

# Routable Transit Network Design

By Ian Ford, Revision October 2006

## **Purpose**

The design of a routable guideway transit system is much different than that of a corridor system. In the typical corridor system, a collection of two-way linear routes creates a network, and travelers must transfer among routes. In a routable system, such as PRT or GRT, the design principles learned from corridor systems do not apply.

The purpose in developing theory about the design of such networks is to aid in design of initial systems so that as they expand, they “scale up” well in terms of connectivity, speed, and cost. Design mistakes in small systems could be very burdensome during expansion.

## **Observations**

A series of initial observations about routable networks follows:

- There are only **one-way segments** in the final analysis. A two-way corridor with two tracks is considered as two separate one-way tracks that happen to be near each other. The use of one track for bidirectional travel is useful in limited circumstances, but it degrades capacity so much that it is not considered relevant at the network scale.
- There are **no endpoints**. Automated vehicles have to stay on the one-way track, and so every track has to loop around to reconnect to some other track. Therefore all **networks must be composed of loops**.
- The theory applies to PRT, GRT to some extent, and dual mode, including freight and passenger applications. The theory would be assumed to apply to dual mode systems because any such system could also be designed to handle driverless vehicles, and so at least a part of the system would consist of one-way loops.
- Various network factors limit segment **capacity**. These are:
  - ◊ bidirectional travel on the same segment
  - ◊ merges
  - ◊ tracks crossing at the same level
  - ◊ a single segment that handles traffic from two or more travel directions, such as a traffic circle segment that must accommodate all east *and* northbound traffic
  - ◊ low-speed off ramps where vehicles must slow down while still on the main line, in order to negotiate a turn or stop; and low-speed on-ramps for the same reason. (This is not a problem if the on/off ramp is long enough to allow all acceleration/deceleration off the main line.)
- Various network factors limit **speed**. The main one is:
  - ◊ Curvature. Depending on the system, banking can alleviate the slow-downs required by curvature. However, near a merge or diverge point, the track cannot simultaneously be banked to optimize the left fork and also banked to optimize the right fork, so banking as a solution to this may be severely limited in some places.
- Various network factors affect **cost**:
  - ◊ Bi-level intersections, and height in general increases cost.
  - ◊ Longer on/off ramps increase cost.

- ◇ Redundancy of coverage increases cost; for example, if a person is with walking distance of two or more segments, the system might have been designed for more coverage for the same cost. (But redundancy is also a good thing for other reasons.)

## **Optimization**

What should be optimized? Here, people will not necessarily agree on the goals and priorities, but the main categories are as follows. The optimizations assume a given land coverage. (Optimizing mode split and system extent are whole different topics, which are not included here.)

- Least cost is probably the most important optimization, generally.
- Height is a concern that may outweigh cost in some places where views are a premium. Lower height is more favorable, and coincides with cost optimization.
- Lowest trip time is the predominant user goal, and network design affects this by how circuitous the routing is. For example, if there is a one way loop of 10 miles in circumference, and in order to go the “wrong” way a short distance on that loop, one had to travel all the way around, this would be a failure of the network design to optimize trip time. This also affects cost optimization.
- Highest segment capacity is a factor that allows expansion with less redundancy.

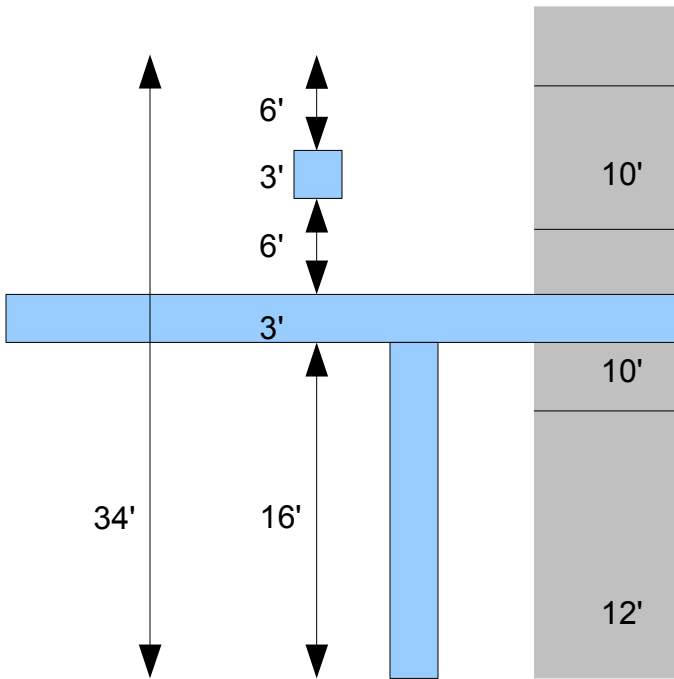
## **Major design choices**

Certain design choices made early in a small system can be easily changed later, while others tend to lock the expansion into the original choices. The choices made that lock in expansion are the ones that are of particular value to study here. Major choices that must be faced early on are:

- **Hierarchical vs. undifferentiated segments.** In a hierarchical network, some segments serve a distance function with restricted access, while others serve a local access function, perhaps with lower speed. We will refer to the distance segments as the level A subnetwork, and the local-access segments as the level B subnetwork. It could be broken further into level C and D and so on, but practically, 2-3 levels are probably enough. Strongly hierarchical networks concentrate traffic on level A, which means that the A subnetwork has to have the capacity to handle all traffic by itself, and the B subnetwork adds almost nothing to the system capacity. A hierarchical network may provide faster trip times. The alternative is undifferentiated segments, in which all segments provide the same service.
- **Speed and curvature** are related, as curvature is the primary limiting factor of speed. Tight curves in a small system could limit speed and capacity of an expanded system. Curvature limits of the A subnetwork of a hierarchical system might be more strict than for the B subnetwork; in other words the B segments could be more curved.
- Related to the choice of hierarchical structure is the choice of **segment classes**. Different classes of segments could have different weight limits, for example. In this case, light vehicles could traverse the whole network, while heavy vehicles would be limited to a subset of the network. Segment classes could also involve different curvatures, clearances, grades, or other constraints. These issues won't be addressed any further in this paper.
- While routable networks must be reducible to loops, one can design **two-way corridors** by flattening loops. A major design consideration is whether to include such flattened loops or not. If the ultimate goal is coverage of a metro area, two-way corridors don't add value, since one-way segments on a given roadway offer nearly the same level of service. The main reason for considering two-way corridors is in initial installation, where the goals may be simply to connect two or three activity centers. In those circumstances, there may be no perceived value in straying from the main travel corridor. This design choice strongly

influences how a network can be expanded, as explained below.

- As with highway intersections, **bi-level intersections** add height and cost, while nearly doubling capacity. The alternative is keeping the entire network on the same level.



