Applications of Operations Research in the Air Transport Industry

Cynthia Barnhart • Peter Belobaba • Amedeo R. Odoni
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139
cbarnhart@mit.edu • belobaba@mit.edu • arodoni@mit.edu

This paper presents an overview of several important areas of operations research applications in the air transport industry. Specific areas covered are: the various stages of aircraft and crew schedule planning; revenue management, including overbooking and leg-based and network-based seat inventory management; and the planning and operations of aviation infrastructure (airports and air traffic management). For each of these areas, the paper provides a historical perspective on OR contributions, as well as a brief summary of the state of the art. It also identifies some of the main challenges for future research.

1. Introduction
During the one hundred years since the first flight of Orville and Wilbur Wright, the air transport industry has grown into a major sector of the global economy. Even more importantly, it has become essential to developing and maintaining cultural and economic links among countries and peoples. The airlines alone generated more than $300 billion in revenues in 2002, a lean year, and carried about 1.6 billion passengers, a number expected to grow at an annual rate of 4%–5% over the next 20 years according to most forecasts. According to the industry “air transport provides 28 million direct, indirect, and induced jobs worldwide” and carries “over 40% of the world trade of goods, by value” (Collaborative Forum 2003).

After spending roughly its first 40 years trying to get off the ground, literally at times, the air transport industry has grown by leaps and bounds during the last 60, especially since the advent of the “jet age” in the late 1950s. Throughout that second period, operations research (OR) has played a critical role in helping the airline industry and its infrastructure sustain high growth rates and make the transition from a novelty that catered to an elite clientele to a service industry for the masses. More than 100 airlines and air transport associations are currently represented in AGIFORS, the Airline Group of Operational Research Societies, which has been active since 1961. Indeed, it is difficult to think of any single sector, other than perhaps military operations, with which operations research has been linked more closely. One of the reasons is that airline operations and, more generally, the air transport environment provide natural contexts for the application of OR techniques and models. A second is that the airline industry has consistently been a leader in the use of information technology and has relied heavily on the intensive use of computers over the years.

The objective of this paper is to present a historical perspective on the contributions of operations research to the air transport industry, as well as to offer an assessment of some of the challenges that will be confronted next. Any reasonably thorough coverage of this subject would probably require an entire issue of this journal because the number of OR papers published on air transport easily exceeds 1,000 over the last 50 years. In view of the severe constraints on its length, the scope of the paper will instead be confined to a selected subset of air transport-related topics, where operations research has made some
of its most significant contributions to date. Examples of important topics that are either not covered at all or are touched on peripherally include: aviation safety and security, airline fleet planning, airline staffing, airline maintenance planning, aircraft loading, and decision support tools for the management of airport operations (e.g., gate assignments). Moreover, the specific topics and contributions that are highlighted are presented in nonquantitative terms and largely reflect the authors’ own interests. In addition to the bibliographic references associated with these contributions, other survey papers, which provide additional details and references, are cited whenever possible.

Section 2 of the paper deals with the classical problems of scheduling, routing, and crew assignment in the airline industry. This is a context that is perfectly suited to the use of large-scale, discrete optimization approaches and, indeed, has motivated several methodological and computational developments in this vibrant area of OR over the years. Section 3 covers airline revenue management, including overbooking, flight leg yield management and network revenue maximization. Through a combination of stochastic and optimization models, OR work in this area has generated significant additional revenues for the airlines ever since the late 1980s. Moreover, revenue management continues to be a field in which airlines are vying intensively for competitive advantage. Section 4 surveys selected applications of OR to the study, planning, and design of the two major pieces of aviation infrastructure, the airports system and the air traffic management (ATM) system. Historically the emphasis here has been on stochastic models, as the questions addressed have focused on capacity, delays, and safety under conditions in which the probabilistic characteristics of the input parameters play a dominant role. However, optimization models, both deterministic and stochastic, have found use in the intensive recent research on air traffic flow management, a topic also reviewed briefly in §4. Finally, §5 summarizes the main conclusions regarding the fundamental challenges faced by future research.

