design that the system should have once the variables are chosen.



2-D Systems: Time-Dependence and Adaptation

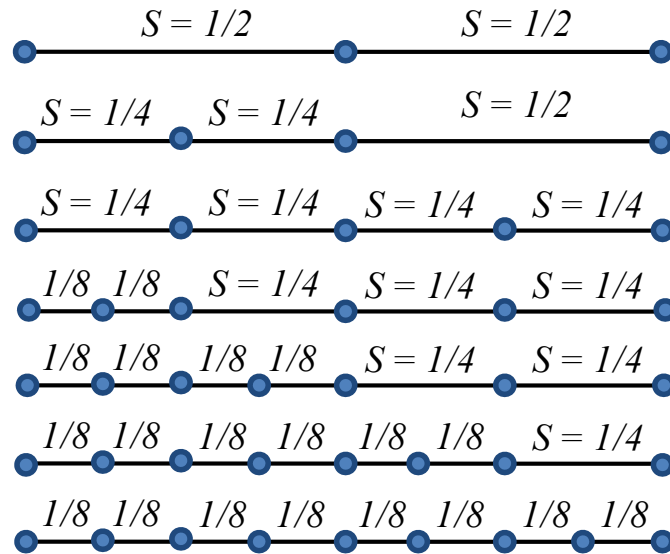
Over time, demand for transportation in a city changes. Some of the decision variables are easier to change over time than others. The headway, H , can be varied very easily even within the course of a day. The stop spacing, s , can be changed with a little more effort, and line spacing, S , is relatively fixed.

Suppose we have a linear city of length one, and we place a station to minimize access distance. A single station divides the city into two halves and should be placed in the center to minimize worst case and average access distance; $S^* = 1/2$.



As demand grows over time, we may want to add stations incrementally to the city, one at a time. If we can pull up old stations and always re-optimize, the placement should make the spacing follow the progression, $S^* = 1/2, 1/3, 1/4, 1/5$, etc. However, if the stations are fixed once they are placed, subsequent placement of stations will not always give us the minimum access cost.

For a worst case analysis, imagine that we have the city above with $n = 2$ spacings ($S^* = 1/2$) and we add one more station. Only half of the city benefits, so the worst case access is unchanged. The worst case access cost is only improved when symmetry is established at $n = 2, 4, 8, \dots$. The incremental addition of stations following this naïve approach is shown below:



Is there a way to place stations so that each incremental addition of a station improves the worst case access? In fact there is!

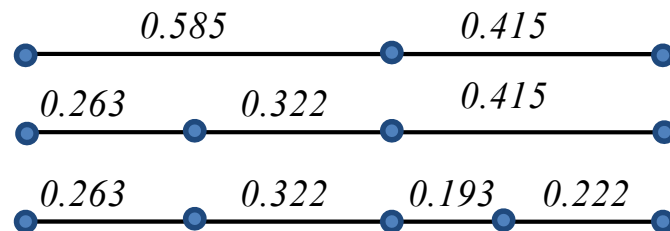
If stations are placed so that for $n+1$ stations, the placement results in spacings of length

$$x_i^{(n)} = \log_2(i+n) - \log_2(i+n-1) \text{ for } i = 1, 2, \dots, n$$

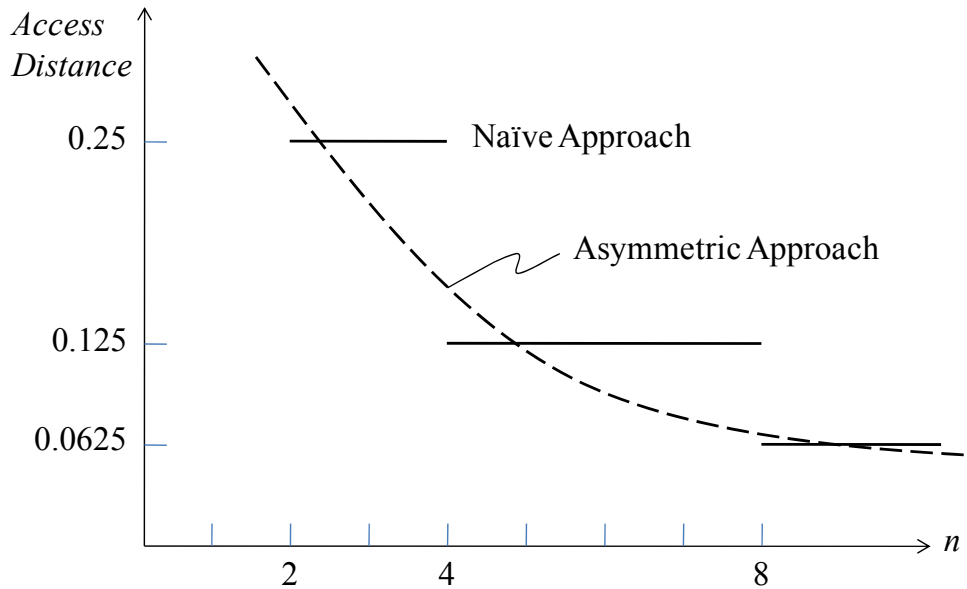
Note that this always satisfies $\sum_{i=1}^n \log_2(i+n) - \log_2(i+n-1) = 1$, and also $x_n^{(n+1)} + x_n^{(n+1)} = x_1^{(n)}$.

The latter equation shows that the largest spacing becomes the sum of the smallest two (after the split). This is in fact how the equation is derived.

The incremental addition of stations would now look like the progression below.



If we plot the worst case access distance against number of stop spacings, this asymmetric approach is better than the symmetric naïve approach most of the time.



For average case analysis, the average access distance when station expansion follows this specific asymmetric recipe is $1.04/4n$. If the stations could be picked up and re-placed optimally every time the system is expanded, the average access distance would be $1/4n$.² So the penalty for not being able to move transit lines once they are placed can be as low as 4%. This is good news, because it means fixed infrastructure that is extremely costly to change (such as subway tunnels) are always near optimal!

Capacity Constraints

One additional point that can be added to this module is the idea of ridership (i.e., the number of passengers in a bus). This can be useful in planning because it allows for the addition of vehicle size (capacity) constraints to the design. For a grid system with spacing S , the average ridership can be roughly approximated by:

$$\text{Ridership} \approx \frac{\lambda S^2 H \ell}{4S}.$$

This expression arises from considering a single $S \times S$ cell of our grid. The numerator is the number of pax-km generated in the cell in a single headway, and the denominator is the number

² As an exercise, derive the average case analysis cost for the naïve case where the stations cannot be moved. The result may surprise you!

of bus-km traveled per cell in one headway. Assuming optimal design parameters developed in eqs. 6.5 and 6.6 the result is:

$$\text{Ridership} \cong \left(\frac{c_d^2 v_w}{16} \right)^{\frac{1}{3}} \lambda^{\frac{1}{3}} \beta^{-\frac{2}{3}} \ell.$$

If we adopt the typical numbers we have been using, we see that for a wealthy city ($\beta \sim 10$) with $\ell \sim 10$ km, a demand of $\lambda \sim 1$ pax/km²-hr would yield an average ridership of about 1 pax per bus. So, we would not expect collective transportation (CT) to be a feasible option for delivering mobility if λ was much smaller than 1 pax/km²-hr. And what to do in this case will be the topic of the next module.

Comparing Collective and Individual Transportation

Recall from the equation immediately following (6.6) that our average cost function for 2-D systems with no hierarchies and allowing for transfers could be expressed as:

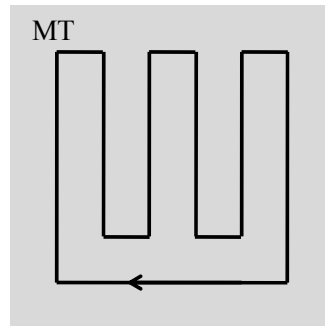
$$\left(\frac{z^*}{\beta} - \frac{\ell}{v} \right) = \Delta + 2 \sqrt{\frac{\ell t_s}{v_w}} + 6 \left(\frac{c_d}{\beta \lambda v_w} \right)^{\frac{1}{3}} \quad (7.1)$$

The RHS is the extra cost of CT over and above the unavoidable (time) cost of overcoming distance. We can see that CT has nice economies of scale. Demand and extra cost are inversely related, so as demand rises, the cost per passenger decreases. (In fact, if Δ and t_s are neglected, the extra cost tends to 0 as $\lambda \rightarrow \infty$; it can be virtually eliminated.) This is the beauty of CT. However, in cases of low demand ($\lambda \rightarrow 0$) the cost can become quite substantial. This is expected, because the system is there regardless of the ridership it generates. In these cases, collective transportation isn't very efficient.

As a point of comparison, consider the cost function for individual transit (IT), something like a taxi, which looks like this:

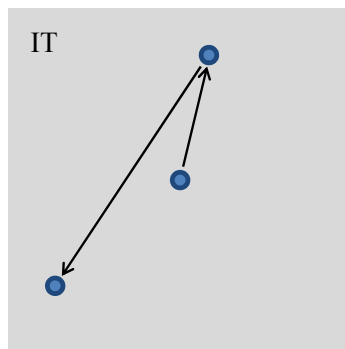
$$\left(\frac{z^*}{\beta} - \frac{\ell}{v} \right) = c_I \frac{\ell}{\beta}$$

where c_I is the cost per unit distance of providing IT. Notice that the extra cost of IT no longer declines with increasing demand (as it did with CT); and that it does not go to infinity in cases of low demand (as it did with CT). To further understand the source of this difference we now show a more detailed comparison between the two systems with one vehicle, assuming $\lambda \rightarrow 0$.



We consider a square of area R with an underlying grid of streets and assume that buses stop on demand. With CT, the decision variable is length of route and we use the idealized analysis of 2D-systems at the start of this module to evaluate the optimum trade-off between access time and in-vehicle travel time. With an optimal length of route, L^* , the expected door-to-door time is³:

$$E(t_m) = \sqrt{\frac{R}{vv_w}}.$$



With IT, the route can be directed where the people are, removing all access time. Assuming that the vehicle starts from a central location every time a request is made, the expected door-to-door time can be shown to be:

$$E(t_I) = \frac{7}{6} \frac{\sqrt{R}}{v}.$$

³ The result differs from our earlier result by a factor of $\sqrt{2}$ because (by stopping on demand) the current system eliminates the access distance parallel to the transit route.

The ratio of expected time between the two systems is approximately 3. Note that the cost of both systems is similar since both use one vehicle and one driver. So, in cases of low demand, IT will outperform CT by a large margin. This is due to its flexibility in routing. Therefore the next module will explore possible ways of delivering this flexibility for systems with higher (albeit still low) demand.

Further Readings

The following readings may be useful to reinforce the concepts you have learned.

Holroyd, E. M. (1967) "The optimum bus service: a theoretical model for a large uniform urban area" in *Vehicular Traffic Science* (L.C. Edie, R. Herman and R. Rothery, editors) Proc. 3rd ISTTF pp. 308-328, Elsevier. (Average-case analysis of non-hierarchical many-to-many grid systems with uniform demand. The passenger routing model in this beautiful reference is complex; unfortunately this complicates the formulae, obscuring possible insights.)

Daganzo, C.F. (2009) "Structure of competitive transit networks" *Transportation Research Part B* (in press). (This reading is of interest because it generalizes the ideas in this module by exploring a general family of systems that includes the hub-and-spoke and the grid concepts as special cases. The simple formulas it gives, are used to compare the performance of different technologies (Bus, BRT, LRT and Metro) in different urban contexts.)

Module 5: Planning—Flexible Transit

(Originally compiled by Eric Gonzales and Josh Pilachowski, March, 2008)

(Last updated 9-22-2010)

Outline

- Ways of delivering flexibility
- Taxis
- Dial-a-Ride (DAR)
- Car-Share

We saw in the last module that conventional forms of collective transportation with fixed access points and routes deliver better service than individual transportation when one or more of the following factors are high: the demand rate, the typical trip length and/or the time value of money. When this happens travelers gain if they trade-off the flexibility in routing and timing of individual transportation for the lower costs of collective transportation. Since there is also a grey area where the choice is not so clear, one may ask whether collective transportation can be made more flexible so it can better compete with individual transportation in situations like these. This is the question addressed in this module.

Ways of Delivering Flexibility with Public Transportation

We divide public transportation concepts depending on whether or not people share rides.

Individual Public Transportation (IPT)

Taxi – A driver takes passengers directly from their origin to their destination

Car-share – Users pick up a vehicle at one of many predetermined locations (pods), then return it to any pod when finished

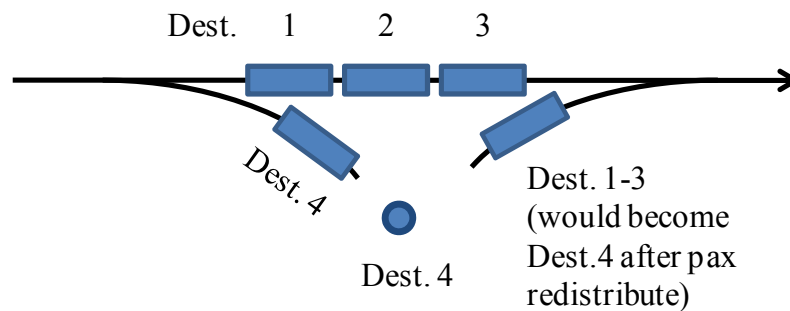
Driverless Taxi (futuristic) – Same as taxi, only without a driver required. The military is currently developing these types of vehicles for urban warfare, but they could also be used during peace time. See <http://www.darpa.mil/grandchallenge/overview.asp> for more information.

Personal Rapid Transit (PRT) (futuristic) – Small occupancy vehicles would travel along existing guideways. See http://en.wikipedia.org/wiki/RUF_%28dual_mode_transit%29 for more information.

Collective Transportation (CT)

Dial-a-ride (DAR) – Same as a taxi, but with multiple users sharing the vehicle. FIFO is not guaranteed.

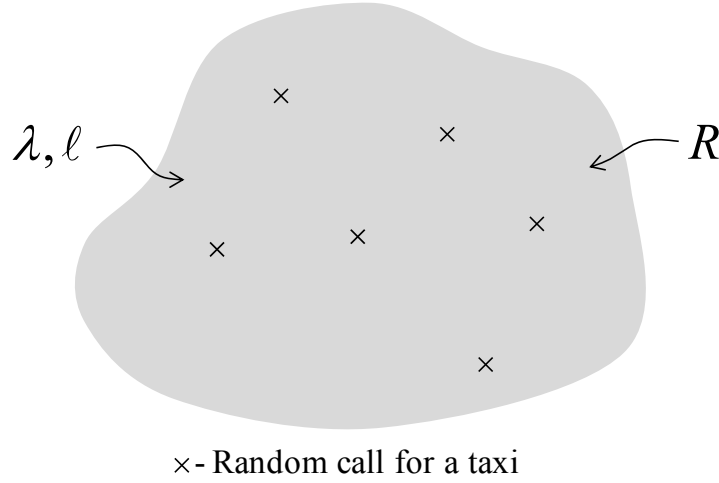
Long Distance DAR with transfer on-the-fly (futuristic) – Train with passengers in cars grouped by destination. Cars would couple and decouple without slowing the train and then drop off and pick up individuals as DAR. Passengers would walk in the train to the appropriate car. This could be useful as a substitute to long distance trains. A t - x diagram can be used to estimate feasible stop spacings.



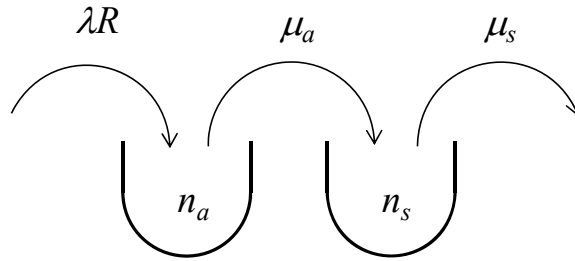
The main advantage of flexible IPT over the automobile is that public service cars are shared; by being used most of the time they require less parking infrastructure. The added promise of flexible CT is that perhaps it can reduce cost with only a small degradation in LOS, and also provide some economies of scale with respect to demand. So, with this in mind, we now examine what existing flexible concepts can do.

Taxis

Consider a region of area R in which we provide radio-taxi service to customers with demand density λ and expected trip length ℓ . We will assume that we provide enough taxis, m , to ensure that every call immediately gets assigned a taxi.



The fleet size, m , can be divided into three types of taxis: n_i idle taxis, n_a assigned taxis, and n_s servicing taxis. The cost of the system will be roughly proportional to m . We measure LOS by the waiting time a user experiences between requesting a taxi and being picked up. The following diagram shows the rate at which taxis switch from one state to another.