This drawing shows a bi-level intersection with 16 foot ground clearance, 3 foot guideways, and 6 foot vehicle clearance. The first three storeys of a building are shown with floor heights as indicated.

Bi-level intersections add about 50% to the single-level height (or 36% if you include vehicle clearances).

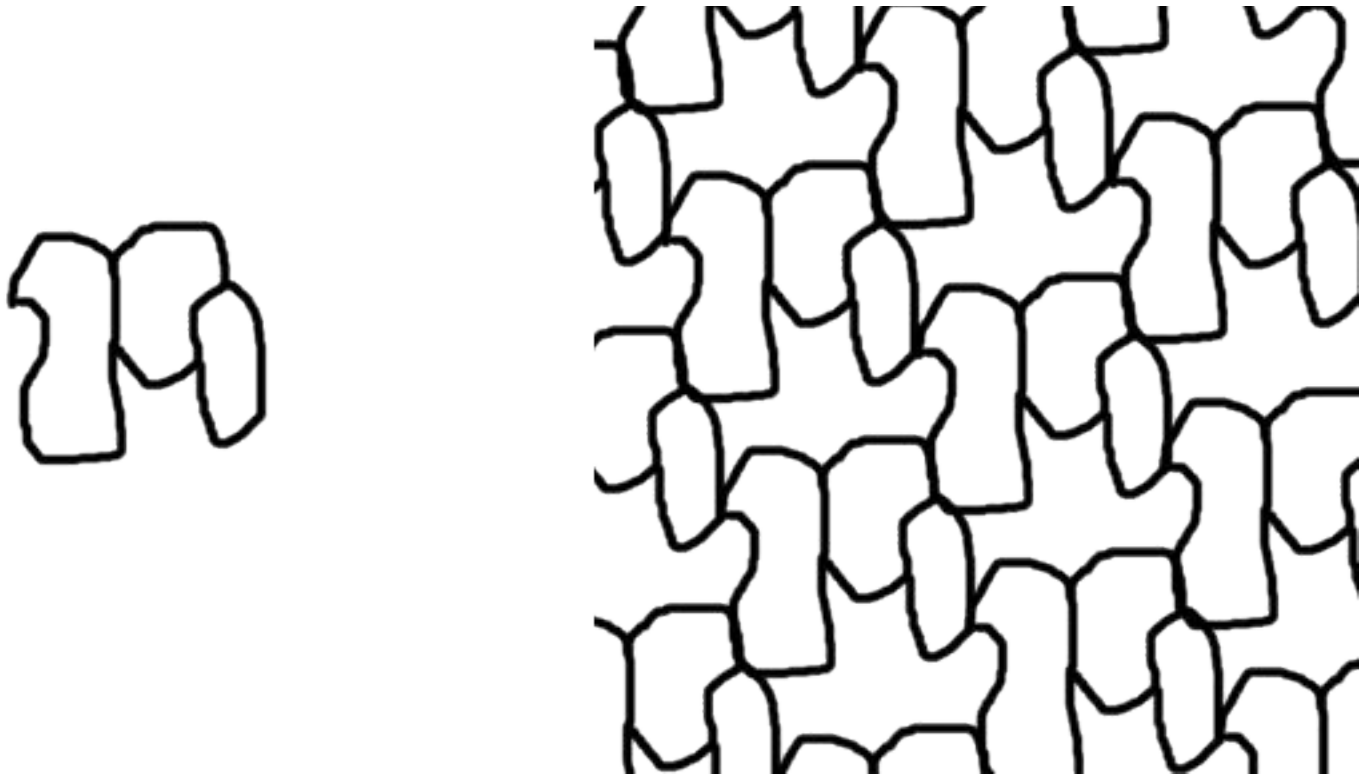
**Figure 1. Example heights of a bi-level intersection**

## **Connecting segments**

To connect segments, there are three building blocks, or connectors: **merge, diverge, and cross**. Crossing segments don't provide connecting service, but they reduce each others' capacity. The benefits of a cross are to avoid the cost of bridging one segment over the other (in an area planned for low capacity), and to avoid curvature.

We have identified that the whole network is by necessity reducible to a set of connected loops. Loops are either clockwise or counterclockwise. Loops connect by sharing one segment. If one loop is inside the other and they have a shared segment, they run the same direction. If the loops are adjacent, they run in opposite directions.

The design task is to make connected loops (which run in circles) useful to people (who generally want to go in straight lines, not in circles). It is easy to design loops that provide circuitous routing and *don't* combine to form any straight routes. Here is an example:



**Figure 2. Loops without straight connections might start out as workable (left) ...but don't scale up well (right)**

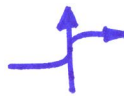
While the system on the left in figure 2 may be doable on the scale shown, it doesn't help define the theory of a scalable network. Multiply this map many times – as on the right – and you will see how absurd it would be to travel long distances.

The need for straight routes implies the need for intersections. In this context, intersections are defined as connections between loops using the three types of connectors, in a way that provides some straightness in the service provided. As noted earlier, we are excluding bidirectional track from this paper because of its limited role. We start with one-level intersections, and list the possible configurations.

- Straight-straight (figure 3) – This uses a cross, two merges and two diverges and likely has the highest cost. It offers the traveler the benefit of the least curvature when going straight in either direction. It probably has the lowest capacity, depending on a number of detail design points.
- Jog-straight (figure 4) – This uses one merge and one diverge. It prioritizes one direction by forcing all curvature on the other direction. There may not be room over a typical street to place this kind of intersection.
- Jog-jog (figure 5) – This also uses one merge and one diverge, and shares the burden of curvature equally on both directions



**Figure 3. Straight-straight intersection**



**Figure 4. Jog-straight intersection**

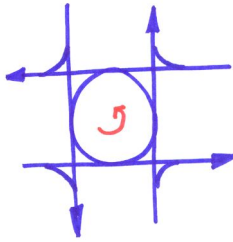


**Figure 5. Jog-jog intersection**

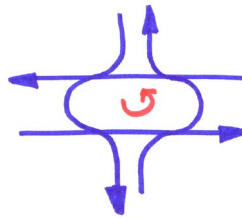
While there are many more ways to combine merge and diverge connectors, these are the three ways that provide the straight-through service that is the object of the traveler.