2. Aircraft and Crew Schedule Planning

Schedule planning involves designing future aircraft and crew schedules to maximize airline profitability. This problem poses daunting challenges because it is characterized by numerous complexities, including a network of flights, differing aircraft types, gate, airport slot and air traffic control restrictions, noise curfews, maintenance requirements, crew work rules, and competitive, dynamic environments in which passenger demands are uncertain and pricing strategies are complex. Not surprisingly, no single optimization model has been solved, or even formulated, to address this complex design task in its entirety. The problem’s unmanageable size and complexity has resulted in the decomposition of the overall problem into a set of subproblems, often defined as follows:

1. **Schedule design**: Defining which markets to serve and with what frequency, and how to schedule flights to meet these frequencies.

2. **Fleet assignment**: Specifying what size aircraft to assign to each flight.

3. **Aircraft maintenance routing**: Determining how to route aircraft to ensure satisfaction of maintenance requirements.

4. **Crew scheduling**: Selecting which crews to assign to each flight to minimize crew costs.

Suboptimal, yet feasible aircraft and crew plans are constructed by solving the subproblems in order, constraining the solutions to subsequent problems based on the solutions to preceding problems. Although smaller and simpler than the overall problem, these subproblems are still large-scale and rich in complexity. In fact, OR theoreticians and practitioners have been developing models and algorithms to solve them for decades and, in so doing, have had significant successes and impacts.

In the late 1960s and early 1970s, United Airlines, American Airlines, British Airways, and Air France, recognizing the competitive advantage that decision technologies could provide, formed OR groups. These groups grew rapidly, developing decision support tools for a variety of airline applications, including aircraft and crew schedule planning, and in some cases began offering their services to other airlines.
as well. According to one executive, the schedule planning system developed at American Airlines and Sabre has generated over $500 million in incremental profits annually (Cook 2000). Most major airlines around the world have formed similar groups by now. Their name and size, as well as their placement in the corporate structure, vary considerably from airline to airline. Many consulting companies are also offering products and services in this general area. Other testimonials on the impact of optimization in the airline industry can be found in Yu et al. (2003), Butchers et al. (2001), Wiper et al. (1994), Smith et al. (1992), and Patty et al. (1991).

Beginning in the 1950s and 1960s, early research on airline schedule planning focused on approaches to assign aircraft to routes (Ferguson and Dantzig 1956a) and crews to trips (Arabeyre et al. 1969). Since then, researchers have continued to work on these problems, refining and expanding the models to better represent the actual problem faced, or developing more sophisticated algorithms to improve the quality of the solutions generated.

In the following sections, we provide an overview of progress, highlighting some of the major accomplishments and describing some of the remaining challenges.

2.1. Schedule Design
The flight schedule, specifying the flight legs to be flown and the departure time of each flight leg, largely defines the competitive position of an airline and is thus a key determinant of airline profitability. Designing a profit maximizing flight schedule, however, is extremely complex. It affects and is affected by essentially all aircraft and crew scheduling decisions of the airline, and competing airlines as well. No single model has captured all these interdependencies, and even if such a model were formulated, it surely would be intractable. Moreover, its input data requirements are impractical, requiring, for example, accurate estimates of itinerary-specific passenger demands, spill costs, and recapture rates.

Notwithstanding this complexity, flight schedule design and variants of the problem have been of interest to researchers for many years, with Simpson (1966), Chan (1972), Soumis et al. (1980), and Etschmaier and Mathaisel (1984) describing early work. Nonetheless, because of the inability of optimization models to adequately capture the scope of the design problem, the typical airline practice today is to build flight schedules manually, with limited optimization. With recent research advances, however, this trend is reversing and optimization is beginning to play a role.