Using Little's Formula, and using T_a for the expected customer waiting time, we see that:

$$\mu_a = \frac{n_a}{T_a}; \text{ with } T_a \sim t_s + \frac{E(d_{n_i})}{v},$$

where $E(d_{n_i})$ is the expected distance the closest idle taxi to a request will have to travel. An expression for this expectation is obtained by imagining a region of area R , with n_i points and a circular disk with diameter $2x$. Because the n_i taxis are randomly distributed we can write:

$$\Pr\{d_n \geq x\} \equiv \Pr\{\text{zero points in the disc}\} = \left(1 - \frac{\pi x^2}{R}\right)^{n_i}.$$

From this distribution function we find (the proof for this can be found in the appendix at the end of this module):

$$E(d_{n_i}) \cong \frac{1}{2} \sqrt{\frac{R}{n_i}}.$$

Also using Little's Formula, and using T_s for the expected customer service time, we see that:

$$\mu_s = \frac{n_s}{T_s}; \text{ with } T_s \sim \frac{\ell}{v} + t_s.$$

If the system is in equilibrium, then:

$$\lambda R = \frac{n_a}{T_a} = \frac{n_s}{T_s}$$

This gives us two equations with three unknowns: n_i , n_a and n_s . To get a third equation we assume a target service level with average waiting time T_0 :

$$T_0(\ell) = T_0 + 2t_s + \frac{\ell}{v} = 2t_s + \frac{\ell}{v} + T_a \Rightarrow T_a = T_0.$$

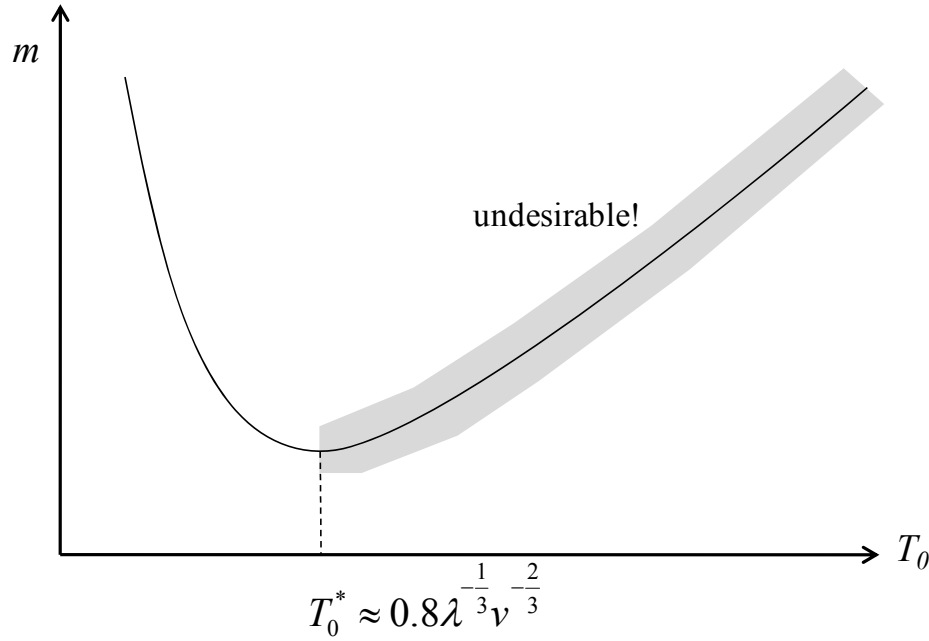
Using this as the third equation, and neglecting t_s as a reasonable first approximation, we find that the equilibrium solution is:

$$\begin{aligned} n_i &= \frac{R}{(2vT_0)^2} \\ n_a &= \lambda R(T_0) \\ n_s &= \lambda R\left(\frac{\ell}{v}\right) \end{aligned}$$

which gives the minimum fleet size required to achieve the target level of service:

$$m = n_i + n_a + n_s = \frac{R}{(2vT_0)^2} + \lambda R\left(T_0 + \frac{\ell}{v}\right).$$

Notice that m declines with T_0 up to a point, and then starts rising again. The rising branch is undesirable.



The shape of the curve means that there is a minimum fleet size required, $m^* \approx 1.2R\lambda^{\frac{2}{3}}v^{-\frac{2}{3}} + \lambda R \ell / v$, to ensure that incoming calls are assigned a taxi without delay, and that the worst LOS this kind of system should provide, T_0^* , is also bounded.

We see that the least possible extra generalized cost (in universal units) of delivering this type of service is:

$$\left(\frac{z_T^*}{\beta} - \frac{\ell}{v} \right) = T_0^* + T_s + m^* \frac{\gamma}{\beta} \frac{1}{\lambda R} \approx \left(0.8 + 1.2 \frac{\gamma}{\beta} \right) \lambda^{\frac{1}{3}} v^{\frac{2}{3}} + \frac{\gamma}{\beta} \frac{\ell}{v}$$

where γ is a taxi's cost per unit time (which should be greater than β). Note that the extra cost is at least ℓ/v even for $\lambda \rightarrow 0$. So taxis do not have significant economies of scale.

Example:

$$\left. \begin{array}{l} T_0 = 0.2hrs \\ t_s = 0.02hrs \\ \lambda = 1pax/hr - km^2 \\ R = 400km^2 \\ v = 20km/hr \\ \ell \sim 10km \end{array} \right\} \Rightarrow \begin{array}{l} n_a = 88 \\ n_i = 7 \\ n_s = 280 \\ \overline{m} = 375 \end{array} \quad \text{such a low number of idle taxis is bad for new users.}$$

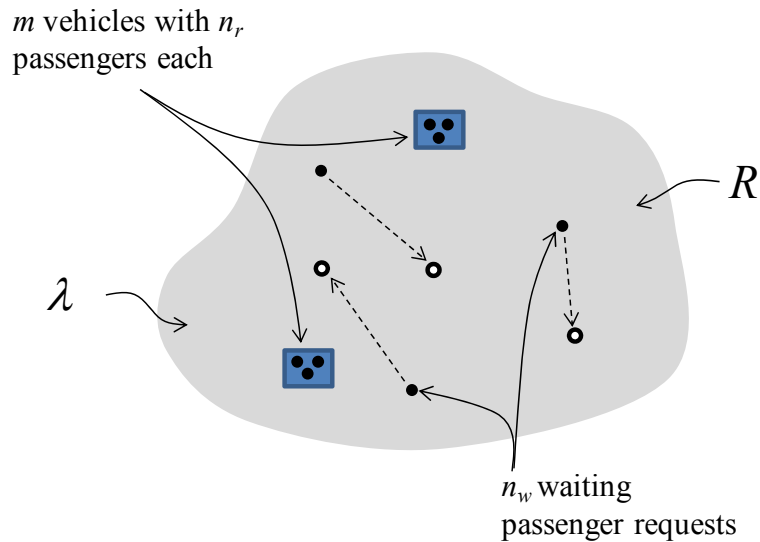
Note, for this system $T_0^* \approx 0.1$, so we should be able to reduce both T_0 and m . So, if we instead change $T_0 = 0.08hrs$ then the new result is:

$$\begin{array}{r} n_a = 40 \\ n_i = 30 \\ n_s = 280 \\ \hline m = 350 \end{array}$$

A better level of service and less taxis needed! Clearly, operating with $T_0 > T_0^*$ should be avoided. If T_0 is too large the trip times to collect passengers will be long, requiring many taxis in collection mode, and the equilibrium becomes unstable.¹

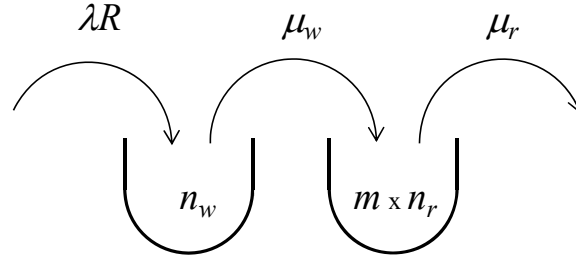
Dial-a-Ride

If taxis can have more than one passenger, then you can remove idle vehicles from the system and reduce cost. Service calls would then go directly to vehicles currently in service. To analyze this system, we again consider a region of area R with demand density λ . We will assume that origins and destinations are uniformly distributed throughout the region.



¹ A surge in demand would: further increase the number of taxis in collection mode; reduce the number of idle taxis, thereby increasing distance to new users and their collection times. With larger collection times, the number of taxis in collection mode would increase some more, etc...

Passengers will be divided into those waiting at home, n_w , and those inside vehicles, n_r . The fleet size is m . The following diagram shows the flow of passengers from one state to another:



We use the following assumptions (favorable to DAR):

- After achieving a desired passenger load, n_r , buses alternate between pickup and drop-off
- Buses pickup the passenger closest to them
- Buses drop of the passenger with the closest destination

Using Little's Formula, equilibrium assumptions, and the expected distance equations from before we can assume that:

$$\mu_r = \frac{m}{\bar{t}_p + \bar{t}_d} = \lambda R, \quad (7.2)$$

where \bar{t}_p and \bar{t}_d are average time to pickup and drop-off respectively, and the distances for each are:

$$d_d = \frac{1}{2} \sqrt{\frac{R}{n_r}}; \quad d_p = \frac{1}{2} \sqrt{\frac{R}{n_w}}. \quad (7.3)$$

For any choice of m and n_r (decision variables) we can use (7.2) and (7.3) to find $n_w(n_r, m)$. We then choose the value of n_r that would minimize the number of people in the system:

$$p = mn_r + n_w(n_r, m).$$

A little bit of algebra shows that the result is:

$$p^* = \frac{c^2}{m}, \text{ where } c = \frac{\lambda \sqrt{RR}}{2v}. \quad (7.4)$$

Note: c has the meaning of number of requests that the system would receive in the time it takes a bus to travel across the region (dimensionless demand).

The generalized cost per unit time for the system would then be:

$$\gamma m + \beta p^*,$$

where γ and β are the costs per unit time of one bus and one passenger. On a “per passenger” basis the cost is:

$$z_{DAR}^* = \$^* + T^* = \frac{1}{\lambda R} \left(\gamma m + \beta \frac{c^2}{m} \right)$$

Optimizing this EOQ expression gives us:

$$m^* = \left(\frac{\beta c^2}{\gamma} \right)^{1/2} \sim c$$

$$\$^* = T^* \sim \frac{\sqrt{\gamma \beta} c}{\lambda R} = \frac{\sqrt{\gamma \beta R}}{2v}. \quad (7.5)$$

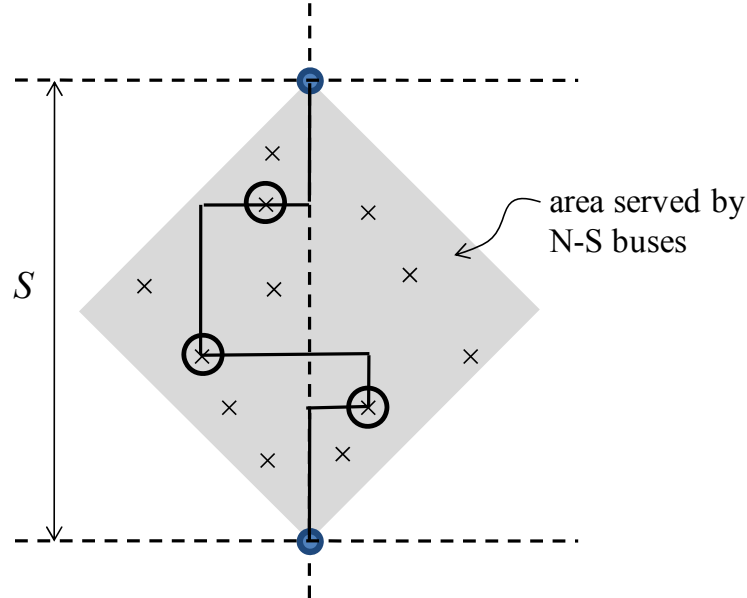
Note: λ does not appear in this equation. So *DAR has no economies of scale*. Recall however that the cost of taxi service would have been:

$$\$^* + T^* \sim \frac{(\beta + \gamma) \sqrt{R}}{v}.$$

So, particularly when $\beta \ll \gamma$ (poor populations), (7.5) is a significant improvement over taxi service.

Could there be a form of DAR with economies of scale? We suspect that introducing transfers may produce a positive answer to this question because, after all, fixed route CT does not have economies of scale without transfers – but it does with transfers.² So, perhaps by introducing transfers into the DAR concept economies could be achieved. One possibility is as follows. Remember our 2D network where we had to trade-off between number of stops and commercial speed. We introduced some flexibility by not stopping at every stop and showed how this could be studied. We can generalize this by allowing buses to detour and serve passengers at their origins and destinations as shown in the figure. (The figure shows an N-S route; if a similar pattern is used for E-W routes then the whole space is covered.). Users could activate a stop near their origin and the bus would detour through the region to pick up each passenger.

² You can convince yourselves of this by introducing demand into the idealized analysis of Module 4.



A trade-off would arise between the detour to pick up passengers and the elimination of access time. For any given detour the served passenger (or “customer”) would gain the benefit of zero access time, while passengers on the bus would be penalized by the extra travel distance.

The average access distance saved by the customer (assuming all his/her locations are equally likely) can be shown to be:³ $S/6$. Thus, the average time benefit is:

$$\text{time benefit of a detour} \approx \frac{S}{6v_w}.$$

Now let us examine the penalty. If the demand is so low that buses rarely make more than one stop per interval, S , then the bus would return to its original route after each detour, and the distance added by a detour would be twice as large as that saved by the customer; i.e., $2(S/6)$ on average. But this situation is pessimistic. If buses make multiple stops they do not have to return to the mainline after every detour (see figure) and this reduces the average detour distance. In the most optimistic case, where buses make a very large number of detours, the average detour

³ By symmetry, it suffices to consider customers in the bottom half of the shaded square. Let $y \in (0, S/2)$ be any such customer's vertical distance from the bottom of the square, and x his/her access distance. Since all locations are equally likely $E(x|y) = y/2$. Now, note that the p.d.f. of y is triangular because the number of possible locations is proportional to y . As a result $E(y) = S/3$. It then follows that $E(x) = E(E(x|y)) = E(y/2) = S/6$.

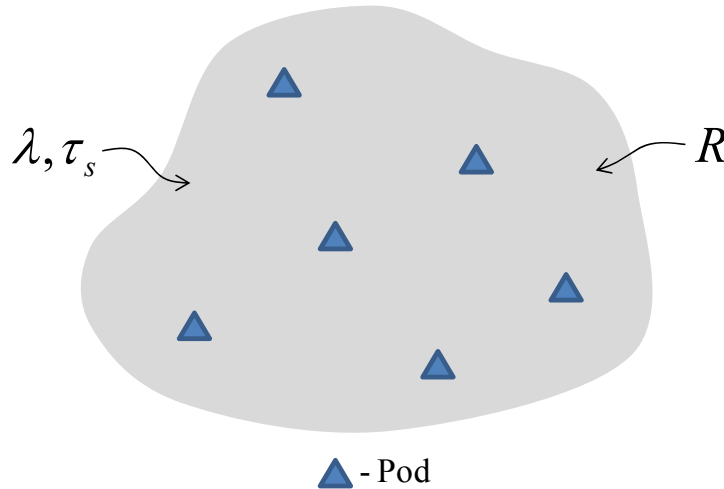
distance turns out to be:⁴ $2S/9$. So splitting the difference between the pessimistic and optimistic cases, we estimate the distance added by a detour as $5S/18$, and the collective time penalty as:

$$\text{time penalty of a detour} \approx \frac{5S}{18v} (\# \text{ of passengers in the bus}).$$

So setting the penalty equal to the benefit we see that the number of passengers in the bus should be no greater than $(3/5)v/v_w$; e.g., 6 passengers if $v/v_w \sim 10$. The second equation at the outset of this module shows when this condition may be met but, clearly, this is a strategy for low demand systems. There are many unresolved practical issues with this type of system, but they would be worth some study.

Public Car-Sharing

We now consider a region of area R in which to provide car-share service, with demand density λ and expected trip time τ_s . Pods where users can pick up cars will be distributed through the region. The density of pods is Δ .



⁴ The argument here parallels the one just given for the access distance, after recognizing that when there are very many customers, consecutive customers have nearly the same y . With this in mind, we see that the average distance added by serving a customer near y is approximately equal to the average distance between two random points in a segment of length $2y$; i.e., $E(x|y) = 2y/3$. Thus, again, since the expected number of customers is proportional to y so that $E(y) = S/3$, we find: $E(x) = E(E(x|y)) = E(2y/3) = 2S/9$.

The fleet will have m cars, which are to be repositioned evenly after time h ; say one day. We need to make sure that there are enough cars at each pod so it is rare that one would run out. For simplicity of explanation, we assume no rush hour. There are three important costs for this system: fleet costs; vehicle repositioning costs; and user access costs.

We propose that the required fleet size is:

$$m \geq \lambda R \tau_s + 2(\lambda R \tau_s)^{\frac{1}{2}} + 2\sqrt{\frac{2\lambda h}{\Delta}} \Delta R$$

which can be broken down as: [the # of cars in use] + [95% CI (assuming Poisson arrivals) of cars in use] + [95% CI for net change of cars for a pod in a repositioning interval, multiplied by the number of pods]. The first two terms are a probabilistic upper bound to the number of cars in use. As such, these terms express the required fleet size for an ideal best case scenario with $h = 0$; i.e., where cars could be constantly and instantaneously reassigned across pods. (This is a logical conclusion because with $h = 0$ a pod would run out of cars only if all pods did; i.e. if the demand for cars in circulation exceeded the total available.) The third term expresses the safety stock that pods must collectively have to compensate for the fact that they are only rebalanced every h time units. This third term is the product of the number of pods and twice the standard deviation of the difference between cars requested and cars returned at a pod in a rebalancing interval. (This is why the demand rate is multiplied by 2 inside the square root.⁵) So if pods are given this extra safety stock the expression can be simplified to:

$$m \geq \lambda R \tau_s \left(1 + \frac{2}{\sqrt{\lambda R \tau_s}} + \frac{2.8}{\sqrt{\frac{\lambda \tau_s}{\Delta}}} \sqrt{\frac{h}{\tau_s}} \right)$$

The repositioning costs in one h will be the product of the number of vehicle moves per pod, which is a number close to the standard deviation of the net change in a pod; i.e.:

$$\sigma \sim \sqrt{\frac{2\lambda h}{\Delta}}$$

the number of pods ΔR , and the distance to reposition the cars (a problem famously known as “the transportation problem of linear programming” (TLP)). Research shows (see references) that for this type of problem:

⁵ This is quite accurate if we assume that customers can return cars to any pod in the system, and somewhat conservative if we assume that cars must be returned to the pod from which they were checked out—as existing systems operate.

$$\text{repo dist} \approx \sqrt{\frac{1}{\Delta}}(1 + 0.078 \log(\Delta R)) \approx \frac{1.1}{\sqrt{\Delta}}.$$

The last approximation is reasonable unless the number of pods is huge.