Intersection of a two-way corridor with a one way segment requires two one-way intersections, which could be

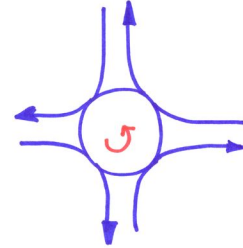
any of the three basic types. Intersections of two two-way corridors require four one-way intersections, assuming full connectivity. Using the notation NxE to denote a northbound intersecting an eastbound track, then two intersecting two-way corridors require intersections for NxE, NxW, SxE, and SxW. Each of the four individual intersections can be any one of the three basic intersection types. Of the  $3^4 = 81$  possibilities, here are some likely examples:



**Figure 6. Straight-straight intersection of two two-way corridors**



**Figure 7. Jog-straight intersection of two two-way corridors**

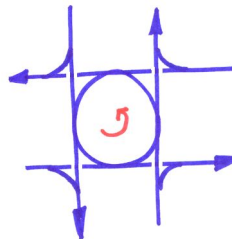


**Figure 8. Jog-jog intersection of two two-way corridors**

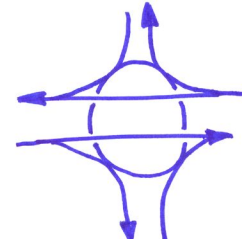
If you add bi-level intersections, the number of possibilities becomes very large. There are 3- and 4-level designs as well. By going to two or more levels, all through directions can be built with no curvature. Examples:



**Figure 9. Simplest bridge**



**Figure 10. Bridge of two two-way corridors**



**Figure 11. One of many other possibilities**

### **Area coverage frameworks basics**

Loops with intersections can provide the dual benefits of area coverage and straightness. By connecting loops at their corners, a grid is formed. Loops that form a grid may be called the “squares” of the grid.

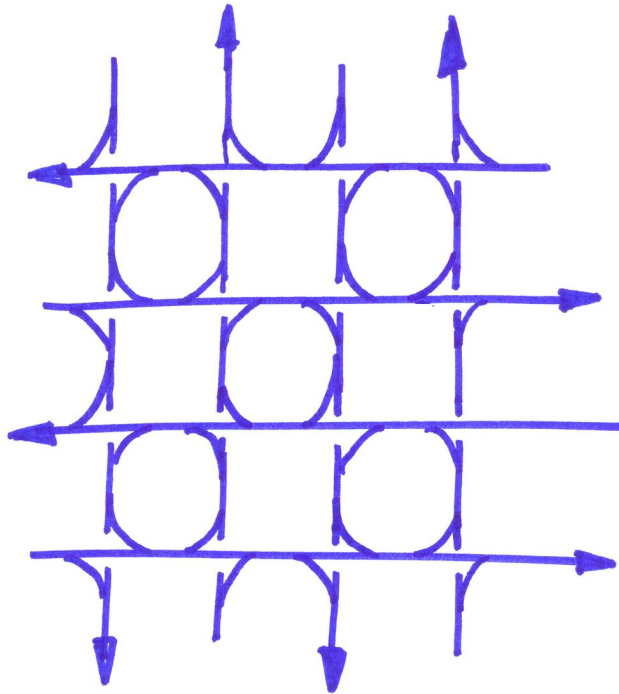
Expansion of any routable network can occur by any of these means:

- Adding a square onto the outside of the network, sharing one side or one corner with some existing square.
- Adding a square onto the inside of an existing square, sharing one or two sides.
- Adding a redundant path onto an existing square (either inside or outside), forming a new square that shares three sides with the existing square.

Any of these three kinds of expansion can be used with any area coverage framework, but it has to be done in a way that maintains the framework.

The following sections elaborate several different area coverage frameworks.

## Simple bi-level area coverage framework

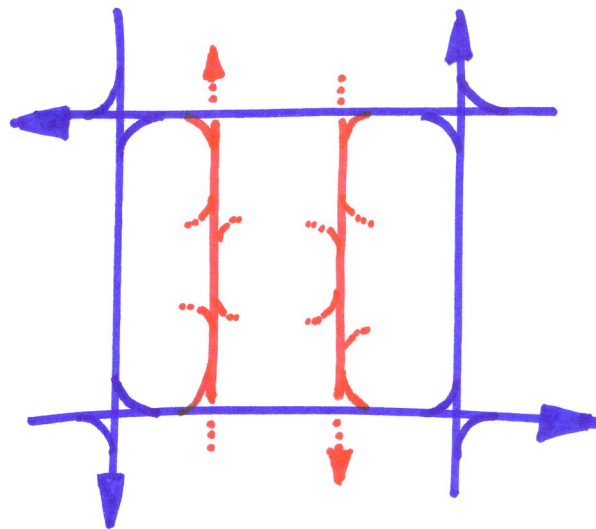


**Figure 12. Simple bi-level framework**

The “simple bi-level area coverage framework” uses simple bi-level intersections. At the expense of height, this kind of network has very low curvature for a typical trip. It also has high capacity compared to any framework with single-level intersections. All the N-S tracks are high, and all the W-E tracks are low, or vice versa.

External expansion can occur by adding a square with the same high/low pattern as the adjacent grid, and building the non-connected corners with a stub of an intersection that can be completed by a later expansion.

Internal expansion can occur by trisecting any grid square by a pair of tracks at the same level, and building four half-intersections that can be completed later.

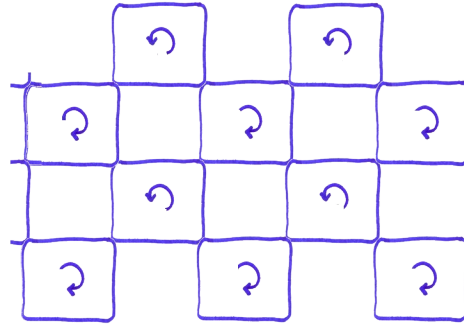


**Figure 13. Internal expansion of simple bi-level framework**

If only one of the two trisecting tracks is built, it should be placed and directed so as to leave room for the other one at a later date.

Internal expansion in this framework has limits. You cannot easily expand in a hierarchical fashion. If the existing track is level A (the top of the hierarchy) and your intent is to add level B (slower, more access points) within a level A square, then the four internal intersections are not so flexible: you either have to stay with more costly bi-level intersections throughout, or bring the height down and have single level intersections, or omit connections at those points.

### ***Jog-jog area coverage framework***

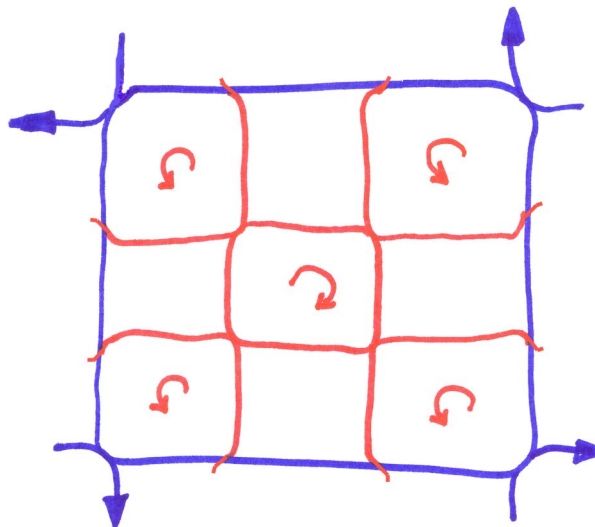


**Figure 14. Jog-jog framework**

The “jog-jog area coverage framework” uses jog-jog intersections and keeps the network at the same height. This reduces the price of the bi-level framework while retaining most of the straightness. Its capacity is much lower since traffic in both directions at each intersection has to merge onto the same track segment.

