Abstract. The rapid expansion of on-demand ride-hailing services has changed the transportation landscape in many cities worldwide. While these services provide a valuable service, as evident by their popularity, there are many questions regarding their scalability, efficiency, equity and negative externalities (e.g. congestion, pollution etc.). One way to mitigate some of these concerns is to consider high-capacity ride-pooling services, where many passengers share a vehicle simultaneously. Operationalizing such a system requires the ability to efficiently match large groups of riders to a fleet of shared vehicles in real time, a computationally challenging task to solve at an urban scale. The first part of this talk presents a framework for solving this problem at the scale of New York City, capable of routing 450,000 passengers a day in real-time using a fleet of two to three thousand shared micro-transit vehicles of capacity four to ten. The second part of the talk will focus on a discussion around how these on-demand services can/should be integrated with more traditional mass-transit systems to exploit the benefits of both types of systems (while limiting their drawbacks), and how to incorporate behavioral and economic considerations into such assessments.