Thus, the repositioning costs per unit time will be:

$$c_d \left(\sqrt{\frac{2\lambda h}{\Delta}} \right) \Delta R \left(\frac{1.1}{\sqrt{\Delta}} \right) \left(\frac{1}{h} \right) \approx 1.5 c_d \sqrt{\lambda h} \frac{1}{\Delta} R.$$

The access (LOS) costs are proportional to:

$$\lambda R \Delta^{-\frac{1}{2}}.$$

Therefore, in summary we see that our cost components depend on h and Δ as follows:

- Fleet costs $\sim (h\Delta)^{1/2}$
- Vehicle repositioning costs $\sim h^{-1/2}$
- User access costs $\sim \Delta^{-1/2}$

The sum of these terms (appropriately weighted) is a GEOQ expression that has a unique h^* and Δ^* and would yield the least possible cost of a car-sharing operation. Test yourselves and see if you can do it. Notice how car-sharing beats taxi for large λ .

References

Daganzo, C.F. (1977) "An approximate analytic model of many-to-many demand responsive transportation systems," *Transportation Research* **12**(5), 325-333. (Average-case analysis of many-to-many DAR system; strategy II of that reference is the one described in this module.)

Daganzo, C.F., Hendrickson, C.T. and Wilson, N.H.M. (1977) "An approximate analytic model of many-to-one demand responsive transportation systems," *Proc. 7th Int. Symp. on the Theory of Traffic Flow and Transportation*, pp. 743-772, Kyoto, Japan. (Average-case analysis of many-to-one DAR system; the strategies in this reference are special cases of the one in the homework.)

Daganzo, C.F. (1984) "Check-Point Dial-a-Ride Systems," *Transportation Research*, **18B**, 315-327. (The system analyzed in this reference is a building block toward the 2-mode DAR described in this module).

Daganzo, C.F. and Smilowitz, K.S. (2004) "Bounds and approximations for the transportation problem of linear programming and other scalable network problems" *Transportation Science* **38**(3), 343-356 (Derives the TLP formula we used in connection with the repositioning costs for car-sharing systems).

Appendix: Determination of Expected Distance to a Taxi

We start with the probability of zero taxis within a disc of radius x :

$$\Pr\{d_n \geq x\} \equiv \Pr\{\text{zero points in the disc}\} = \left(1 - \frac{\pi x^2}{R}\right)^{n_i}.$$

By integrating this over the range of possible radii, we can calculate the expected distance from a point to the closest taxi as:

$$E(d_{n_i}) = \int_0^{\sqrt{\frac{R}{\pi}}} \left(1 - \frac{\pi x^2}{R}\right)^n dx.$$

With a change of variable $y = \sqrt{\frac{\pi}{R}}x$, the integral can be changed to a simpler form:

$$\int_0^1 (1 - y^2)^n \sqrt{\frac{R}{\pi}} dy.$$

With another change of variable $y = \cos \theta$, and using the identity $(1 - \cos^2 \theta) = \sin^2 \theta$, the integral can be further simplified to a known form:

$$\sqrt{\frac{R}{\pi}} \int_0^{\frac{\pi}{2}} (\sin \theta)^{2n+1} d\theta = \sqrt{\frac{R}{\pi}} \frac{\sqrt{\pi}}{2} \frac{\Gamma(n)}{\Gamma(n + \frac{1}{2})}.$$

For values of $n \gg 3$, a reasonable assumption for a region that would employ a fleet of taxis, we can use the approximation:

$$\frac{\Gamma(n)}{\Gamma(n + \frac{1}{2})} \cong \frac{1}{\sqrt{n}},$$

which results in an answer of:

$$E(d_{n_i}) \cong \frac{1}{2} \sqrt{\frac{R}{n}}.$$

For a region with where vehicles move along a rectilinear grid, the same method can be followed with a couple of substitutions:

$$E(d_{n_i}) = \int_0^{\sqrt{\frac{R}{2}}} \left(1 - \frac{2x^2}{R}\right)^n dx$$

$$y = \sqrt{\frac{2}{R}}x$$

With a final result of:

$$E(d_{n_i}) \cong \frac{\pi}{2\sqrt{2}} \sqrt{\frac{R}{n}} \cong 0.63 \sqrt{\frac{R}{n}}$$

Module 6: Management—Vehicle Fleets

(Originally compiled by Eric Gonzales and Josh Pilachowski, April, 2008)

(Last updated 9-22-2010)

Outline

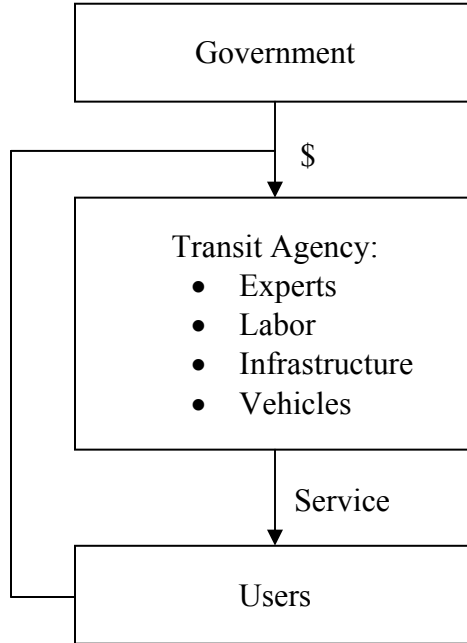
- Introduction
- Schedule Covering 1 Bus Route
 - Fleet Size: Graphical and Numerical Analyses
 - Determination of Terminus Locations and Bus Runs
- Schedule Covering N Bus Routes
 - Single Terminus Close to a Depot
 - Multiple Termini and Deadheading Heuristics
- Discussion: Effect of Deadheading
- Appendix: The Vehicle Routing Problem and Meta-Heuristic Solution Methods

The block diagram below illustrates that a transit agency is in essence a mechanism (the middle box) for transforming money inputs from both, government and users into transportation service. In this course we focus on the workings of this mechanism, treating the arrows pointing into the middle box as constraints and those pointing out as the objective function. We have just finished a set of (planning) modules that explore the ideal structure of this mechanism, focusing on the long term.

We are now about to start a set of modules that will explore what needs to be done in the short and medium term to execute the long term plans. This involves medium-term investment and deployment decisions of the transit agency's manageable resources, which mainly consist of vehicles and personnel. Invisible to the public, we call these actions "management decisions". This module deals with vehicle fleet management. Module 7 will deal with personnel management. A transit agency also needs to make other medium-term and short-term operational decisions that are visible to the public. Module 8 will deal with these.

Although our attention will continue to be focused in the middle box we should not lose sight that it is only part of the whole picture. A transit agency is also concerned with the arrows. The issues of finance and governance (inbound arrows) and public relations and information dissemination (outbound arrows) are of much importance to the success of a transit operation. These issues, however, are not transit-specific and will therefore not be addressed in these notes. So let us now return to the inside of the box, with a focus on management.

Public Transportation Systems: Management—Vehicle Fleets



Introduction

To this point, in the planning part of the course, we have assumed that agency costs (including operating costs and amortized capital costs) are proportional to the vehicle hours and vehicle kilometers of service provided. This would be exact if vehicles could be rented for only the time that they are needed for use in service, and drivers could be hired and fired so that people only worked the hours that buses are in operation. In reality, vehicles are purchased or leased for more than a few hours at a time and labor unions place restrictions on the number of hours that drivers can work. In this and the next module we will develop vehicle operating plans and driver staffing plans recognizing these limitations. Homework exercises will compare these more realistic operating costs experienced by the agency and those assumed in the planning stage. We will find that the assumptions made during the planning part of the course were not bad approximations.

Definitions

These definitions will be used in the two management Modules.

Schedule – set of routes and scheduled services advertised by the transit agency.

Depot – location where buses are stored without drivers.

Public Transportation Systems: Management—Vehicle Fleets

Run – time-space path of one specific transit vehicle from and returning to a depot. The vehicle needs a driver during the entire run. The run may include coverage of more than one transit line.

Terminus – part of a transit line (i.e., route) where buses can be changed.

Loop – part of the run between consecutive visits to a route's terminus; it must be covered by the same bus.

Driver Task – indivisible part of a run that must be covered by the same driver—to be specified later.

Job – set of tasks covered by one specific worker in a single day.

Worker Type – common work pattern characterized by pay rate and properties of their shift.

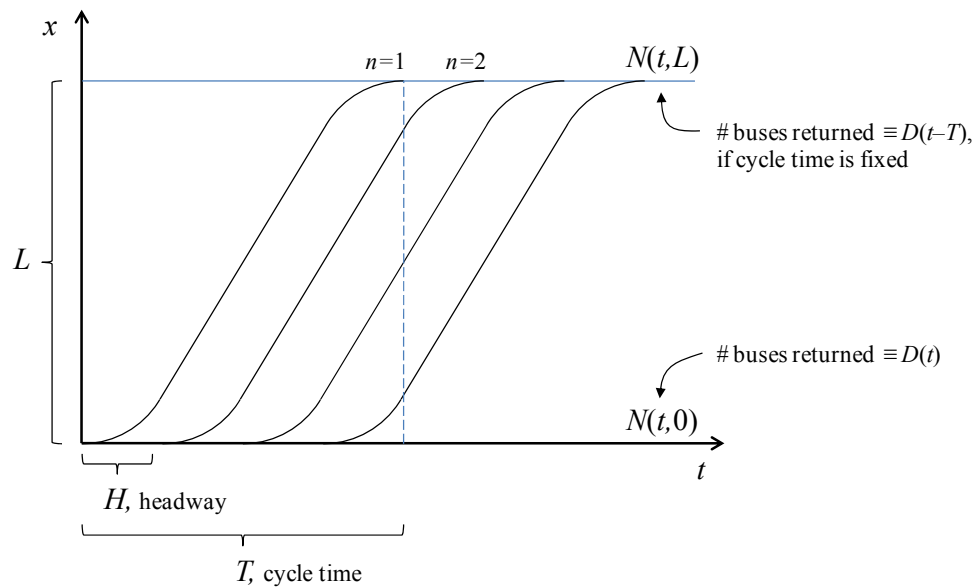
Allocating vehicles and drivers to provide the schedule promised by the transit agency is a two step process:

- 1) Find a fleet operating plan to cover the schedule \rightarrow {# of vehicles, specific runs}. To do this, some vehicles may have to sit unused for part of the time.
- 2) Find a staffing plan to cover the runs \rightarrow {# of workers by type, specific jobs}. This involves cutting the given runs into tasks and then allocating workers to cover the tasks.

These steps are parallel in structure. Both answer the question: how many items are required to cover a set of requirements? This Module is concerned with step 1.

Schedule covering: 1 bus route

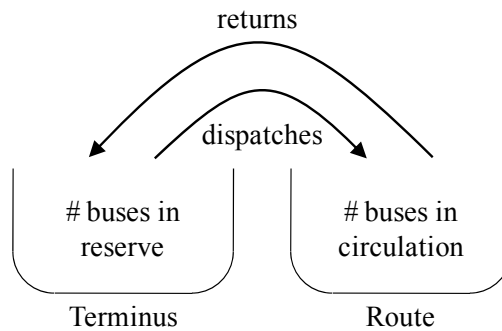
The data for this problem (the schedule) can be represented in a time space diagram showing each of the buses traveling along a route from a terminus (at $x = 0$) and looping back to the terminus. Each bus requires a cycle time T to make a full loop of length L and return to the terminal. Then, $N(t, 0)$ is the cumulative number of dispatches from the origin over time (also denoted $D(t)$), and $N(t, L)$ is the cumulative number of returns, which is $D(t-T)$ if the cycle time is fixed.



Fleet Size: Graphical Analysis

We analyze this system as a queuing system from the perspective of the terminus—imagining for the time being that the depot is on top of the terminus and that the depot supplies buses when needed. What is the minimum number of buses needed to sustain the schedule?

Each bus can be classified as waiting in reserve at the terminal until dispatch or circulating in service. The transitions between reserve and circulation are the dispatches and returns.



A cumulative plot of buses available $A(t)$, buses dispatched $D(t)$, and returned $R(t)$ shows graphically how the number of buses in reserve and service evolves over time. See below. The curve $D(t)$ is given and the other two are derived. The number of available buses is equal to those initially available, M , and those returned: $A(t) = M + D(t - T)$. Note how for this closed queuing system the sum of reserve and circulating buses is the total number of vehicles, M , which remains constant over the course of the day. Note: curve A is obtained from curve R with a vector shift (T, M) .

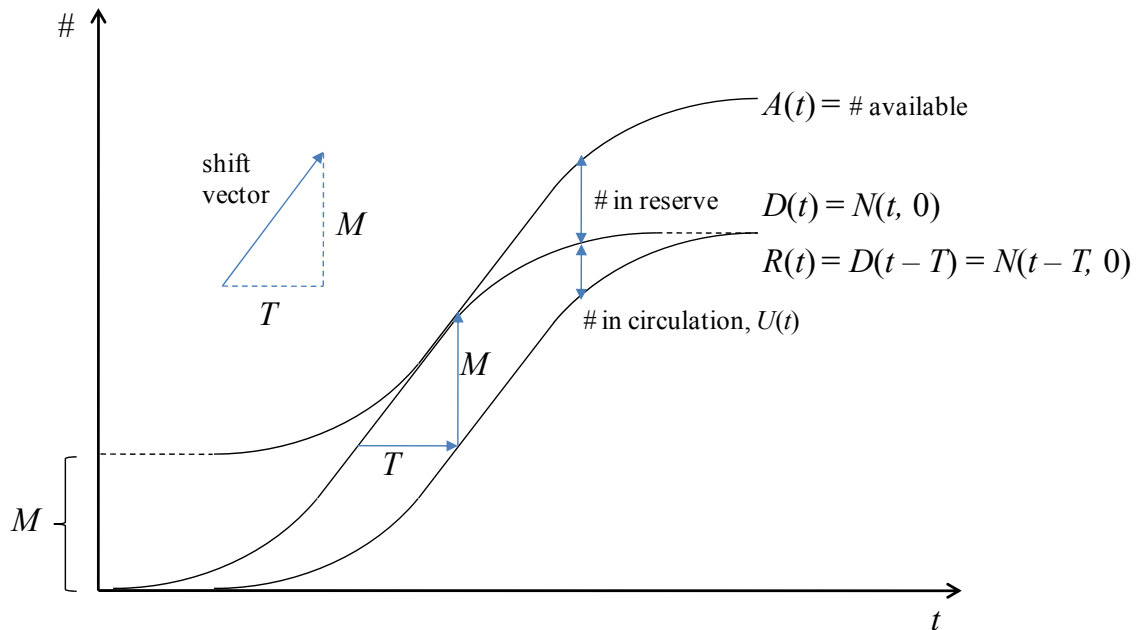
Since the number of buses in reserve has to be positive, we require: $A(t) \geq D(t)$; so the minimum fleet size is obtained when $A(t)$ and $D(t)$ are tangent, as shown.

Note: The tangency point is where the cumulative dispatch and cumulative return curves are maximally separated; i.e., where the number of circulating buses, $U(t)$ is maximum. So, we have:

$$M = \max\{U(t)\}.$$

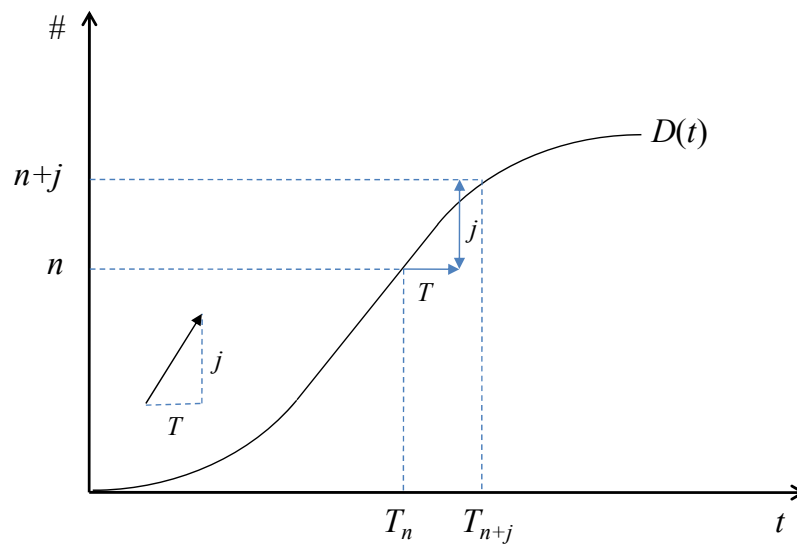
As an exercise, you can prove that this formula reduces in the time-independent case to the result we know:

$$D(t) = \left\lceil \frac{t}{H} \right\rceil \Rightarrow M = \left\lceil \frac{T}{H} \right\rceil.$$



Fleet Size: Numerical Analysis

Given a cumulative dispatch curve $D(t)$ with (tentative) fleet size j , we see from the picture below that if $T_{n+j} - T_n \geq T$ for all n , then $A(t)$ is always to the left of $D(t)$ for all t . The condition is a precedence condition ensuring that every bus is available before it is dispatched.