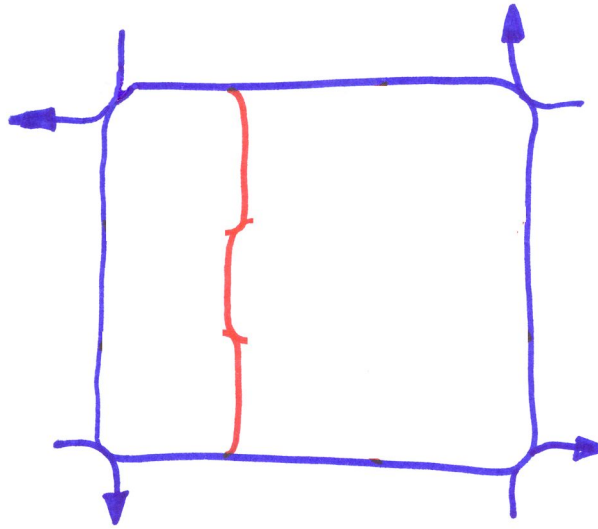
External expansion can occur by adding a square with curved corners onto any outer corner of the network, or by completing a partial square.

Internal expansion is a bit more complex and interesting. Basically, you add five squares to the inside of any existing square as shown in figure 15. You can then add four squares to the space between existing squares, as shown in figure 17 in green. Either way, the effect is to trisect the square in both directions.



**Figure 15. Complete internal expansion of one square of a jog-jog framework**

You can add as little as one of the five squares, or one redundant path, but the redundant path should contain the jogs and intersection stubs where expansion would occur.

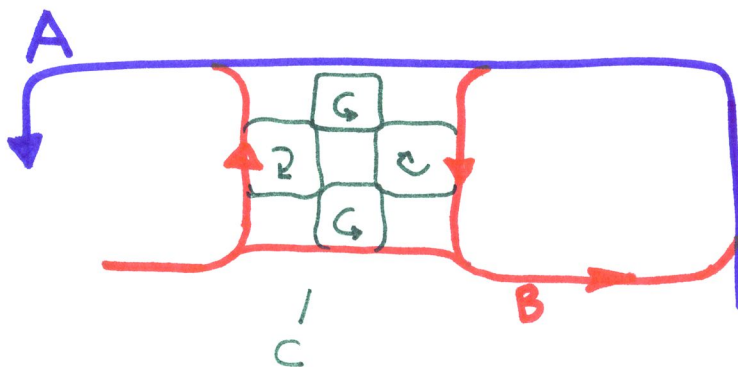


**Figure 16. Minimal internal expansion of a jog-jog network by adding a single redundant path**

Internal expansion of a jog-jog framework allows for hierarchical track. Regardless of scale, the five new squares are one level below that of the surrounding square. For example, if there is an existing A square, then the five squares built inside it would be B squares. The four new B-B intersections would be jog-jog. In fact, all intersections between the same hierarchical level are jog-jog. The eight A-B intersections would be jog-straight: since the A is the higher level (and was there first), all curvature is handled by the B segment. Note that A-B intersections may not fit over existing streets so well, so the B grid may have to switch between two parallel roads each time it crosses an A track.

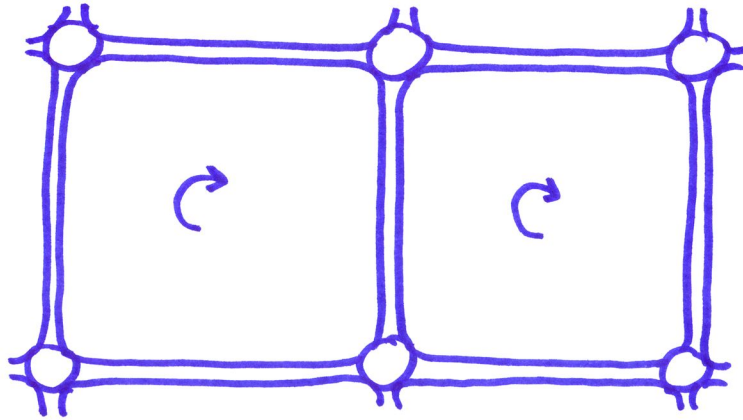
Within any B square, a set of 5 C squares can be built in exactly the same way as B squares are built inside A squares. Each hierarchical level can have a design speed – for example A=40 mph, B=30 mph, C=20 mph. Therefore the curves could be tighter for the lower levels.

A special problem occurs when building a C square onto an A square. In order to maintain priority and capacity of the A segment, the lower speed of the C segment should not affect the A segment. To handle this, a separate non-connected side of the C square can be used. If connectivity is important at the location, an accelerating entrance and/or a decelerating exit could be built.



**Figure 17. Adding a C square adjacent to an A square**

## ***Two-way corridor area coverage framework***



**Figure 18. Two-way corridor framework**

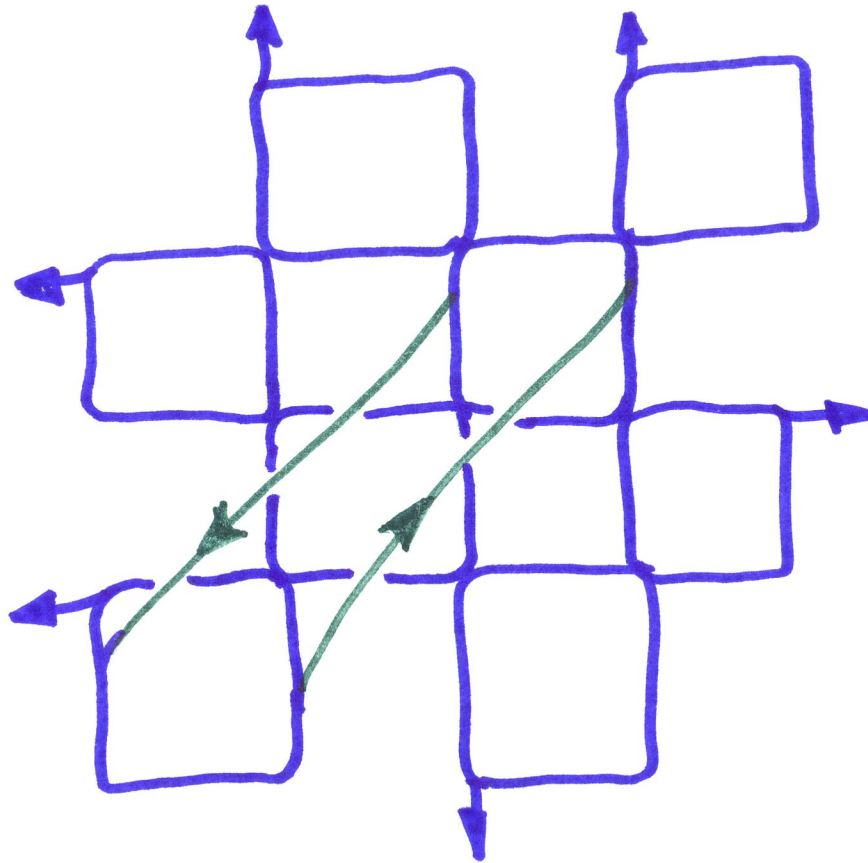
The “two-way corridor area coverage framework” uses two-way corridors and bidirectional intersections. It is shown here with jog-jog intersections. This mimics the car/road system in which all traffic is on the right (or mirror image). In this type of area network, redundancy is very high with only a small savings in trip length.