Success has been achieved by defining a simplified design problem involving only incremental changes to existing flight schedules. Berge (1994), Marsten et al. (1996), and Lohatepanont and Barnhart (2001) develop models and algorithms that select from a subset of candidate flights legs, those that will be added to or removed from a given (often existing) flight schedule. Their approaches are incremental in that the changes from one published flight schedule to the next are limited. The reported impacts, however, are significant. Lohatepanont and Barnhart solve problems at one major airline that contain about 800 potential flight legs, 65,000 itineraries, and 165 aircraft, resulting in formulations with 30,000–60,000 rows and 50,000–65,000 columns, and solution times ranging from 12 hours to more than 3 days. They report potential improvements in aircraft utilization and significant increases in revenue, with an estimated impact exceeding $200 million at that airline.

A nonincremental, clean-slate approach to airline schedule design is described in Armacost et al. (2002). They present models and algorithms to generate (near-) optimal flight network designs for express parcel delivery, and report savings of about 7% in operating costs and potential reductions of 10% in the required fleet size.

Another schedule design application involving a charter airline is described in Erdmann et al. (1999). By exploiting the special characteristics of the problem, they are able to achieve near-optimal solutions in minutes.

Schedule generation, with its important strategic and financial implications, represents an important area for future research, one rich in opportunity and challenge. The successes to date are just first steps in addressing the myriad of questions surrounding schedule design. Future research is needed to capture the critical interactions among the various resources of the airline, its competitors, and airports.
2.2. Fleet Assignment

With the flight schedule determined, the fleet assignment problem is to find the cost-minimizing assignment of aircraft types to legs in the flight network. Fleeting costs are comprised of:

(1) **Operating costs**: Specified for each flight leg—aircraft type pair, representing the cost of flying that flight leg with that aircraft type.

(2) **Spill costs**: Measuring the revenue lost when passenger demand for a flight leg exceeds the assigned aircraft’s seating capacity.

One of the earliest research efforts on airline scheduling was that of Ferguson and Dantzig (1956a). They developed a linear programming model to allocate aircraft to roundtrip routes to minimize operating plus spill costs. In a follow-up paper (Ferguson and Dantzig 1956a), they expanded their approach to handle uncertain demands. These first models, while applied to simplified flight networks and tiny problems by today’s standards, identified key problem attributes and launched decades of research on fleet assignment.

As a result, some 30 years later, researchers developed the capabilities to solve fleeting problems representative, both in complexity and size, of those truly faced by airlines. Abara (1989) and Hane et al. (1995) formulated the fleet assignment problem as a multicommodity network flow problem with side constraints. The underlying network (depicted in Figure 1) has:

(1) **Nodes**: Representing the times and locations of flight leg departures and arrivals.

(2) **Flight arcs and ground arcs**: Each flight arc corresponds to a flight leg with its arrival time adjusted to include the minimum amount of time required on the ground for disembarking and embarking passengers, unloading and loading baggage, and refueling. Ground arcs represent idle aircraft on the ground between flights.

The objective is to flow commodities, that is, aircraft types, through the network feasibly and with minimum cost. Side constraints enforce the following requirements and restrictions:

(1) **Cover**: Each flight leg is assigned to exactly one aircraft type.

(2) **Aircraft count**: Only available aircraft are assigned to the network.

(3) **Balance**: Each aircraft type is assigned to the same number of flight legs arriving at a station as departing that station.

Each possible assignment of an aircraft type to a flight leg is represented by a binary variable, with value 1 if the aircraft type is assigned to that flight leg, and 0 otherwise. The resulting formulation for the fleet assignment problem of a major airline contains about 20,000 rows and 30,000 columns and can be solved typically within minutes.

2.2.1. Impacts and Challenges. Multicommodity flow-based fleet assignment models are widely used by the industry and are credited with achieving significant cost savings, measuring in the millions of dollars annually. For example, Rushmeier and Kontogiorgis (1997) report realized savings of at least $15 million annually at USAir. Using fleet assignment models, Wiper et al. (1994) report annual savings of $100 million at Delta Airlines, and Abara (1989) reports a 1.4% improvement in operating margins at American Airlines.