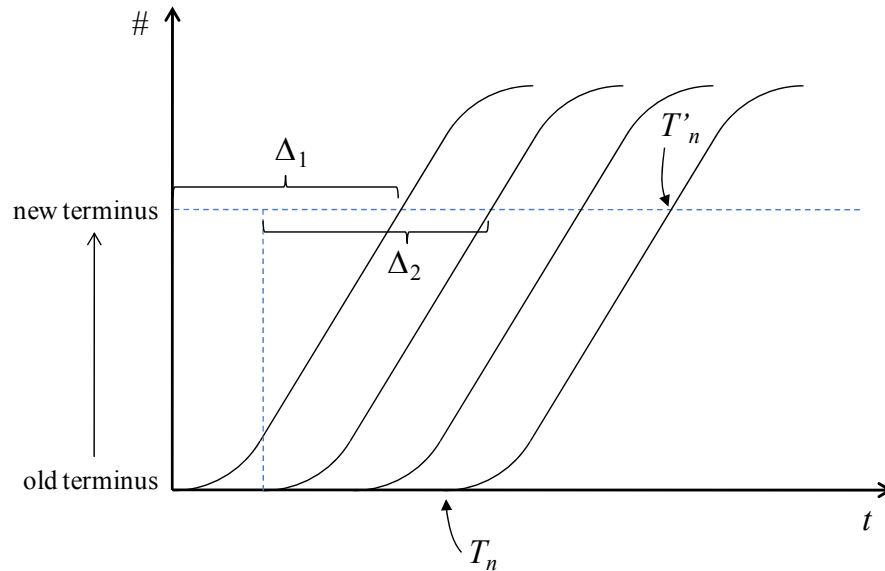


The minimum fleet size can be easily determined with a spreadsheet that checks the precedence condition for different tentative fleet sizes, j . The lowest value of j corresponding to column with all values greater than or equal to 0 is the minimum fleet size which ensures that the reserve of buses is never empty.

		$j = 1$	$j = 2$	
n	T_n	$T_{n+1} - T_n - T \geq 0$	$T_{n+2} - T_n - T \geq 0$...
0	time data			
1	time data			
2	time data			
\vdots	\vdots			

Terminus Location

So far, we assumed that the terminus was at $x = 0$. Could it be possible to reduce the number of buses by locating the terminus in a different place along the route? The answer is no if bus trajectories are the same through the day. Here is why: Let Δ_n be the travel time of bus n from the old to the new terminus (see figure) and note: $T'_n = T_n + \Delta_n$. Now, if $\Delta_n = \Delta$, then $T'_{n+j} - T'_n = T_{n+j} - T_n \geq T$ because Δ cancels out and we can put the terminus wherever we want without a penalty. This is good because it gives us the feasibility to put the termini at favorable locations (e.g. where buses are nearly empty).

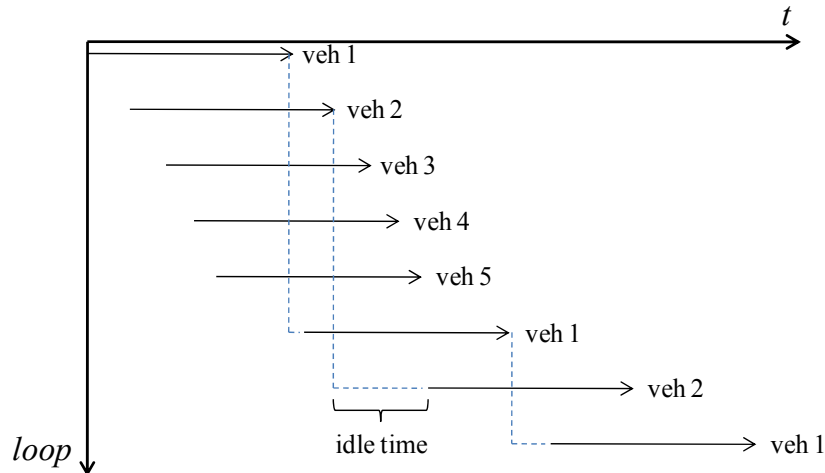


Run Determination

The bus scheduling problem is to determine which bus is associated with each loop; i.e. figuring out what each bus will do: the bus runs. We discuss two heuristic methods for determining the specific run for each vehicle, but there are many more. In method 1 we use the last-in-first-out (LIFO) strategy; new buses are only introduced when absolutely necessary. This method steps through time, so when a loop is scheduled to start the most recent bus that returned is dispatched. If there is no such bus a bus is selected from the initial pool. This strategy is good because it keeps some individual buses running while others experience long periods of idling. The latter can be returned to the depot for driver relief.

Public Transportation Systems: Management—Vehicle Fleets

An alternative strategy with the same goal is a greedy strategy that steps through buses, assigning to each bus as many loops as possible. After each bus, the loops covered by the bus are removed and the next bus again covers as many loops as possible. To do this, each bus is redeployed as early as possible after returning to the terminus. This can be performed graphically by hand by plotting each scheduled loop from the route's terminus against time as shown below. Could you organize this in a spreadsheet?

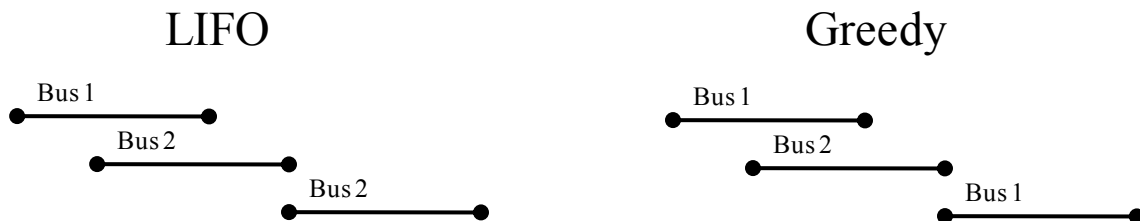


It is perhaps intuitive, and can be proven, that both the LIFO and greedy methods are feasible with the minimum fleet size we calculated earlier:

$$M = \max\{U(t)\}$$

Example: Qualitative difference between LIFO and Greedy Methods

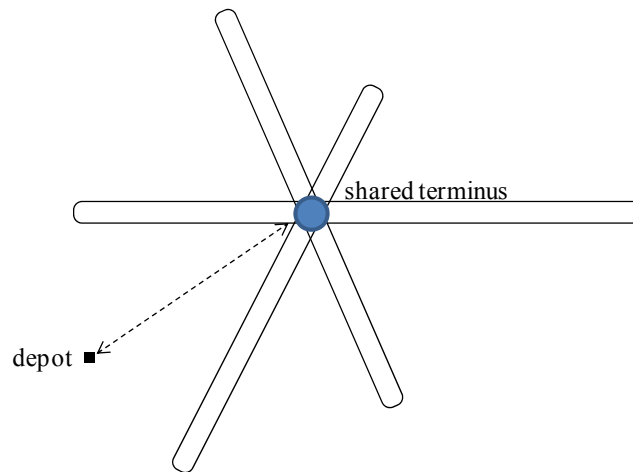
The simple three-run, two-bus system below shows why the two methods differ:



The LIFO method assigns the third loop to the bus returning most recently (Bus 2). The Greedy method assigns it to the lowest indexed available bus (Bus 1). The methods only differ in the bus that is selected from the pool of idle buses for the next run. Since we are assuming that all buses are identical, the choice has no future repercussions on the availability of buses. In fact, one could have selected the bus at random, or with any other rule, and the strategy would perform similarly. Thus, the specific bus choice can be made with other (non-bus) criteria in mind.

Schedule covering N bus routes

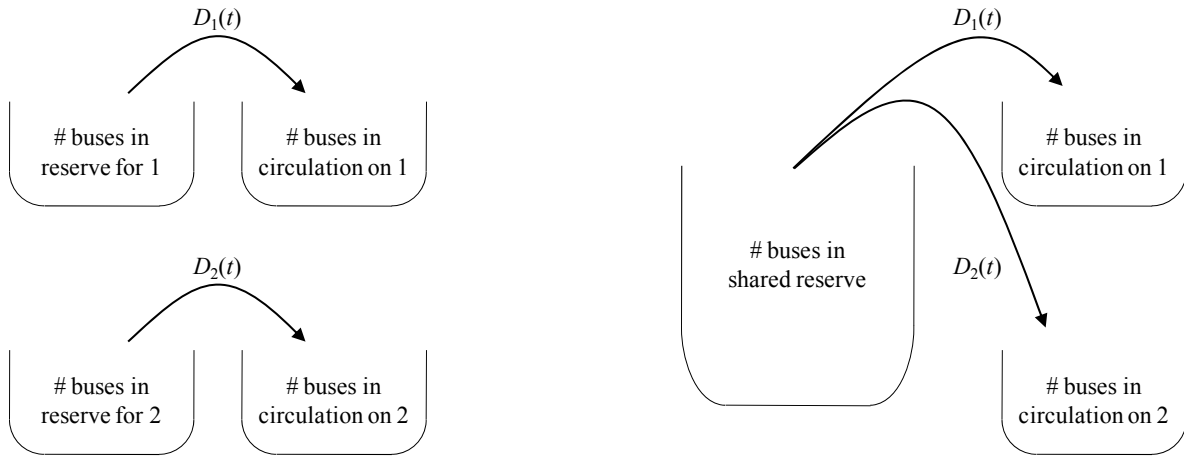
Imagine a map of many routes all passing converging at a centrally located terminus. We could imagine this is a bus station in the center of a city or at a busy rail transfer station. The terminus may be close or far from the depot. We imagine for now that it is close.



Single Terminus Close to the Depot

We could treat each route independently as before, but on the other hand, it may be possible to reduce the fleet size by sharing buses between routes. On the left below is the model for dedicating bus fleets to separate routes in isolation. On the right, this model is modified so that rather than a reserve of buses for each route, the terminus holds a reserve of available vehicles for all routes. Each route, i , is characterized by loops of different cycle times, $T(i)$, so the time until a dispatched bus returns is no longer uniform but depends on which route the bus has been dispatched.

Public Transportation Systems: Management—Vehicle Fleets



The aggregated cumulative count of dispatched and returned vehicles is now expressed as

$$D(t) = \sum_i D_i(t)$$

$$R(t) = \sum_i D_i(t - T_i)$$

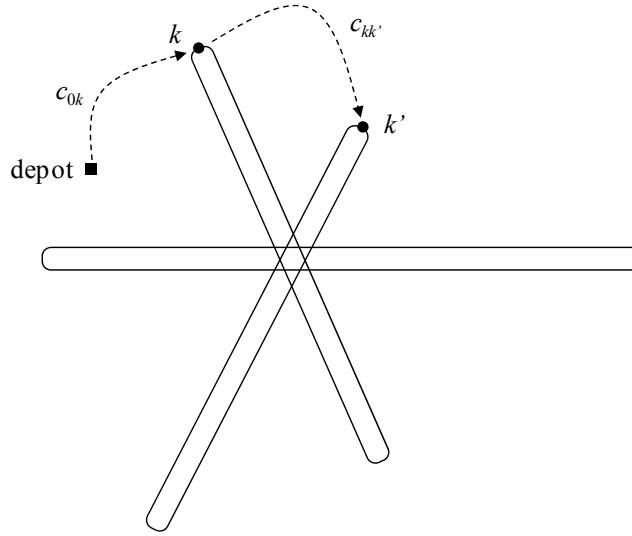
And since the fleet (M) is shared, the cumulative number of buses made available for collective use is still:

$$A(t) = M + R(t)$$

Curves $D(t)$, $R(t)$, and $A(t)$ can be plotted as before to determine the minimum fleet size, and the formula $M = \max\{D(t) - R(t)\}$ continues to hold. The only difference is that $R(t)$ is no longer related to $D(t)$ by a shift.

Dispersed Termini and Deadheading Heuristics

Consider now the case where the termini and depot are dispersed. Perhaps the termini are at ends of the lines, and there may be some cost, $c_{kk'}$, of moving a bus from loop k to loop k' . To include deadheading from the depot, we use $k = 0$ for the depot and $k = 1, 2, \dots$ for the loops.



A simple heuristic method can be used to solve this problem approximately. This method is good if $c_{kk'} \ll T_{(k)}$. Otherwise, it produces solutions that may need improvement. We imagine that all buses on route k are requested a time

$$\Delta_{(k)} \geq \max_{k'} \{c_{kk'}\}$$

ahead of their real dispatch time, recognizing that they could be coming from any other terminus. If we build this slack into the schedule, i.e. we define:

$$T'_{(k)} = T_{(k)} + \Delta_{(k)}$$

We can treat this new problem (with $T'_{(k)}$) as previously (ignoring deadheading). This is a way to obtain a tentative fleet size and set of bus runs which can be improved using a computer.

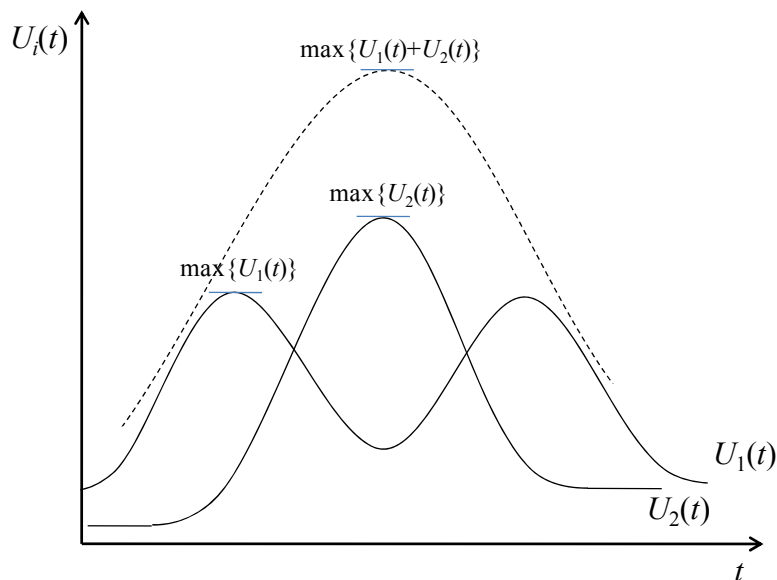
Fortunately, the problem we are solving is analogous to the vehicle routing problem (VRP); a famous problem that has been extensively studied. So, we don't have to do this from scratch. (The appendix gives some background on the VRP and a simple computer method that can be used to improve tentative solutions.) The VRP is analogous to the schedule covering problem that we want to solve because we are looking for the least costly way to cover a set of requirements. The analogy is presented in the table below.

The penalty can be defined by any function that maps idle time between loops to a penalty. This may be a function that increases as the idle time wastes money until some point when the bus can be returned to the depot.

<i>Vehicle Routing Problem (VRP)</i>	<i>Schedule Covering Problem</i>
points i, j	loops k, k'
distance, c_{ij}	penalty, $p_{kk'}$ $\begin{cases} \infty & \text{if impossible} \\ 0 & \text{if } c_{kk'} = 0 \\ > 0 & \text{if feasible, but } c_{kk'} > 0 \end{cases}$
vehicle	Bus
vehicle load	loops covered by a bus; i.e., the bus run
capacity	∞

Discussion: Effects of Deadheading

To illustrate the potential benefits of deadheading, suppose we have two bus lines with different peaking patterns, such as a commuter route running heavily in the morning and evening, paired with a route that is run most heavily during the middle of the day for something like an athletic event. The figure below displays these patterns by means of two solid curves. The dotted line (not drawn to scale) is the sum of these curves.



Compare the fleet requirement if the routes were considered separately (the sum of the maximum route requirements considering each route individually) to the fleet requirement if the two routes

share resources even if deadheading is required (the maximum of the sum of route requirements given by the dotted curve). It always happens that:

$$\sum_i \max_t \{U_i(t)\} \geq \max_t \sum_i \{U_i(t)\}$$

So, some savings are possible with deadheading, but these are offset against the cost of deadheading itself. The greatest benefit is from routes that peak at different times.

Appendix: Introduction to the “Vehicle Routing Problem” and Meta-Heuristic Solution Methods.

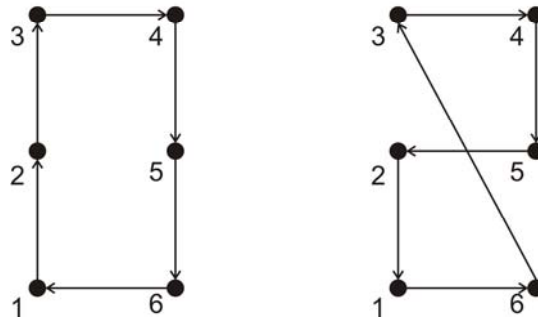
Here we describe some combinatorial optimization concepts that are useful for scheduling public transportation workers and vehicles.