External expansion is straightforward.

Internal expansion is very impractical if two-way corridors are to be used internally. However, the internal square can be expanded the same way as the jog-jog framework is expanded, as explained in the previous section. The main downside of this is the impracticality of B segments to intersect the A segments, because the B segments would have to make two jog-straight intersections and turn 180 degrees within the width of the corridor.

### ***Special features***

- The most detailed hierarchical level of track would be an on-street low speed level. In this concept, vehicles would require systems to avoid hitting people and objects, and such technology already exists. An on-street track could be very inexpensive, particularly if not electrified.
- A single-level network can be expanded vertically in a free-form way without breaking the framework structure. This could be done in a diagonal to offer express route, as shown in figure 19.



**Figure 19. Example diagonal express track, bridged over the grid**

### ***Recommended strategies***

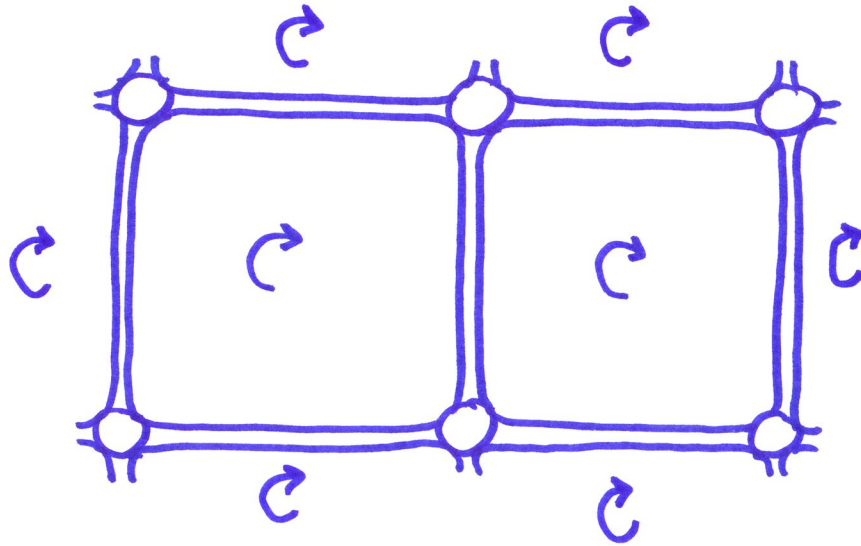
The strategies will be explored in a case study, using Albuquerque, NM. The choice is based solely on the author's familiarity with the city.

The design philosophy that will be used is “think big and build small”. It is easier to take away from a plan than to add to it, so the plan should be as comprehensive as possible. On the other hand, it is easier to add onto a structure than to take away from it, so the structure should be as small as possible, initially, while meeting some transport need.

Which framework is the best? There is no clear answer to that, because it would depend on local priorities. The author's recommendation is that if speed and capacity is felt to be worth the extra height and cost, then the simple bi-level framework is the best; otherwise the jog-jog framework is the best. There is room for much more work in this area before making any final recommendations: First, there may be more useful frameworks that have not been discovered. Also, this paper is not informed by simulation modeling, which could add useful information.

One promising option is to make the A grid bi-level, spaced at 1-3 mile intervals, and expand internally using the jog-jog framework.

Any framework based on single-level two-way corridors can never exceed the price and performance of either of the other two frameworks, so it should be avoided. After several rounds of expansion, the corridors become a liability – adding cost and reducing connectivity. It is impractical to abandon the corridors because the directionality pattern when using two-way corridors is that all squares (at the top hierarchical level) are clockwise. In the other frameworks, any two connected squares have opposite directionality. Once the all-clockwise pattern is set, there is no place to put a counterclockwise square.



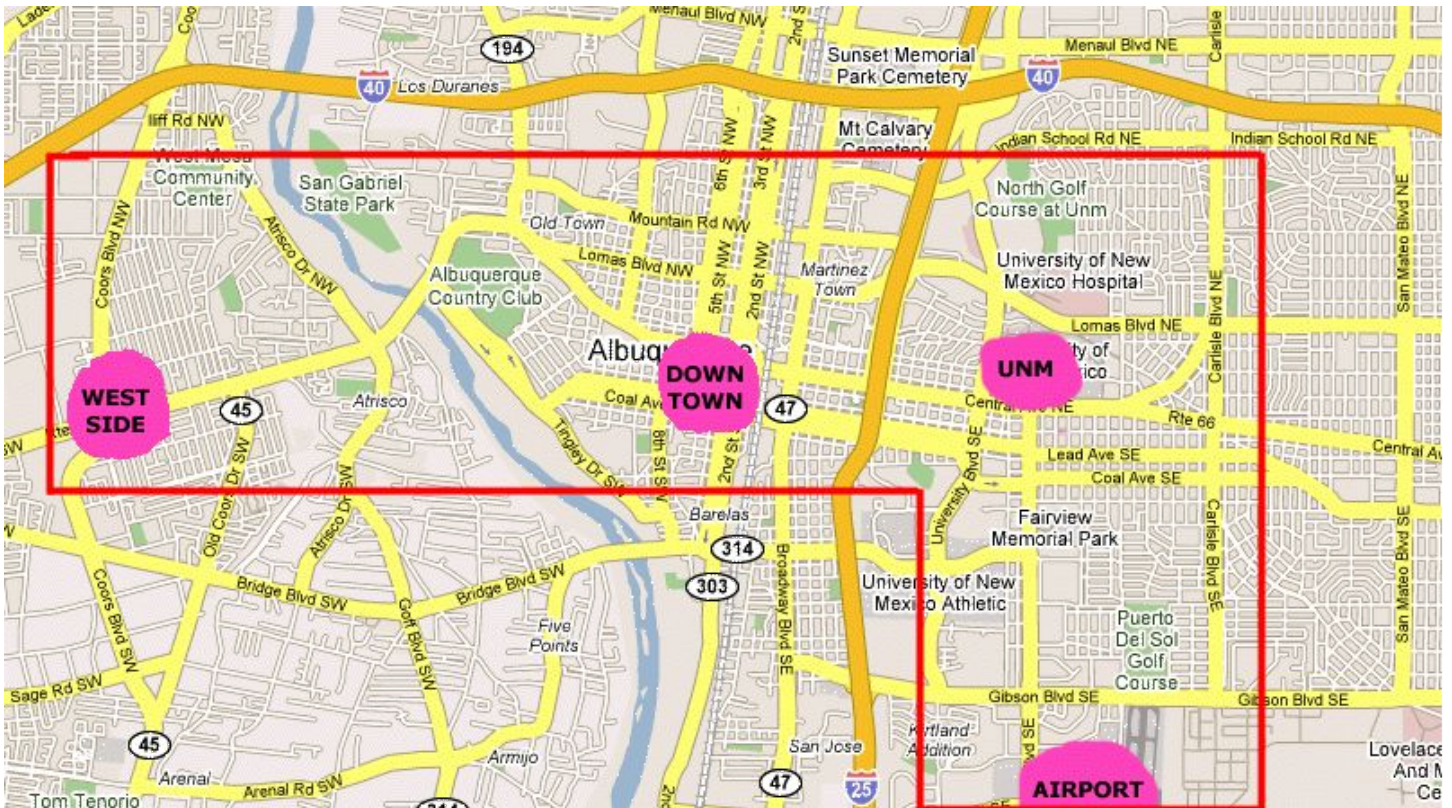
**Figure 20. In an all-clockwise pattern, all expansions must be clockwise.**