Beyond providing economic benefits, research on fleet assignment problems has led to advanced techniques for solving general linear programs. For example, the node consolidation idea introduced in Hane et al. (1995), which reduced formulation size by more than 40% for the problems of one major airline, has been generalized and incorporated into commercial solvers, allowing more efficient solution of large-scale optimization problems.
These impressive results notwithstanding, there remain several critical challenges in fleet assignment. Many of these challenges stem from modeling assumptions that include:

1. Many fleet assignment models assume that the flight schedules repeat daily, even though most airlines operate different schedules on the weekend.
2. Most fleeting models assume flight leg demand is known and does not vary by day of week, but historical data show that day-to-day demand variations are present.
3. Flying times and ground times are typically assumed to be deterministic in fleet assignment models; however, congestion on the ground and in the air, weather conditions, and new security practices produce large variations in flight and ground times.
4. Most fleet assignment models assume that the number of spilled passengers, and their associated spill costs, can be computed at a flight-leg level. In fact, passenger demand, spill, and the revenue associated with each passenger are itinerary specific, not flight-leg specific. As a result, it is possible to estimate leg-specific spill costs only approximately.

To improve on leg-based models, researchers introduced itinerary- or origin-destination-(O-D) based fleet assignment approaches (Jacobs et al. 1999; Barnhart et al. 2002a, b). These approaches consider passenger fares, demand, spill, and recapture to be itinerary, not flight-leg specific. Inaccuracies resulting from the allocation of fares and demands to flight legs are thus not introduced into these enhanced fleeting approaches. Using their O-D-based fleet assignment approach, Jacobs et al. (1999) report annual improvements of 0.54%–0.77% in revenue compared to those obtained with a flight-leg-based model. Similarly, in a case study involving a major U.S. airline, Barnhart et al. (2002a) report that itinerary-based fleet assignment improves on leg-based fleeting significantly, with estimated savings ranging from 30 million to over 100 million dollars annually.

In a first step at integrating fleet assignment and schedule design decisions, Rexing et al. (2000) simultaneously assign an aircraft type to each flight leg and select each flight leg’s departure time, allowing retimings of 5–20 minutes from the current schedule. Departure retiming allows additional aircraft assignments that can lead to better matches between flight leg demand and assigned capacity. The result, according to Rexing et al. (2000), is reduced operating costs and improved revenue capture, with savings for one major airline of $20–$50 million annually.

2.3. Aircraft Maintenance Routing

With schedule design and fleet assignment decisions made, the flight network decomposes into subnetworks, each one associated with aircraft of a single type. The assignment of individual aircraft to flight legs in a subnetwork occurs in the aircraft maintenance routing step. The goal is to determine routings, or rotations, for each aircraft in a fleet. A routing is a sequence of flight legs, with the destination of one flight leg the same as the origin of the next leg in the sequence. A rotation is a routing that starts and ends at the same location. Each aircraft’s rotation visits maintenance stations at regular intervals. More details on the maintenance routing problem are contained in Feo and Bard (1989), Gopalan and Talluri (1998), and Clarke et al. (1996b). For restricted maintenance routings of three or four days, Gopalan and Talluri (1998) and Talluri (1998) describe graph-theoretic approaches to maintenance routing.

In general, the aircraft maintenance routing problem can be modeled as a network circulation problem with side constraints. The decision variables correspond to sequences (strings) of flight legs, with each sequence beginning and ending at maintenance stations and satisfying the rules governing the maximum time between maintenance. If a string is included in the solution, a single aircraft flies each flight in the sequence and then undergoes maintenance. Side constraints include cover constraints and count constraints. Cover constraints ensure that each flight leg is contained in exactly one selected string, and count constraints limit the number of assigned aircraft to the number available. Additional details are provided in Barnhart et al. (1998a).