Local Search Methods and Meta-Heuristics

The basic idea behind local search methods is to guess solutions that get increasingly better as the procedure develops. Solutions are characterized by a “state” which is a string of numbers. This can be illustrated with the TSP. Given are N points (or cities), $i = 1, 2, \dots, N$, and a matrix of distances $\{e_{ij}\}$ between every pair of points. In the TSP, we look for a tour that visits all the points with the least total distance (or “cost”). Since the positions in which the cities appear in a tour are uniquely defined by an ordering of the first N integers (a permutation), any such ordering is a state of the TSP problem. Example 3 below shows 2 possible states for a 6 point TSP problem: (1, 2, 3, 4, 5, 6) and (1, 6, 3, 4, 5, 2). It is assumed in this example that costs are given by the Euclidean distances of the links. Thus, the cost of each state is the length of the tour one would measure with a ruler.

Any “local search” is based on perturbations that transform a state into a similar state, hopefully with lesser cost. For the TSP, a perturbation could be choosing 2 consecutive cities and swapping their order. For example, from (1, 2, 3, 4, 5, 6) we could go to (1, 3, 2, 4, 5, 6,) and from this to (1, 3, 4, 2, 5, 6). The set of states that can be reached in one step (one perturbation) is the state’s local neighborhood. Perturbations should be simple (so they are easy to make and evaluate), but also comprehensive, in the sense that they should allow the system to reach any state from any other state. Consecutive city swaps have these two properties and are therefore acceptable perturbations for the TSP.

Example 1



Given a current state, a “greedy” local search would evaluate the cost of all the states in its neighborhood and move to the one with the least cost if such a state exists; otherwise the search ends. This procedure is then repeated using this new state as the current state, and then repeated iteratively until the search ends because no improvement can be found. The termination point is called a “local” optimum. Local optima are generally not unique for the TSP. For example, you can verify that the two tours of Example 3 are locally optimal, even though tour (1, 6, 3, 4, 5, 2) on the right is quite bad.

In view of this, people have created “meta-heuristic” methods that in theory can avoid being trapped in local optima and converge to the global optimum. The simplest meta-heuristic method is called simulated annealing (SA). It differs from the greedy method in that it randomly chooses a single perturbation from the current state to identify a single new state. A coin is then flipped to see whether the new state is accepted and becomes the new current state, or one stays put. The probability of success “ p ” is chosen to be the following function of the change in cost, Δe , and the iteration number, n : $p = 1$, if $\Delta e \leq 0$; but if $\Delta e > 0$ then $p = \exp\{-\Delta e/(n+a)\}$, where “ a ” is a positive constant. Note that at the start of the search ($n = 1$) there can be a significant probability of accepting a more costly state (with $\Delta e > 0$) but this probability declines as the simulation progresses. This probabilistic feature of SA allows the algorithm to jump out of local optima and, given enough time, to converge to the global optimum. Unfortunately convergence is slow for problems with more than (say) 100 points. Even in these cases, though, the method can be used to fine tune solutions obtained with other methods. A large value of “ a ” is normally chosen for this type of application.

The Vehicle Routing Problem (VRP)

The VRP arises in practice more often than the TSP, and many variants of it exist (e.g. with route length restrictions, time-windows, etc.). In its most basic forms it seeks vehicle routes to serve a

set of N customers distributed in space. Customers have items to be carried, which take up vehicle space. Vehicles have finite capacity.

Given are:

N points, $i = 1, 2 \dots N$

M vehicles, $m = 1, 2 \dots M$

A depot at $i = 0$

A matrix of distances, e_{ij}

A demand d_i for every point (city) (in units of “quantity”)

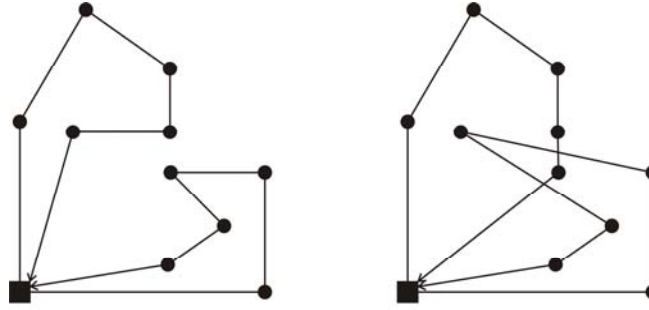
A vehicle capacity, V_m for every vehicle (also in units of quantity)

We look for:

An allocation of points to vehicles and a set of vehicle routes ending and beginning at the depot that minimizes either vehicle distance, number of vehicles or a combination of the two.

The VRP can also be attacked with meta-heuristics such as simulation annealing (SA), and these techniques still give reasonable results for problems with up to (about) 100 points. Instead of a single permutation, a “state” now consists of an ordered allocation of cities to vehicles. Note, some of these states may be infeasible--if the total demand for vehicle m exceeds V_m .

The SA algorithm would work as before. One defines perturbations, which can be swaps of points (also called “customers”) within a tour, or swaps of groups of customers among tours. Example 1 shows the result of swapping the last customer of the tour on the left with the middle customer of the tour on the right. It should be clear that any state whatsoever can be reached from any other state if one uses a proper sequence of swaps. Therefore, the SA approach with random swaps should (theoretically) work. In practice, experience with the VRP has been good with problems as large as ~100 points. For larger problems SA can be used as a fine-tuning tool with a large value of its parameter “ α ”. A demonstration of this approach can be found in Robuste, et al. (1990), which applied the SA annealing algorithm to a problem with about 200 points).



Example 2. Tours Before and After a Swap of Points Between Tour Tours

As is explained in the text, many transit problems can be cast in the form of a VRP-like problem that can be solved or fine-tuned with SA. This technique can be quickly mastered and applied. The case study in Robuste et al (1990) took less than 1 week from conception to completion.

More Information:

The following elementary readings could be of use. Section 10.9 of “Numerical Recipes: The Art of Scientific Computing” by W. Press et al., Cambridge 1987, pp. 326-334, describes simulated annealing in the context of the Traveling Salesman Problem (TSP), and shows some computer code. A short description can also be found in Appendix B of Daganzo (2005), *Logistics Systems Analysis*, Springer. Section 4.5.2 of this reference (*Fine-tuning Possibilities*) summarizes the case study in Robuste et al, (1990).

Module 7: Management—Staffing

(Originally compiled by Eric Gonzales and Josh Pilachowski, May, 2008)

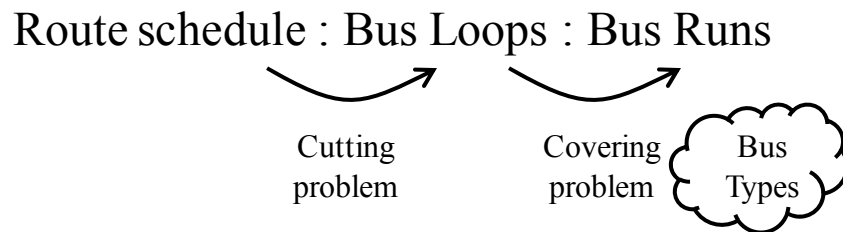
(Last updated 9-8-2010)

Outline

- Recap
- Staffing a Single Run
 - Effect of Overtime
 - Effect of Multiple Worker Types
- Staffing Multiple Runs
 - Run-Cutting
 - Covering
- Choosing Worker-Types
- Dealing with Absenteeism
- What is Still Left to be Done

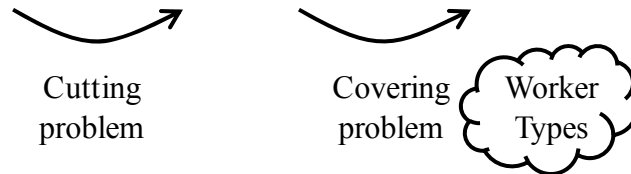
Recap

Recall from last lecture the 2-step process we used to cover a schedule:



We start with a route schedule and cut it into loops that can be covered by the buses. Buses, categorized by speed, capacity, etc... are then assigned so that each loop is covered. This can be solved as a Vehicle Routing Problem (VRP). The solution consists of all the runs for each bus. If buses were automated vehicles this would be the end of the problem. However, since they are not, we must figure how to cover the bus runs with drivers. To do this we consider the drivers and the constraints they add, using the following sequence:

Bus Runs : Driver Tasks : Driver Jobs

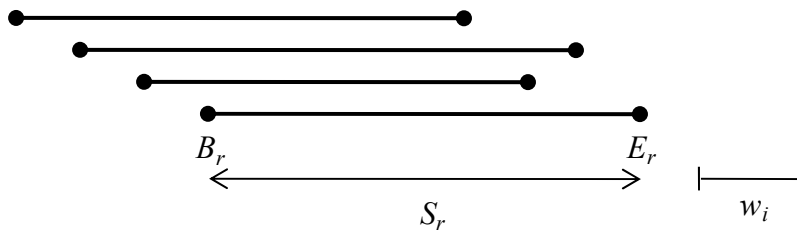


Notice the similarities between the two formulations. The bus runs obtained from the fleet scheduling step are now cut into elementary driver tasks and drivers, categorized by shift length and wage rate, are then assigned to feasible sets of tasks (jobs) so that each task is covered. This is what the focus of this lecture will be.

Because of the similarities between the schedule covering and staffing problems we can analyze them in a similar way: first by considering each run separately and then by pooling them, allowing drivers to cover multiple runs.

Staffing a single run

The result of the schedule-covering analysis from last lecture is a series of runs, r , each with a beginning time, B_r , an ending time, E_r , and a duration S_r . We also take as given the (continuous) work interval for workers of type i , w_i . Normally, $w_i < S_r$ for many runs.

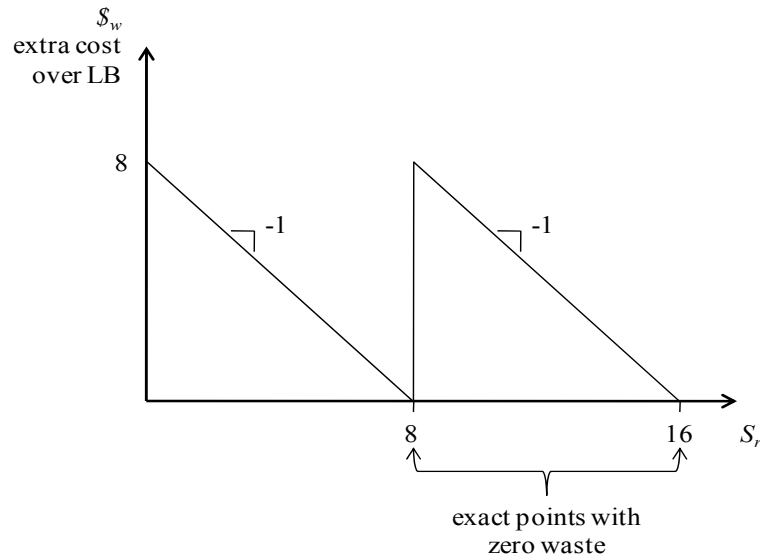


We also assume that we know the wage rate of drivers of type i , as well as the premium added to their wage rate for working overtime. To simplify the formulas we use monetary units so that the wage is 1. In those units, we define the overtime premium as π , and the overtime wage as $(1+\pi)$.

Consider now the extra cost of wages over the lower bound (LB) obtained by assuming that $\pi = 0$ (i.e. that drivers can be hired and paid only when needed). We will examine how this extra cost depends on the types of shifts that are used - assuming that the transit agency has the flexibility

Public Transportation Systems: Management--Staffing

to ask workers to start their shifts at any time. First we look at a single run of length S_r and set $w = 8$ hrs and $\pi = \infty$ (no overtime offered). Then the extra cost over the LB, $\$w$, is described by the following curve, which is our base case; see figure below.



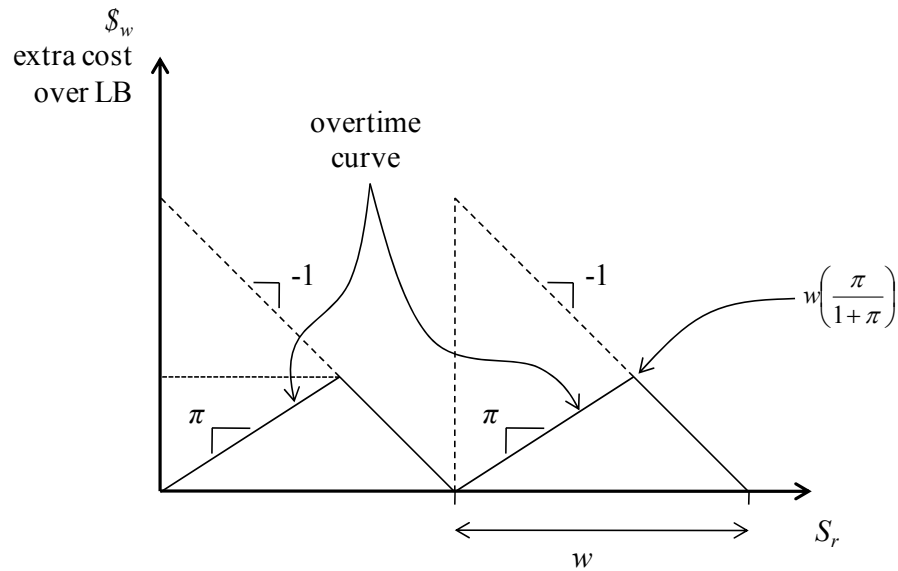
You can see that the largest cost, w , is paid for runs with lengths slightly longer than a multiple of w . Over multiple runs of random duration, the average wasted cost should be about $w/2$. We waste about one half of a shift per run.

Effect of Overtime

By allowing for overtime (and hiring short-term drivers at the overtime rate), a new extra cost curve with slope π becomes possible, see figure below. The least cost is then the minimum of the two curves. One would use overtime only when the overtime curve is beneath the regular curve (i.e., where the latter is dotted.)

Simple algebra reveals that the maximum cost is now: $w \left(\frac{\pi}{1 + \pi} \right)$. Thus, the excess (wasted) cost

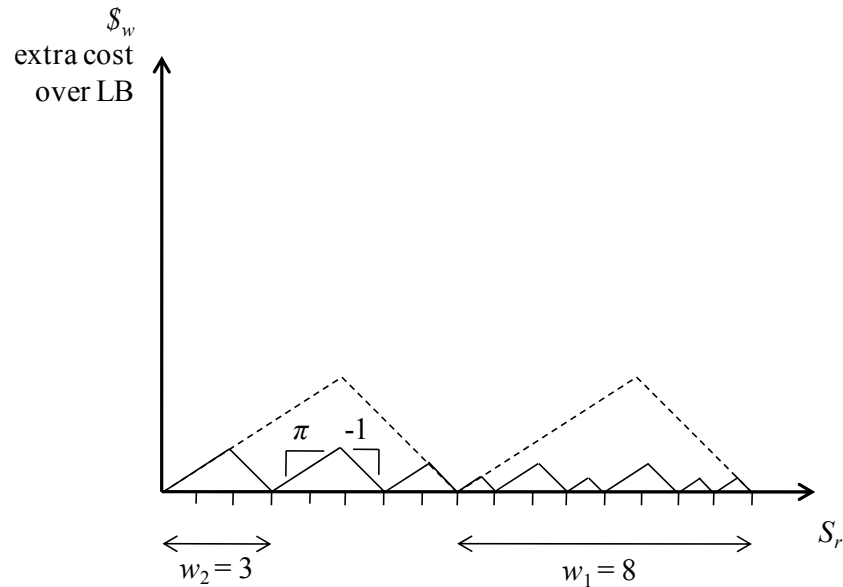
per run should be about $\frac{w}{2} \left(\frac{\pi}{1 + \pi} \right)$ on average. Note how overtime reduces waste.



Effect of Multiple Worker Types

If we introduce more worker types to the system we can further reduce our cost. For example, by offering a shorter shift, $w_2 = 3$ hrs, we can find exact points of zero extra cost for runs with the following lengths: 3, $6 = (3 + 3)$, 8, $9 = (3 + 3 + 3)$, $11 = (8 + 3)$, $12 = (3 \times 4)$, $14 = (8 + 3 + 3)$, $15 = (3 \times 5)$... (continuing for all integer values greater than 15). You can see from the figure below that the resulting extra should be about 0.5 without overtime and about 0.25 if overtime with $\pi \approx 1$ is allowed.

If we know the cumulative distribution, $F(x)$, of all runs, the extra expected cost across all runs can be expressed more precisely as: $E(\$_w) = \int_0^{24} \$_w(x) dF(x)$.

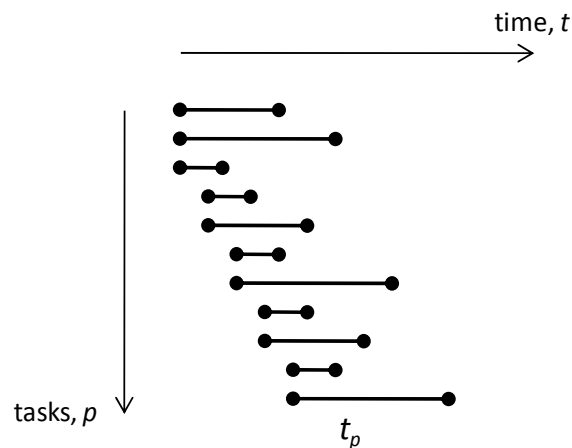


Staffing Multiple Runs

As in the case of the fleet scheduling problem, the situation can still be improved by pooling: considering all runs together and allowing drivers to serve more than one run.

Run-Cutting

To do this, we first cut the runs into elementary tasks, p , that can be covered by different drivers - although each task must be done by the same driver. These tasks should be as short as possible, recognizing the practical constraints that apply to the agency. (Perhaps, for example, drivers can only be switched at certain points on the routes.) The results of the cutting process can be expressed graphically as in the fleet scheduling problem:



Covering

Data for the problem would include the task duration, t_p , the times, $t_{pp'}$, between the end of tasks p and the beginning of p' , and the real cost of moving a driver from task p to task p' , $c_{pp'}$ (this cost is set to ∞ whenever the move is infeasible; e.g., for all moves where $t_{pp'} < 0$).