If a city builds a two-way corridor and then runs into the expansion problem shown in figure 20, it can be partially isolated. It can be isolated to a single swath that bisects the whole metro region. Suppose the initial corridor was east-west. This corridor could be expanded to the eastern and western edge of the city, retaining its two-way mode. On the north and south sides of it, a jog-jog or bi-level framework could be built. However, if the corridor pattern was also extended north or south, then it would have to be expanded that way in a wide east-west swath of that pattern. A one-way framework can then be built to the north and south of the swath.

The jog-jog framework is chosen for this Albuquerque example because height is a major factor: the region appreciates its mountain views. A pure passenger PRT system will be used in this example; in other words, there will be no group/mass transit, freight, or dual-mode features.

Each design step is explained so that the reader may apply the process to a different city. This should be understood as illustrative only; it may turn out that a different framework is a better choice.

## Step 1. Determine the maximum service area of a pilot project



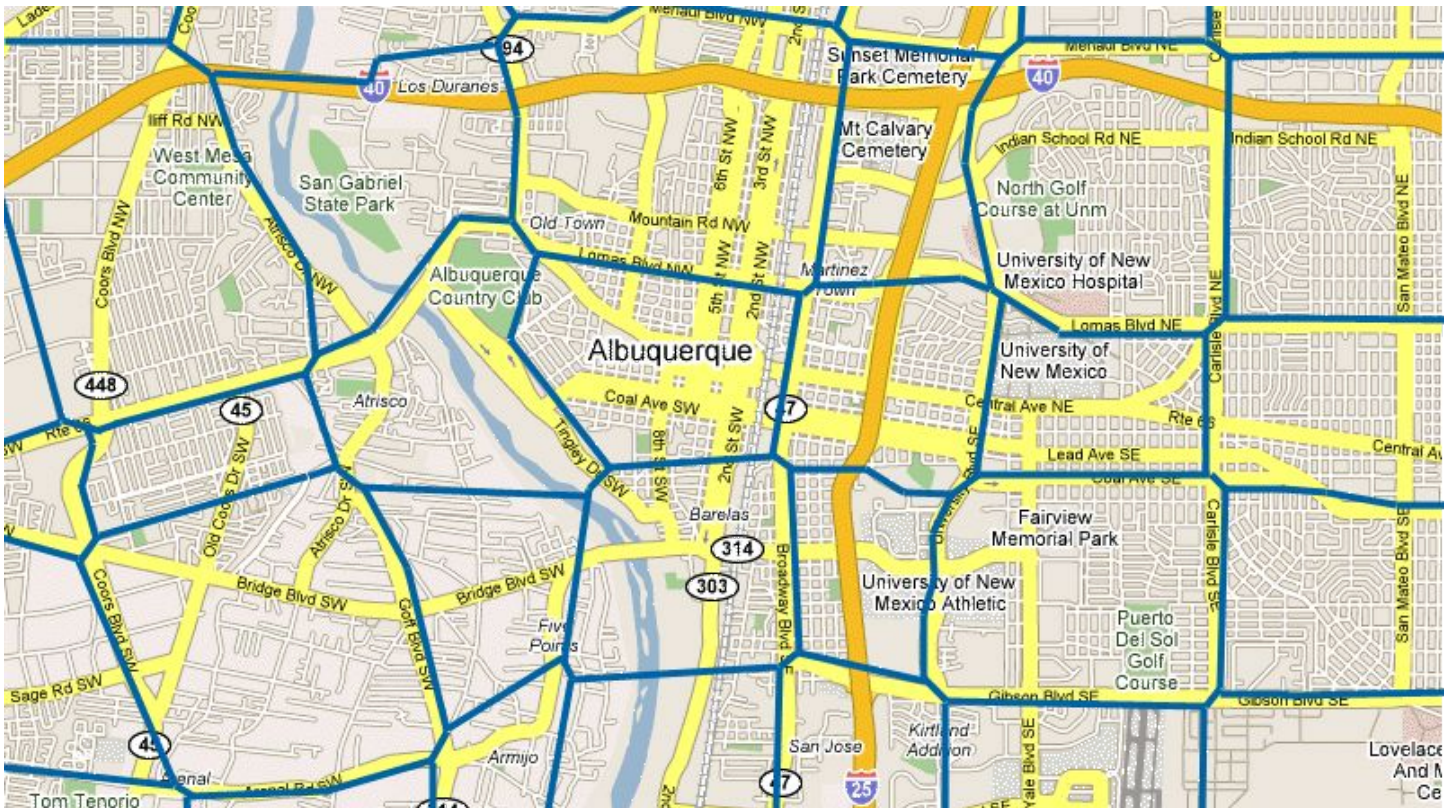
**Figure 21. Service area of a pilot project for Albuquerque**

Which activity centers are to be connected? Is there a need for local circulation within an activity center? These are largely political questions. For Albuquerque (see figure 21), the initial network might connect a park and ride lot on the west side of the river with downtown, the airport, and the university area. Internal circulation would be needed in downtown and UNM (the university).

There are different stages of testing, implementation, and acceptance. A minimum *test* facility could be very small – less than one mile. A test facility could be certified for revenue operation, and then expanded. A “pilot project” in this paper denotes a project that has already been certified for revenue operation, and is large enough to be useful in its own right. Figure 21 shows the “maximum service area” as a red line, which is the area in which the pilot would get built, but the pilot would not occupy the whole service area.