This covering problem can also be formulated as a VRP, albeit a variant with different constraints. The analogy from the previous lecture is continued below:

<i>Vehicle Routing Problem (VRP)</i>	<i>Schedule Covering Problem</i>	<i>Staffing Problem</i>
depot	depot	worker home (or depot)
points i, j	loops k, k'	tasks p, p'
vehicle	bus	worker
vehicle load	loops covered by a bus; i.e., the bus run	tasks covered by a worker; i.e., a job
capacity	∞	∞
time for a stop, t_i	n/a	t_p
time between stops $t_{i,j}$	n/a	$t_{pp'}$
distance for a leg, $c_{i,j}$	$p_{kk'}$	$c_{pp'}$
time constraint, T	n/a	w_i

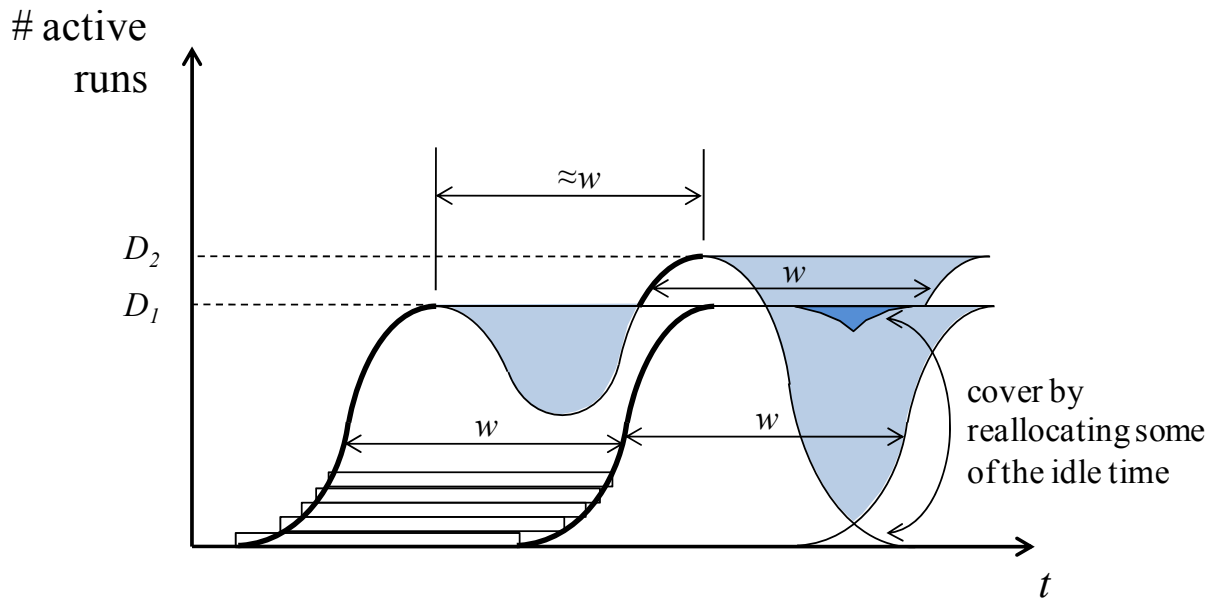
Despite the complications, the SA method can still be used. We would use the same “state” as in the previous lecture (i.e., ordered strings of numbers, p , for each worker), and would treat runs independently. We could also use the same set of perturbations. Only the cost evaluation step would be slightly different since violation of the time constraint would imply an infinite cost.

The SA approach can be used even if worker types are characterized by specific beginning and ending times of their shifts, and not only by their shift durations.

Simplified estimation of cost

We can use a graphical method in order to obtain a quick estimate for a LB of total cost. To do this we assume that workers can start their shifts at any time and can be reallocated across runs without a time penalty. This allows us to focus on the number of runs, ignoring specific runs and where they take place in space. By graphing the number of active runs over time, the problem

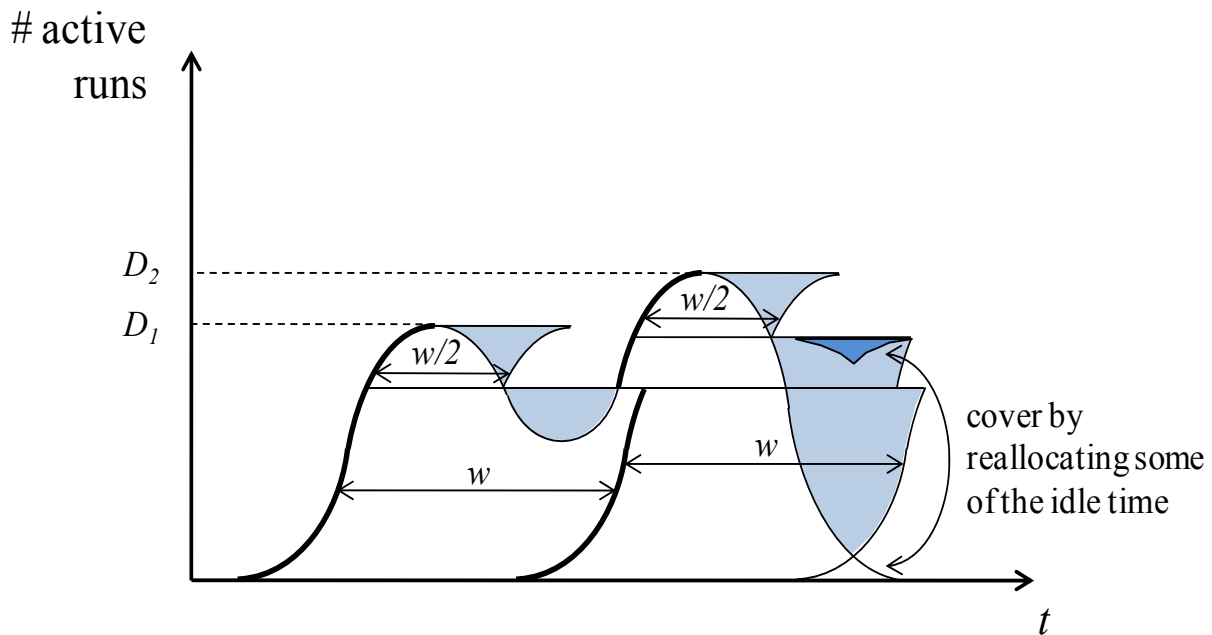
simply becomes one of covering the area under the graph with “strips” representing a work shift, with a height of one bus and a length, w .



In the above figure, D_1 is the number of active runs during the morning peak, and D_2 is the number of active runs during the afternoon peak. The least possible number of strips have been used to cover the area under the graph. The darker parts of the curve mark the beginning of the strips. The shaded portion of the graph represents the wasted time such that a driver is employed without having a bus to drive. The small portion at the end of the day can be covered by idle workers from the afternoon peak.

Note: The two peaks are separated by the length of a workday which of course is close to w . This presents a problem, since drivers who start their shifts at the beginning of the peaks must be idle most of their shift. From the picture you can see that there will be $(D_1 + D_2)w$ driver-hours hired.

By allowing some shorter shifts of duration ($w/2 = 4$) in addition to the regular shifts ($w = 8$) we could use the following solution:



This results in no driver being idle for more than 4 hours even though most of the shifts are of regular length. This is very interesting: a large cost reduction is achieved by hiring just a few 4-hr workers. If we continue this idea by allowing shorter shifts and overtime, the amount of waste can be reduced even more. The homework illustrates this effect.

Choosing Worker Types

We have shown that it is beneficial to have shorter shifts; however there is the question of how to induce workers to choose these shifts. There is the possibility of paying higher wages, but we can possibly provide an incentive other than money. You could offer a shift schedule such that a driver can work 9 hours a day for four days, and then have a 4-hour shift on their fifth day. By partitioning 4-hour shifts for each weekday over all the workers, each day would have a 4-hour shift for every four 9-hour shifts. The normal and 9/4 rotations are shown in the table below:

Public Transportation Systems: Management--Staffing

	M	T	W	Th	F
rot 1	8	8	8	8	8
rot 2.1	9	9	9	9	4
rot 2.2	9	9	9	4	9
rot 2.3	9	9	4	9	9
rot 2.4	9	4	9	9	9
rot 2.5	4	9	9	9	9

You could then allow drivers to choose their preferred rotation, in order of seniority, with the standard 8-hour shift being the default. This strategy has not been put into practice regarding transit staffing, however it exists in other fields¹. Research suggests that the idea could have merit for transit systems².

Dealing with Absenteeism

The above analysis assumes that drivers show up for work reliably. It does not take into account sick leave, vacation time, or absenteeism. Given m , the number of jobs needed, we assume a probability, f , that people will show up to work. We will also assume that absentees are paid. In the best-case scenario, with the same number of absentees every day, we would need to hire $n = \lceil m / f \rceil$ drivers; e.g., if $m = 60$ and $f = 0.9$, we would need $\lceil 60 / 0.9 \rceil = 67$ drivers.

¹ Coleman, R. M. (1995) "The 24 hr business", AMACOM, N.Y.

² Muñoz, J. C. (2002) "Driver shift design for single-hub transit systems under uncertainty" PhD thesis, Dept. of CEE, U. C. Berkeley, CA.

Public Transportation Systems: Management--Staffing

If we leave n as a decision variable but the number of people who show up to work, N , is random; and if the on-call workers who come in last minute to cover a shift, are paid at a higher rate ($\$_0 > \$$), then the expected cost would be:

$$E(\text{cost}) = \$n + \$_0 E(m - N)^+$$

Here, N is a binomial random variable $B(n, f)$. For large n , N can be approximated as a normal random variable with mean $(m - nf)$ and variance $nf(1 - f)$.

Note: The selection of n presents a tradeoff similar to the well-known “newsboy problem” in which a newsboy maximizes his expected profit by buying just enough newspapers to balance the risks of either running out or having some unsold inventory at the end of the day.

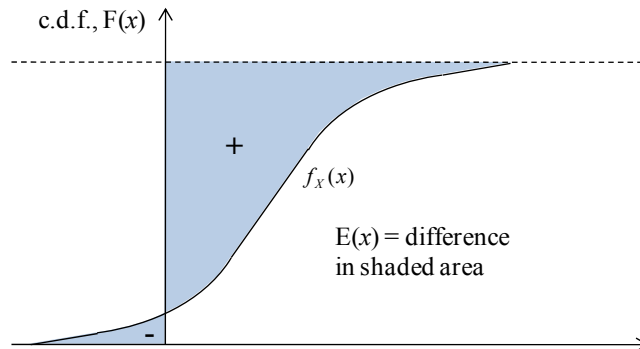
The formula for the expected cost is:

$$E(\text{cost}) = \$n + \$_0 \int_0^{\infty} \left(1 - \Phi \left(\frac{x - m + nf}{\sqrt{nf(1 - f)}} \right) \right) dx$$

where n is a decision variable, m and f are data, and x is a dummy argument. Here we have used the well known result for the mean of a random variable in terms of its cdf:

$$E(x) = - \int_{-\infty}^0 F_X(x) dx + \int_0^{\infty} (1 - F_X(x)) dx ; \text{ see Figure.}$$

There exists a closed form solution to expected cost integral above in terms of Φ and ϕ 's, exploiting the fact that $\int \Phi(x) dx = \phi(x) + x\Phi(x)$. However since the cost can be found numerically, the formula is omitted.



Since for our application the area to the left of the axis in the above figure is small, we will take the mean to be the area to the right of the axis. For the case given at the beginning of this section with $m = 60$, $f = 0.9$, and a cost ratio of on-call drivers $\$/\$_0 = 2/3$, the optimal solution would be $n = 64$. This would give an expected cost of 67.1 drivers, which is only marginally higher than our best-case scenario! Of course, worse results would be obtained with smaller pools of drivers. So, having a flexible workforce that can do many tasks is better than having many small pools of specialized workers. The formula we have given quantifies these effects.

What is still Left to be Done

Besides vehicle fleets and personnel, a transit agency also needs to manage other medium- and short-term problems. Some are quite visible to the public and may have to be handled adaptively in real time. We call them operational decisions. They will be examined in the last Modules, albeit not exhaustively. They include:

Real-Time Control of Vehicle Schedules and Response to Service Disruptions (Module 8)

Throughout this course we have dealt mainly with the assumption that buses run on schedule. In reality there are many possible disruptions that can prevent this from happening. These include vehicle breakdowns, delays caused by signals and congestion, and passengers with special needs, all of which can cause buses to divert from schedule. We need to know how best to recover from these disruptions. For example, using GPS and communications technology it is possible to introduce real-time controls which can minimize the effect of these disruptions.

Interaction between Transit and Other Modes (Module 9—in the works)

Because many forms of transit exist within the traffic stream, there are many interactions between transit and other modes. We need to learn how better to manage all modes together. We should learn when and how to segregate different modes on the surface streets so the available street space is better used.

Special events (Module 9—in the works)

Transit can be useful for special events such as the Olympics or emergency evacuations and we need to know how to plan ahead for these.

Module 8: Reliable Transit Operations

(Originally compiled by Vikash V. Gayah, April, 2010)

(Last updated, 9-22-2010)

Outline

- Reliability
- System of Systems (books on classical dynamics and non-linear oscillations)
- Uncontrolled Bus Motion (references 1, 2)
- Conventional Schedule Control (references 3, 7)
- Dynamic/Adaptive Control (references 4-8)
- List of References

Reliability

Reliability in a transit network refers to consistency in vehicle headways, arrival times, and schedules. When transit users are asked about the most important issues relating to transit, the number one response is the reliability of the system. Therefore, it is important for agencies to design systems that have consistent headways and vehicle arrival times.

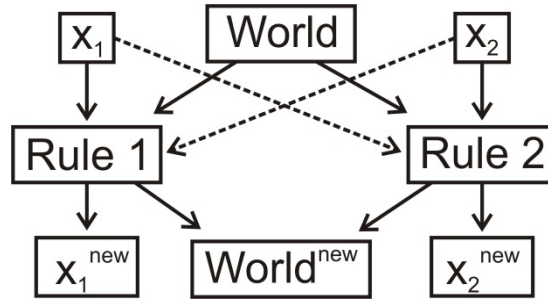
As will be shown here, most transit system are inherently unreliable – vehicles tend to bunch or pair, creating gaps in service. For an animation that explains why, see:

http://www.ce.berkeley.edu/~daganzo/Simulations/Bus_Bunching.html

We will learn in this module how to overcome this difficulty.

Systems of Systems

Transit systems can be analyzed as a “system of systems” (SoS). A system of systems is a group of interconnected systems (known as agents) that interact with decentralized agent-specific rules. Our goal will be to understand the macroscopic behavior of the SoS based on the individual rules governing the agents. The rules governing the behavior of a particular agent (i) depend on the current (and past) state, x_i , of the particular agent, outside factors (representing the world) and the state of other agents with which the agent interacts. The figure below graphically displays the generic structure of a 2-agent system: Agent 1 is on the left and agent 2 on the right; arrows denote the inputs and outputs of each agent’s rules. A SoS is characterized by the mathematical function embodying these rules (called the dynamic equations of the system).



Although the number of possible interactions in these systems increases quadratically with the number of agents, in transportation applications we typically encounter systems in which each agent only interacts with a limited number of agents. In this case, the number of interactions is comparable to the number of agents. Can you think of examples of SoS's in the transportation field and what the agents and rules would be for them? E.g., cars in traffic, airplanes nearing an airport, etc.

Since SoS's are decentralized, we need to understand their behavior over time. An important question to ask about such systems is: if the world is fixed at a steady state, does the system have an equilibrium state which is invariant in time? And if so, is this equilibrium unique? And is it stable? Stability means that the system tends to the equilibrium state when the overall state $\mathbf{x} = (x_1, x_2, \dots)$ is close to it.

These questions can usually be answered in three steps: 1) determine the dynamic equations for the system; 2) determine if one or more equilibrium states exist, and find them; and 3) determine which equilibria are stable.

To get some insight into the meaning of stability and this type of analysis, some examples of SOS are now presented.

Example 1 – A stable single-agent SoS: A parking lot with a fixed demand of vehicles entering the lot ($\lambda = 1000$ vehicles per time period) where 10% of the vehicles in the lot at the beginning of each time period leave by the end of the period. Here, the agent is the parking lot, the world is the entity supplying the demand and the state of the system is the number of vehicles in the parking lot at the beginning of any time period t , $x(t)$.

Step 1: the dynamic equation describes how $x(t)$ changes. It can be written using the given demand and supply rules. Note: the number of vehicles parked in the lot at the beginning of time period $t + 1$ is simply equal to the number of vehicles in the lot at the beginning of time period t , plus the number that enter during the time period, minus the number that leave. Therefore,

$$x(t + 1) = x(t) + 1,000 - 0.1x(t). \quad (1)$$

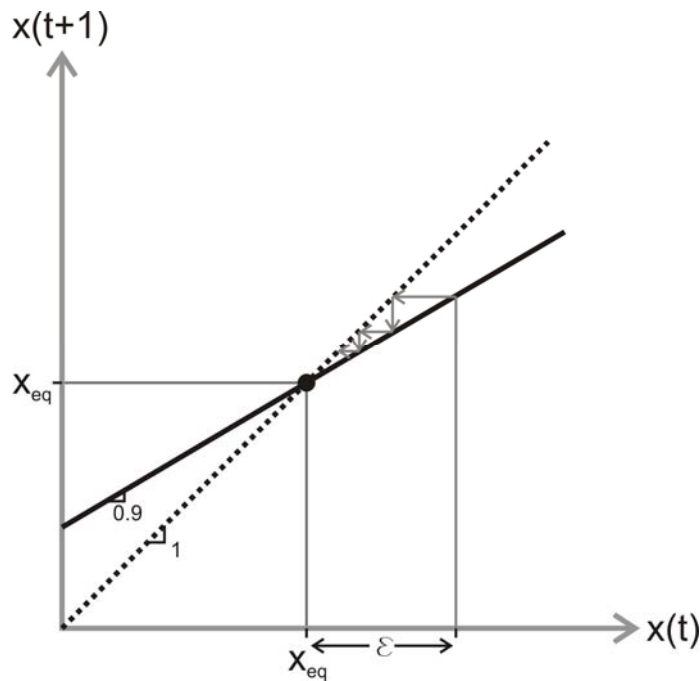
This is the dynamic equation.

Now on to Step 2: at equilibrium, the number of vehicles in the lot will not change with time. Therefore, the equilibrium solution can be found by setting $x(t) = x(t + 1) = x_{eq}$.

$$x_{eq} = x_{eq} + 1,000 - 0.1 x_{eq} \quad (2)$$

Solving this for x_{eq} , we find $x_{eq} = 10,000$. This solution is unique. So we are now done with step 2 and now check for stability.

Step 3: to determine if the equilibrium is stable, we need to examine what happens when $x \neq 10,000$ but $x \approx 10,000$, i.e., $x = x_{eq} + \varepsilon$. The equilibrium will be stable if $x \rightarrow x_{eq} = 10,000$. To graphically see what happens, plot the state of the system at time $t + 1$ as a function of the state of the system at time t . This is the dark line below, given by (1):



Imagine now that the system is perturbed by a value ε from the equilibrium state as in the figure. Using the dynamic equation along with the dotted line $x(t) = x(t + 1)$, we can see how the system evolves through time. A moment of thought reveals that it follows the grey arrows in the figure above. Clearly, the system moves back to the equilibrium state. This is also true if $\varepsilon < 0$. (Check it for yourself.) Therefore, we say that this equilibrium is stable since the system returns to x_{eq} after any minor initial perturbation, ε .

Stability can also be determined analytically by performing the same steps algebraically. To do this, define the residual perturbation after t steps: $\varepsilon(t) = x(t) - x_{eq}$. The dynamic equation can then be rewritten in terms of $\varepsilon(t)$ by subtracting (2) from (1). The resulting equation is:

$$\varepsilon(t + 1) = 0.9\varepsilon(t). \quad (3)$$

The reason for doing this is that we remove the independent constant from (1) and then the resulting equation becomes homogeneous and easier to analyze. Note, $\varepsilon(t) = 0.9^t \varepsilon(0)$. Thus, it is now clear that the perturbation decreases with time and tends to zero no matter the value of $\varepsilon(0)$. Therefore, any perturbation will be reduced in subsequent time steps and the system will move back towards the equilibrium.

Example 2 – an unstable single-agent SoS: The second example is a queuing system where the customers' service times increase with queue length. (This could happen, for example, if the length of the queue reduced the server's efficiency.) In our example, customers arrive at a rate of 1,000 per time step. The server can process $2,000 - 0.1x(t)$ customers in each time step, where $x(t)$ is the number in queue at the beginning of time step t . For this SoS, the server is the unique agent, the world supplies the demand and the state is $x(t)$.

Step 1: noting that the number of customers in the system cannot be negative, we see that the dynamic equation is simply:

$$x(t+1) = [x(t) + 1,000 - 2,000 + 0.1x(t)]^+ = [1.1x(t) - 1,000]^+. \quad (4)$$

Step 2: Now, replace $x(t)$ and $x(t+1)$ with x_{eq} and solve this equation. The system is found to have two equilibria: $x_{eq} = 0$ and $x_{eq} = 10,000$.

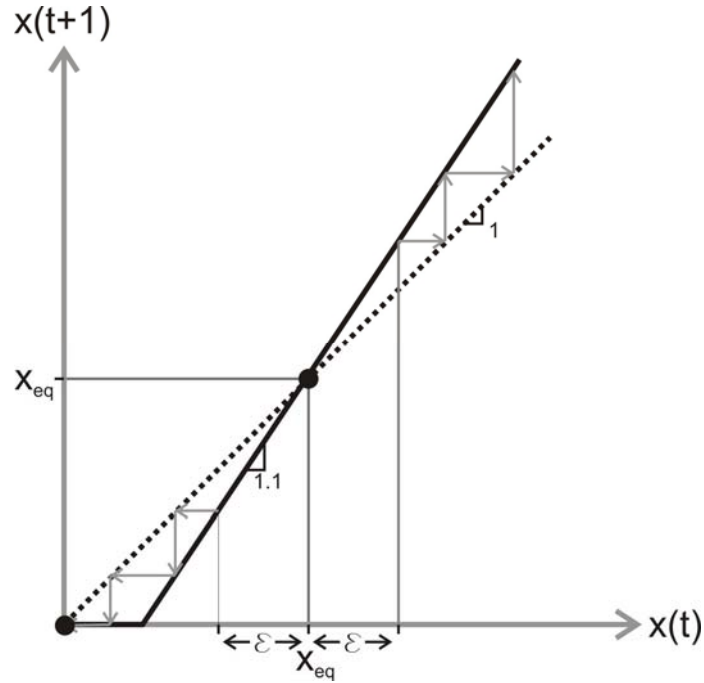
Step 3: the stability of all the equilibria should now be checked. Start with $x_{eq} = 10,000$. As a first sub-step we rewrite the dynamic equation as a function of the residual perturbations from equilibrium as we did for (3). To do this (4) must be "linearized"; i.e. we must remove the $[\cdot]^+$ truncation and eliminate the independent constant. The truncation has no effect close to $x_{eq} = 10,000$ since its argument is about 10,000. Thus, it is omitted. To eliminate the constant, we repeat the same step of example 1 (this is always done); i.e., we rewrite (4) for the equilibrium solution of interest:

$$x_{eq} = [1.1x_{eq} - 1000]^+ = 1.1x_{eq} - 1000 \quad (5)$$

and subtract from (4). The result, in terms of $\varepsilon(t) = x(t) - x_{eq}$ is:

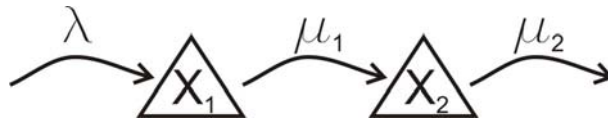
$$\varepsilon(t+1) = 1.1\varepsilon(t). \quad (6)$$

The second sub-step is analyzing (6). In this case, $\varepsilon(t) = 1.1^t \varepsilon(0)$. Thus, it is clear that any perturbation will continue to grow in size with time, and the system will move further and further away from the equilibrium state. Therefore, this equilibrium is unstable. This is also confirmed with the graphical construction of the state of the system. As seen in the graphical construction below, the equilibrium state $x_{eq} = 10,000$ is unstable. Minor perturbations move the system away from the equilibrium.



This whole analysis would have to be repeated for $x_{eq} = 0$, but we don't do it here. The figure clearly shows that the system moves towards the equilibrium $x_{eq} = 0$ when close to it, so this particular equilibrium is stable.

Example 3 – 2 agents: This is an example with multiple (2) agents. The system being studied is a system of two queues in series as shown below where the demand λ is a constant and the service rates depend on the queue lengths as follows: $\mu_1 = \beta x_1 + \gamma(x_1 - x_2)$ and $\mu_2 = \beta x_2 - \gamma(x_1 - x_2)$ where $\beta, \gamma > 0$. The γ terms indicate that work processing resources are constantly being moved from the small pile to the large pile, presumably to balance them. In this case, the queues are the agents, the world supplies the demand and the states are the queue lengths, x_1 and x_2 .



Since we have 2 agents, the graphical solutions we have given cannot be used. Therefore, we analyze the system algebraically. The method used can be applied to any number of agents. For our example, we normalize the units so that $\beta = 1$ and assume that in this system of units $\gamma = \lambda$.

Step 1: the dynamic equations of the system become:

$$x_1(t+1) = x_1(t) + \lambda - \mu_1 = -\lambda x_1(t) + \lambda x_2(t) + \lambda, \quad \text{if } x_1(t+1) \geq 0, \quad (7a)$$

$$x_2(t+1) = x_2(t) + \mu_1 - \mu_2 = (1 + 2\lambda)x_1(t) - 2\lambda x_2(t), \quad \text{if } x_2(t+1) \geq 0. \quad (7b)$$

Step 2: from these equations, an equilibrium solution of the system with $x_{1,eq}, x_{2,eq} > 0$ is:

$$x_{1,eq} = x_{2,eq} = \lambda.$$

Step 3, sub-step 1: the dynamic equations are now rewritten as a function of the perturbations from equilibrium. The equilibrium version of (7) is:

$$x_{1,eq} = -\lambda x_{1,eq} + \lambda x_{2,eq} + \lambda, \quad (8a)$$

$$x_{2,eq} = (1 + 2\lambda)x_{1,eq} - 2\lambda x_{2,eq}. \quad (8b)$$

Now, letting $\delta_i(t) = x_i(t) - x_{i,eq}$ and subtracting (8) from (7) we find:

$$\delta_1(t + 1) = -\lambda\delta_1(t) + \lambda\delta_2(t), \quad (9a)$$

$$\delta_2(t + 1) = (1 + 2\lambda)\delta_1(t) - 2\lambda\delta_2(t). \quad (9b)$$

We are done with sub-step 1 and must now check for stability. Note, we cannot draw a picture of (9) and it is not immediately obvious what happens to $\delta_i(t)$ if the equations are iterated. Fortunately, linear algebra comes to the rescue! Equation (9) can be rewritten in matrix form as:

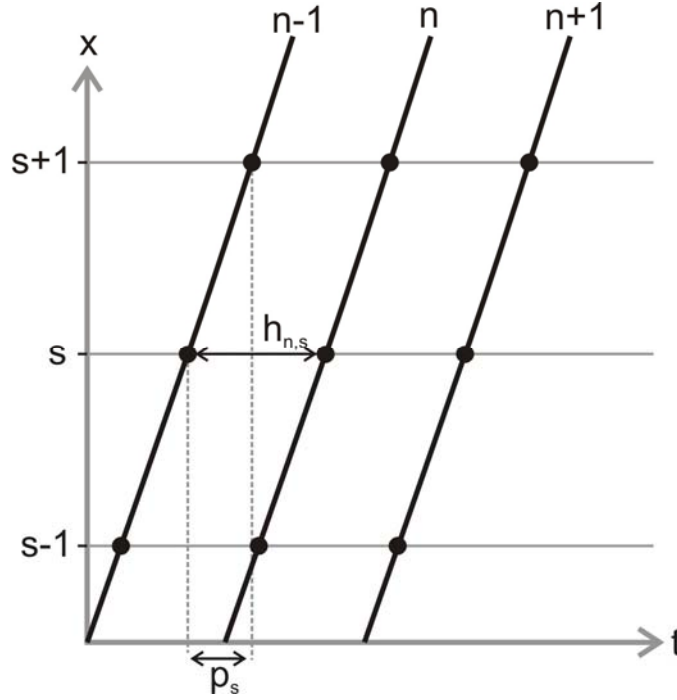
$$\boldsymbol{\delta}(t + 1) = \mathbf{L}\boldsymbol{\delta}(t) \text{ where } \mathbf{L} = \begin{bmatrix} -\lambda & \lambda \\ 1 + 2\lambda & -2\lambda \end{bmatrix}, \quad (10)$$

and in this form, the equation is very similar to (3) and (6). In this case $\boldsymbol{\delta}(t) = \mathbf{L}^t \boldsymbol{\delta}(0)$. Thus, as in the previous cases, if $\mathbf{L}^t \rightarrow \mathbf{0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ as $t \rightarrow \infty$, then the equilibrium is stable. This condition can be checked by analyzing the eigenvalues of the matrix \mathbf{L} . If the absolute values of all the eigenvalues are less than 1 then, as you may recall from linear algebra, $\mathbf{L}^t \rightarrow \mathbf{0}$ and perturbations will shrink with time. Thus, the equilibrium will be stable. If any of the eigenvalues are greater than one, however, the perturbations will grow with time and the equilibrium will be unstable. As an exercise, check what happens for $\lambda = 1$. You will find that the system is unstable. Also, see if you can determine for which values of λ the system is stable. (If you can answer this second question, you have reached a very good understanding of this method.)

Uncontrolled Bus Motion¹

We now apply these ideas to study an uncontrolled bus system, as shown below. In this system, the bus travels between points $(\dots, s-1, s, s+1, \dots)$ known as control points.

¹ Much of what follows is based on [7]



The ideal motion can be defined by the bus schedule in terms of arrival times at succeeding control points. This can be written in the following form:

$$t_{n,s+1} = t_{0,0} + nH + \sum_{i=0}^s p_i, \quad n, s = 0, 1, 2, \dots \quad (11)$$

where n is the bus number, s is the control point, H is the target headway, and p_i is the travel time from control point i to $i + 1$, including stops.

The first term on the RHS of (11) is the time at which the first bus arrives at the control point at the origin. The second term is the time separation between the first and the n^{th} buses at the origin. The last term is the bus travel time from the origin to $s + 1$.

We treat the buses, n , as agents, the control points, s , as “time” and the actual arrival times, which we denote $a_n(s)$, as the state of the agents. We want to see if the $a_n(s)$ stay close to the $t_{n,s}$ as the buses proceed forward (with $s \rightarrow \infty$). We shall use the notation $a_{n,s}$ instead of $a_n(s)$ for consistency with [7].

We are now ready to start the analysis. Equation (11) defines the equilibrium conditions of the system. However, the dynamic equations still need to be derived. As a preliminary step, note that the uncontrolled travel time for bus n between stops s and $s + 1$, $u_{n,s}$, should obey:

$$u_{n,s} = c_s + \beta_s(h_{n,s} - H) = c_s + \beta_s(a_{n,s} - a_{n-1,s} - H) \quad (12)$$

where $h_{n,s} = a_{n,s} - a_{n-1,s}$ is the headway ahead of bus n at control point s , c_s is the target travel time including stops at equilibrium and β_s is an experimentally determined constant (typically between 0.01-0.1 if the distance between stops is 1km). This constant captures the

extra time that boarding and alighting passengers add to the bus trip. Since $a_{n,s+1} = a_{n,s} + u_{n,s} + \gamma_{n,s+1}$, where $\gamma_{n,s+1}$ is a random noise due to traffic or the type and number of passengers arriving at s , it follows that the actual arrival time for bus n at point s is:

$$a_{n,s+1} = a_{n,s} + c_s + \beta_s(a_{n,s} - a_{n-1,s} - H) + \gamma_{n,s+1}, \quad n, s = 0, 1, 2, \dots \quad (13)$$

These are our dynamic equations. Note, (13) includes our exogenous noise term contributed by the “world”. So now we proceed with the linearization step. As before, we subtract (11) from (13) to get the DE in terms of perturbations away from the equilibrium, $\varepsilon_{n,s} = a_{n,s} - t_{n,s}$. The result after a little algebra is:

$$\begin{aligned} \varepsilon_{n,s+1} &= \varepsilon_{n,s} + \beta_s(\varepsilon_{n,s} - \varepsilon_{n-1,s}) + \gamma_{n,s+1} \\ &= (1 + \beta_s)\varepsilon_{n,s} - \beta_s\varepsilon_{n-1,s} + \gamma_{n,s+1}, \quad n, s = 0, 1, 2, \dots \end{aligned} \quad (14)$$

Consider now the stability of bus n , treating $\gamma_{n,s+1}$ and $\varepsilon_{n-1,s}$ as exogenous inputs. We are hoping that if these inputs are turned off, then $\varepsilon_{n,s} \rightarrow 0$ as $s \rightarrow \infty$. We see from (14) that $\varepsilon_{n,s}$ and $\varepsilon_{n,s+1}$ satisfy:

$$\varepsilon_{n,s+1} = (1 + \beta_s)\varepsilon_{n,s}, \quad (15)$$

which is unstable since $(1 + \beta_s) > 1$.

Therefore, uncontrolled bus systems are inherently unstable! When one bus gets behind, even just a little bit, the bus will tend to get further and further behind until it becomes paired with the bus behind it. The opposite happens if the bus runs ahead of schedule. Let us now see what can be done.

Conventional Schedule Control

This section will examine a typically used method to reduce the bus pairing phenomenon: conventional schedule control. In this type of control, slack is added to the bus schedule at predetermined control points along the bus route. Buses are held at these control points if they arrive early to get them back on schedule. Nothing is done to buses that arrive late, but the system is designed so that this is a rare event. There is a tradeoff in applying schedule control, however. The slack added to the schedule reduces the commercial speed of the buses, increasing the in-vehicle travel time that passengers experience to their destinations. So let us examine this tradeoff.

To recognize explicitly that the addition of slack changes the travel time between control points, let us write the scheduled travel time as $p_s = u_s + d_s$, where d_s is the amount of “slack”. The value of d_s should be selected to be greater than typical disturbances that arise in the bus movement. For example, if $\gamma_{n,s+1} \sim N(0, \sigma_s)$ then d_s can be set to $4\sigma_s$ so that the bus will be able to arrive on schedule over 99% of the time.