In light of the fact that routable networks have no track record, a new system might have to be a very small pilot project. In other words, a network would have to start out as not being a network at all. (Many people might not see the point in building a “non-network” that can expand to become a network. This may explain why routable networks have not been built to date.)

## Step 2. Enlarge the study area, and plan the A squares in that larger area

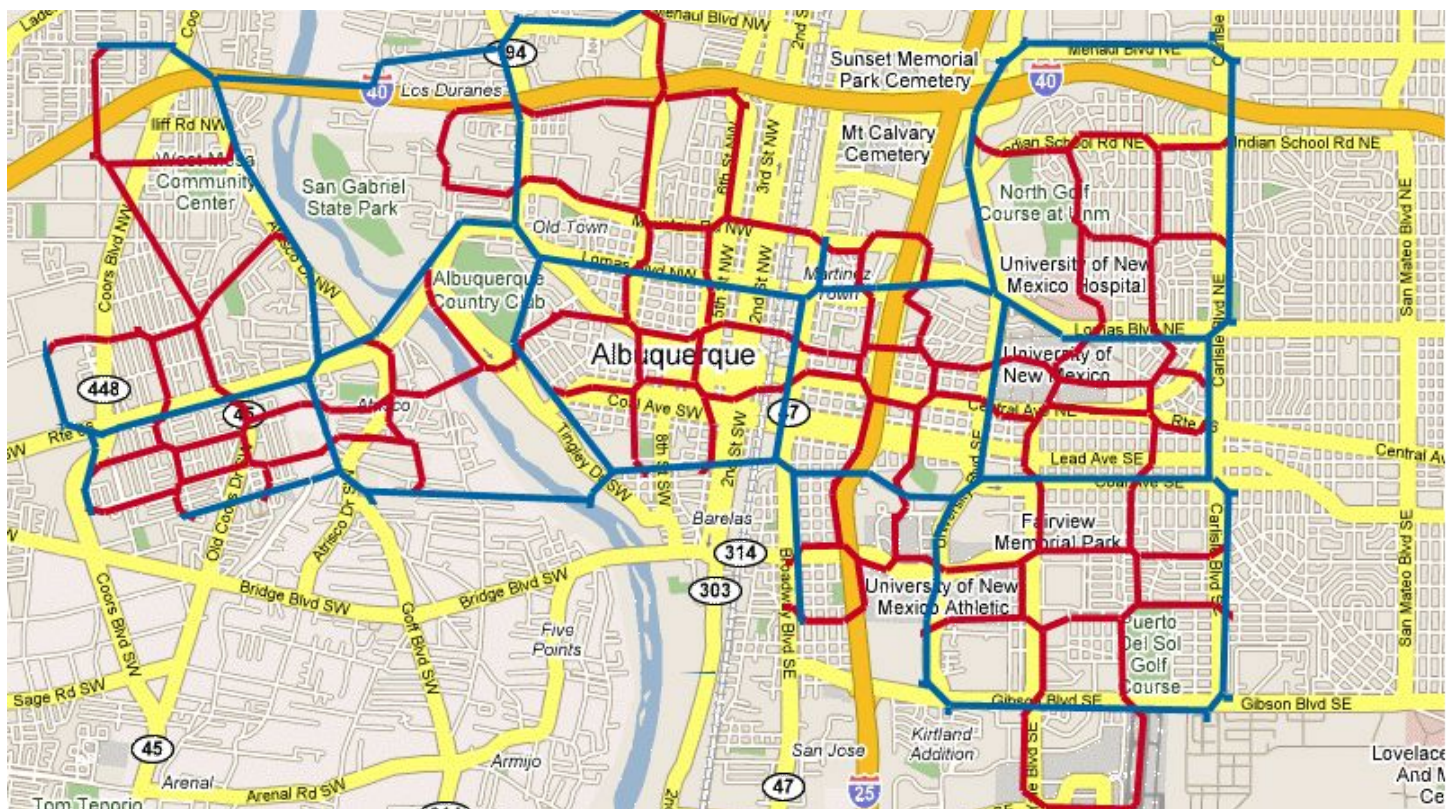


**Figure 22. “A” squares encompassing the pilot service area**

“A” denotes the top hierarchical level of the framework. Attempt to make A squares 1-3 miles on each side. This means that travelers on the A subnetwork will experience jogs at 1-3 -mile intervals. At this point, there is no need to be site-sensitive. The map developed here prepares for step 3, but it is not meant to be a strict guide for actual build-out. In fact, some of the segments shown go over historic or other sensitive features.

Figure 22 shows one way to draw A squares for the service area. Some squares are much larger than others, because that's the best compromise that was found given the irregularity of the roads. Many of the A segments shown will never be built, for various reasons: some go through neighborhoods, some make a long river crossing in a low-demand area. But all of these limitations are OK at this point. We are only trying to lay out the top level framework in order to hang the more detailed framework on it.