Since the slack has been added, the equilibrium bus travel times are given by (11) with $p_s = c_s + d_s$. They satisfy:

$$t_{n,s+1} = t_{n,s} + c_s + d_s, \quad n, s = 0, 1, 2, \dots \quad (16)$$

We assume that buses are allowed to run free between control points but are held immediately before the control points so they will not pass through them ahead of schedule². With this control strategy the dynamic equations are:

$$a_{n,s+1} = \max\{t_{n,s+1}, a_{n,s} + u_{n,s} + \gamma_{n,s+1}\}, \quad n, s = 0, 1, 2, \dots \quad (17)$$

We now express (17) in terms of deviations from the schedule as we did in the derivation of (14). To do this, subtract (16) from (17). The result is:

$$\varepsilon_{n,s+1} = \max\{0, \varepsilon_{n,s} + u_{n,s} + \gamma_{n,s+1} - (c_s + d_s)\}, \quad n, s = 0, 1, 2, \dots \quad (18)$$

Now remember that $u_{n,s}$ is related to the deviations by:

$$u_{n,s} = c_s + \beta_s(a_{n,s} - a_{n-1,s} - H) = c_s + \beta_s(\varepsilon_{n,s} - \varepsilon_{n,s+1}), \quad n, s = 0, 1, 2, \dots \quad (19)$$

and substitute this expression into (18). The final result is:

$$\varepsilon_{n,s+1} = [(1 + \beta_s)\varepsilon_{n,s} - \beta_s\varepsilon_{n-1,s} + \gamma_{n,s+1} - d_s]^+. \quad n, s = 0, 1, 2, \dots \quad (20)$$

This expression is very similar to (4) if we treat not just $\gamma_{n,s+1}$ but also $\varepsilon_{n-1,s}$ as input from the “world”. Like (4), equations (20) have a stable equilibrium at $\varepsilon_{n,eq} = 0$. In fact, even if we allow for small perturbations in $\gamma_{n,s+1}$ and $\varepsilon_{n-1,s}$ the term $[\cdot]^+$ equals zero if $\varepsilon_{n,s}$ is small. This means that bus n returns to schedule immediately. This is good. However, like (4) equations (20) also have an unstable equilibrium, $\varepsilon_{n,eq} = d_s/\beta_s > 0$. This instability means that if a bus is late arriving by more than d_s/β_s , then it cannot recover and it forever loses time. This result shows that conventional schedule control is not resilient to large perturbations, such as those caused by bus breakdowns. It explains why it is so difficult for transit agencies to keep buses on schedule, despite the agencies’ best effects, and why improved methods are necessary.

Before we look at these methods, however, let us now examine how d_s and the length of the control intervals should be chosen for schedule control.

Optimizing the Slack

In implementing conventional schedule control, a pertinent question becomes: how far apart (in number of stops) should the control points be placed? We now assume that the control points are spaced m stops apart where m is a decision variable.

The travel time between control points is now written as $p(m) = c(m) + 4\sigma(m)$ to stress the dependence on m . Recall $\sigma(m)$ is the variation in travel times between control points (not between stops). If all the stops are similar we expect $c(m) = mc$, where c is the time per stop. Now, let σ^2 be the variance of the travel time noise between successive stops. If the noise was independent across stops and all the stops were similar, we would also expect $\sigma^2(m) \approx m\sigma^2$ and

² If the control point is a stop, the slack can be introduced at the stop itself while the bus doors are open.

$\sigma(m) \approx \sigma\sqrt{m}$. This independence assumption breaks down for large m (as buses would begin to pair), but simulations in reference [7] show that it is good as long as $m \leq 0.25/\beta$. We now use these formulae to evaluate the bus average inter-stop travel time, $p = p(m)/m$, as a function of m . The result is:

$$p = c(m)/m + 4\sigma(m)/m = c + 4\sigma/\sqrt{m}, \quad \text{if } m \leq 0.25/\beta. \quad (21)$$

If the constraint is violated, p will be greater.

Note, p is the inter-stop travel time passengers experience in the bus. Users also experience waiting time for the bus, and this needs to be accounted for. Let $\sigma^2_{m'} \approx m'\sigma^2$ be the variance of the bus arrival at stop m' after a control point. Since a headway involves two buses, the headways vary at that stop with variance $\approx 2\sigma^2_{m'} \approx 2m'\sigma^2$. Then, the average waiting time for random arrivals at this bus stop, $\bar{w}_{m'}$, can be written as:

$$\bar{w}_{m'} = H/2[1 + 2m'\sigma^2/H^2]. \quad (22)$$

To get the average wait across all stops, \bar{w} , average the above formulae across all integer m' contained in the interval $[0, m - 1]$. The result is:

$$\bar{w} \approx H/2[1 + (m - 1)\sigma^2/H^2] = C + m\sigma^2/(2H), \quad (23)$$

where $C = H/2(1 - \sigma^2/H^2)$.

If a passenger rides for r stops, the average travel time including riding and waiting is:

$$T = rp + \bar{w} = r(c + 4\sigma/\sqrt{m}) + C + m\sigma^2/(2H), \quad \text{if } m \leq 0.25/\beta \quad (24)$$

Although this can be optimized for m , we also wish to prevent the buses from catching up with each other just by chance. To ensure that this is extremely rare, we use the constraint: $4\sigma(m) \leq H$, which in terms of m becomes $16\sigma^2 m \leq H^2$; i.e., $m \leq (H/\sigma)^2/16$.

To reduce one degree of freedom in the form of the mathematical program resulting from the combination of this constraint and (24), let $A = (H/\sigma)$. The result is:

$$z = \min\{-A + 8r/\sqrt{m} + Am\} \text{ s.t. } 1 \leq m \leq 1/(16A^2), 0.25/\beta$$

The solution of this mathematical program gives the optimal length of the control segment. You can verify that the unconstrained solution is: $m_u^* = 2.52(r/A)^{2/3}$. Thus, the actual solution recognizing the constraints is $m^* = \text{mid}(1, \min\{1/(16A^2), 0.25/\beta\}, 2.52(r/A)^{2/3})$. For small β and $A = 0.1$ (an intermediate value), the ratio z^*/H is: 0.5 for $r = 3$, 1.1 for $r = 7$ and 3.2 for $r = 20$. So it looks like travel time has to be increased by 1 or 2 headways for reasonable values of r if we want to achieve regularity. This indicates that conventional schedule control achieves regularity with a large travel time premium. And we saw earlier that it is not resilient to large disturbances. Therefore, a better control scheme would be appealing. The following are some ideas in this respect.

Dynamic (Adaptive) Control

Schedule control is rigid. Buses are oblivious to what other buses are doing. But recent advances in GPS and communication technology have allowed the possibility of information-sharing and adaptive control schemes. Perhaps this can improve performance. There are two approaches that can be used:

Approach 1 (e.g., reference [6]) – optimize the holding times and total travel times based on the current state of the system and its expected evolution. This is like a DP with recourse approach. Lots of literature on this topic but it is heuristic because nobody knows the recourse function and actions depend on assumptions about demand. Proofs have not been given that this approach prevents bus bunching.

Approach 2 – Control theory approach. Does not optimize travel time; instead it focuses on guaranteeing standards of headway variance and commercial speed. Proofs can be given showing that it prevents bus bunching and the standards are met. Some examples are:

- Forward looking headway control (reference [7]): it is adaptive so it will not ask a bus to slow down unnecessarily if the bus ahead is also ahead of schedule. Results in higher commercial speeds but still susceptible to the “escape” problem (low resiliency).
- Two-way looking spacing control (reference [8]): buses respond to their front and back spacing; buses can cooperate by slowing down to help following buses. Remedies escape problem. More difficult analysis.

Forward looking method

Recall from the analysis of (14) that the motion of bus number n is unstable (perturbations grow), if $|1 + \beta_s| > 1$. Clearly, if $\beta_s \in (-1, 0)$ this problem would be eliminated. But in car-following theory, this does not guarantee that perturbations could not grow across buses.

To establish this result we would have to express (14) in matrix form and then look for eigenvalues, as we did with (10). This is difficult to do in this case because n is unbounded (so the dimension of our matrix is infinite). Instead of dealing with eigenvalues, we use a more specialized result that applies to equations of the form (25) given below:

$$\varepsilon_{n,s+1} = \varepsilon_{n,s}\alpha_1 + \varepsilon_{n-1,s}\alpha_2, \quad n, s = 0, 1, 2, \dots \quad (25)$$

The result says that if $|\alpha_1| + |\alpha_2| \leq 1$ and (25) is iterated with increasing s , then the maximum error across all buses at any control point, $M_s = \max_n \{|\varepsilon_{n,s}|\}$, is bounded and cannot increase with s .

Proof:

$$M_{s+1} = \max_n \{|\varepsilon_{n,s+1}|\} = |\varepsilon_{n^*,s+1}| = |\varepsilon_{n^*,s}\alpha_1 + \varepsilon_{n^*-1,s}\alpha_2| \leq |\varepsilon_{n^*,s}\alpha_1| + |\varepsilon_{n^*-1,s}\alpha_2| \quad (26)$$

where n^* is the worst bus at location $s + 1$ and the last inequality holds due to the triangle inequality. Clearly, the last member of (26) cannot exceed $\max_n \{|\varepsilon_{n^*,s}|, |\varepsilon_{n^*-1,s}|\}(|\alpha_1| + |\alpha_2|)$ and since $|\alpha_1| + |\alpha_2| \leq 1$ it follows that $\max_n \{|\varepsilon_{n^*,s}|, |\varepsilon_{n^*-1,s}|\}(|\alpha_1| + |\alpha_2|) \leq M_s$. ■

In view of this result, we see that (14) would be well behaved if there was no noise and we could find a way of choosing $\beta_s \in (-1, 0)$ because then we would have $\alpha_1 = 1 + \beta_s \geq 0$, $\alpha_2 = -\beta_s \geq 0$, $\alpha_1 + \alpha_2 = 1$ for $s = 0, 1, 2, 3, \dots$. Therefore, the worst deviation in the noiseless system would not grow!

Reference [7] shows that how this analysis is extended if there is noise; i.e., if the equation is of the form:

$$\varepsilon_{n,s+1} = \varepsilon_{n,s}(1 - \alpha) + \alpha\varepsilon_{n-1,s} + \gamma_{n,s+1}. \quad (27)$$

In this case, reference [7] shows that the variance of the perturbations from the schedule grows but does so very, very slowly. And more importantly, the variance of the headways is uniformly bounded by this simple formula:

$$\text{var}(h_s) \leq \frac{\sigma^2}{\alpha(1-\alpha)} \quad (28)$$

So armed with this knowledge, we need to see if β_s can actually be changed to a value in $(-1, 0)$. To do this, we introduce slack into the travel times of the buses so buses can be accelerated when their headways (and, therefore, their workloads) are high and vice versa.

Propose: $p_{n,s} = u_{n,s} + D_{n,s}$ where:

$$D_{n,s} = [d_s - (\beta_s + \alpha)(h_{n,s} - H)]^+ \quad (29)$$

is the extra delay added. Equation (29) is truncated because buses cannot take less time than $u_{n,s}$. This control law can be displayed graphically as in the figure below.

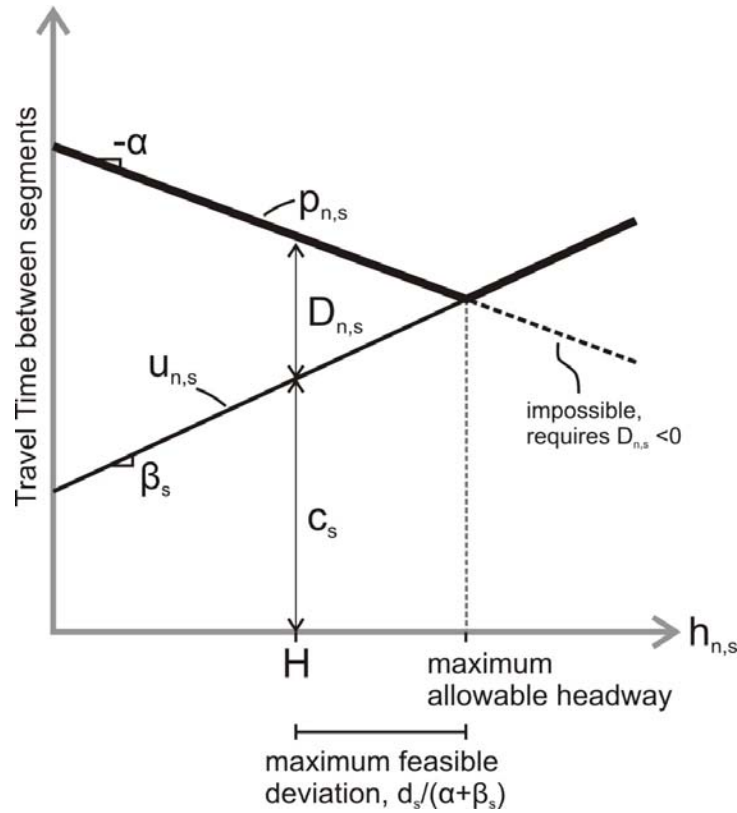
When (29) is truncated, buses run uncontrolled. We assume we choose a large enough d_s to ensure that this rarely happens. Thus, we now write the dynamic equations assuming no truncations. Note that $a_{n,s+1} = a_{n,s} + p_{n,s} + \gamma_{n,s+1} = a_{n,s} + u_{n,s} + D_{n,s} + \gamma_{n,s+1}$. Now use (19) and (29) to write:

$$\begin{aligned} a_{n,s+1} &= a_{n,s} + c_s + d_s + [\beta_s - (\beta + \alpha)](h_{n,s} - H) + \gamma_{n,s+1} \\ &= a_{n,s} + c_s + d_s - \alpha(\varepsilon_{n,s} - \varepsilon_{n-1,s}) + \gamma_{n,s+1}, \quad n, s = 0, 1, 2, \dots \end{aligned} \quad (30)$$

Since $t_{n,s+1} = t_{n,s} + c_s + d_s$ at equilibrium, we subtract this from the above and get:

$$\varepsilon_{n,s+1} = (1 - \alpha)\varepsilon_{n,s} + \alpha\varepsilon_{n-1,s} + \gamma_{n,s+1}, \quad n, s = 0, 1, 2, \dots \quad (31)$$

Note this matches (27). Thus, the bound (28) applies to control law (29).



Recall from the figure above that the maximum feasible deviation one should allow for the control to be in force is $\frac{d_s}{\alpha + \beta_s}$. Since the headway standard deviation is $\sigma/\sqrt{\alpha(1-\alpha)}$, we should set $\frac{d_s}{\alpha + \beta_s} \geq 3\sigma/\sqrt{\alpha(1-\alpha)}$ to assure that the system remains in the controllable regime most of the time. Therefore, set $d_s = \frac{(\alpha + \beta_s)3\sigma}{\sqrt{\alpha(1-\alpha)}}$. This is the expected holding time; i.e., the in-vehicle delay riders experience for every control segment.

Now the average waiting time will be $\frac{H}{2}\{1 + \sigma^2/(\alpha(1-\alpha)H^2)\}$ and the average riding time will be $r\left(c_s + \frac{(\alpha + \beta_s)3\sigma}{\sqrt{\alpha(1-\alpha)}}\right)$. With these functions, the average travel time can be minimized with respect to α .

It is shown in [7] that the added holding delay can be cut by a factor of 3 in a typical case. So this method is promising.

Two-way looking (cooperation)

Reference [8] contains an analysis. It introduces a spacing-based, two-way looking, linear control law (as if buses were attached to each other through springs) similar to the model in the homework with space as the state, and time as the parameter. The following are some key points:

- The coefficients of (25) are a Bernoulli pdf and the denominator of (28) is the variance of said pdf. This result is also true for versions of (25) with more terms and pdf coefficients.
- Physics: a control law that looks forward and backward leads to a dynamic equation like (27) but with 3 terms. The coefficients form a pdf with larger variance → better control under small disturbances,
- It can also introduce cooperation → no critical gap and no “escape” problem.
- It can reduce d_s slightly relative to headway-based laws but provides continuous monitoring with GPS; therefore, it recognizes large problems sooner.

List of References

- [1] Newell, G.F. and Potts, R. B. (1964) Maintaining a bus schedule. Proc. 2nd Australian Road Research Board, Vol. 2, pp. 388-393.
- [2] Newell, G.F. (1977) Unstable Brownian motion of a bus trip. Statistical Mechanics and Statistical Methods in Theory and Applications (ed. U. Landman). Plenum Press, 645-667.
- [3] Daganzo, C.F. (1997) Schedule instability and control. In: Daganzo, C.F. Fundamentals of Transportation and Transportation Operations, Elsevier, New York, N.Y. pp. 304-309.
- [4] Barnett, A. (1974) On controlling randomness in transit operations. Transportation Science 8(2), 102-116.
- [5] Newell, G.F. Control of pairing of vehicles on a public transportation route, two vehicles, one control point. Transportation Science 8(3), 248-264.
- [6] Eberlein, X.J., Wilson, N.H.M. and Bernstein, D. (2001) The holding problem with real-time information available. Transportation Science 35(1), 1-18.
- [7] Daganzo, C.F. (2009) A headway-based approach to eliminate bus bunching. Transportation Research Part B 43(10), 913-921.
- [8] Pilachowski, J. and Daganzo, C. F. (2010) Reducing bus bunching with bus-to-bus cooperation. Transportation Research Part B (submitted)