### Step 3. Draw the maximum pilot network using the framework from the previous step



**Figure 23. Maximum pilot network**

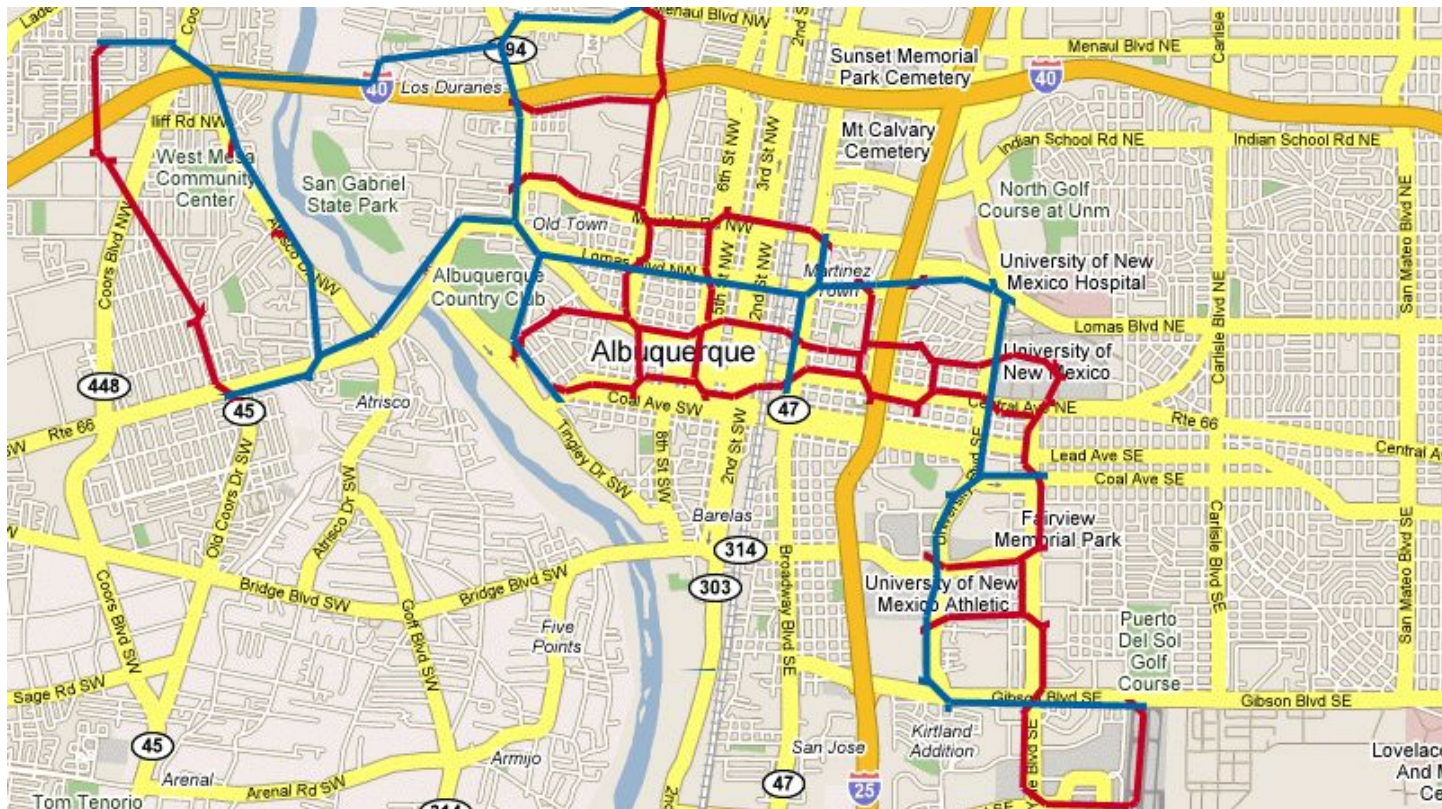
All segments built for the pilot project should also be latent segments of the larger potential network. The pilot project may contain some A and some B squares, or parts thereof. Any B segments should contain all the jogs and stubs so that its latent role in the potential expansion may be expressed at a later date without an interruption in service. See figure 23.

Note that the jogs are coordinated with the street network: each jog takes you to the opposite side of the street. The hierarchical levels are color coded (blue is A, red is B). Notice that the pilot contains incomplete A squares and incomplete B squares. Without an understanding of how we got to this step, someone seeing the pilot map would probably not see why the eastbound track near Central Avenue has jogs in it, while the westbound one on Lomas does not. Readers of this paper will see that the eastbound track is a latent series of B squares connected by jog-jog intersections, while the westbound track is one side of a larger A square.

Each step up to this point involves different actors. Step 1 is a political decision of where to start. Step 2 is a technical exercise that should be done by an informed engineer and reviewed by a small team of experts. Step 3 is much more involved because it has to be site-sensitive. Sensitivity factors include:

- Important historic buildings (which cannot be obscured by elevated track)
- Pedestrian circulation
- Location of stations, and parking at stations
- Many more environmental factors that are beyond the scope of this paper

#### Step 4. Choose a minimum functional network to actually build first.



**Figure 24. Minimum functional network**

Figure 24 is an example of an initial network. In the figure, the stubs are still visible where segments were erased. The reasoning behind choosing to keep certain segments from figure 23 and erase others has to do with Albuquerque land use factors, such as commercial and residential density. Also, the river crossings were limited to existing bridges, which kept the northern part of the network intact. The southeastern part of the network (which goes to the airport) is about as simple as you would want, because you need both directions, and you need several places where vehicles can turn around. Without the east-west segments, some trips would be routed far out of the traveler's way. The downtown grid is expanded to more squares to provide circulation within downtown and to improve capacity of the middle part of the system. The west side loop is possibly too large, but its main function is to serve park-and-ride lots for people commuting to the east side, which is where most of the jobs are.

This case study was undertaken as an example and based on some knowledge of Albuquerque, but is not meant to be taken as a proposal.

#### Step 5. Expand!

Fit each expansion into the framework by reviewing and amending steps 1-4 for each expansion. Build all corners with expansion stubs so that you don't have to shut down the network for the next expansion. Avoid adding redundant paths.

As a final note about the recommended five steps, please notice that the lack of two-way corridors turned out to be a convenience and a great asset for expansion in the Albuquerque example. There is likely to be political pressure to start out with two-way corridors because it appears to be simpler and more like traditional transit. But this pressure should be countered by the arguments in this paper about expansion, and also by the greater

service area that can be achieved, such as the area served in figure 24.

## **Remarks on planning**

The question of the appropriate sizing and capacity of a routable network has often been raised. Given that construction of a network would occur over a long period of time, and travel behavior changes over that time, it is impossible to plan ahead for a known level of demand.

For the same reason that new roads aren't built until after there is intolerable congestion, routable network expansion may not occur until after there is an intolerable wait time to board. On the other hand, expansion may occur prior to saturation as a way to attract more business. It all depends on the power held by those who stand to profit from these decisions.

The planning aspect is difficult for several reasons:

- For planners and engineers, routable networks are a paradigm shift from roads and traditional (non-routable) rail transit. People can't rapidly shift to different methods of infrastructure planning.
- According to many PRT developers, PRT systems could be profitable without public subsidy. This raises the problem that democratic political control of the infrastructure may be lost. Public control of development is tenuous, often legally challenged, and often corrupted by private interests. Any further loss of public control to the profit motive could have unforeseeable consequences.
- Many people believe that the style of development that has been practiced since the 1950s has been oriented exclusively to the private car and has been detrimental to cities. If routable networks offer a new level of speed and freedom, and development starts to become oriented to that, it could result in more of the same kind of detrimental consequences, such as further sprawl and environmental damage, and less walking leading to declining health.

The conclusion from the above remarks is that planning for routable networks must be ongoing, can never be done perfectly, and it directly interfaces with basic questions of development, land use, power and politics.

Until recently, the option of routable transit networks was not available. Now we have the option. Will we refuse the option simply out of fear of unintended consequences? Or, will we take democratic control of the kind of development and infrastructure that we want and use the option as one of many tools to achieve public goals?

## **Summary points**

The critical advice of this paper is: "Plan big and build small". Another way to say this is that every segment that is built must be the expression of one piece of a larger, latent framework. This kind of thinking leads to decisions such as the following, which might not be apparent otherwise:

- Avoid two-way corridors.
- Build on every *third* arterial in the street grid, initially, so that internal expansion can trisect the initial grid.
- Build expansion stubs so that expansion can occur without an interruption in service.
- Set curvature limits for long term speed.