2.3.1. Impacts and Challenges. Solving the fleet assignment problem first and then the resulting aircraft routing problems can lead to violations of aircraft maintenance requirements. To guarantee feasible solutions, particularly in low-frequency,
point-to-point networks, researchers have integrated fleet assignment and aircraft routing models, as in Desaulniers et al. (1997) and Barnhart et al. (1998a). This has the effect of substantially expanding the aircraft routing model to include multiple fleet types, instead of a single aircraft type. The number of constraints for large airlines is often less than a few thousand, but there are many billions of possible variables. For hub-and-spoke networks, however, the increased size and complexity of this integrated model is rarely warranted. Instead, feasible solutions are typically generated using a slightly modified sequential solution approach in which an altered fleeting model is first solved and then the routing model is solved. The modified fleet assignment model contains pseudo-maintenance constraints to ensure that sufficient numbers of aircraft of each type are located at maintenance stations periodically. In hub-and-spoke networks with banks containing many aircraft together on the ground at the same time, these maintenance constraints are often sufficient to ensure that the resulting fleet assignment has associated feasible maintenance routings.

In addition to the possibility of generating infeasible solutions, a disadvantage of the sequential solution approach to aircraft and crew planning is that aircraft routing solutions limit possible crew scheduling opportunities, potentially causing crew costs to increase significantly. The linkage between aircraft routing and crew scheduling occurs because a crewmember can connect between two flight legs separated by less than the minimum required connection time only if the same aircraft is assigned to both legs. To account for this, Klabjan et al. (2002) swap the order of the problems and solve the crew pairing problem before the maintenance routing problem. This approach has the advantage of generating optimal crew solutions, but it does not ensure that for the optimized crew solution there is a corresponding maintenance-feasible solution. To achieve crew optimality and maintenance feasibility, Cordeau et al. (2000) and Cohn and Barnhart (2003) integrate the basic maintenance routing and crew pairing models.

2.4. Crew Scheduling
Crew scheduling problems with numerous, complex rules, and well-defined costs are particularly amenable to optimization. With so many possible decisions, it is difficult to find feasible—let alone optimal—solutions manually. Moreover, crews represent the airlines’ second highest operating cost after fuel, so even slight improvements in their utilization can translate into significant savings.

For these reasons, airline crew scheduling has garnered considerable attention, with research spanning decades. Arabeyre et al. (1969) published a survey of early research activities on the topic. At that time, because of large problem size and lack of advanced techniques and sufficient computing capabilities, heuristics were employed to find improved crew solutions.

Since then, researchers have continued to work on the crew scheduling problem, seeking more efficient solution approaches, expanding models to capture more of the intricacies of the crew scheduling problem, and integrating the crew problem with other airline scheduling problems. Recent detailed descriptions of the airline crew scheduling problem are included in the survey papers by Desaulniers et al. (1998), Clarke and Smith (2000) and Barnhart et al. (2003).

Even today, the crew scheduling problem is typically broken into two sequentially solved subproblems, the crew pairing problem and the crew assignment problem:

1. The crew pairing problem. The problem generates minimum-cost, multiple-day work schedules, called pairings. Regulatory agencies and collective bargaining agreements specify the many work rules that define how flight legs can be combined to create feasible schedules. Work-rule restrictions include limits on the maximum number of hours worked in a day, the minimum number of hours of rest between work periods, and the maximum time the crew may be away from their home base. Even with these limitations, the number of feasible pairings measures in the billions for major U.S. airlines. The cost structure of pairings adds further complexity, with cost typically represented as a nonlinear function of flying time, total elapsed work time, and total time away from base.

2. The crew assignment problem. This problem combines these pairings into equitable and efficient
month-long crew schedules, called *bidlines* or *rosters*, assigning them to individual crewmembers. With rostering, schedules are constructed for and assigned to specific individuals, taking into account their particular needs and requests. With bidline generation (a practice commonly used in the United States), a generic set of schedules is constructed and individual employees reveal their relative preferences for these schedules through a bidding process. The specific assignment of schedules to employees is based on seniority, with the more senior crewmembers most often assigned their preferred schedules. More details on the bidline and rostering problems can be found in Kohl and Karisch (2003), Cappanera and Gallo (2001), Caprara et al. (1998), Christou et al. (1999), Dawid et al. (2001), Day and Ryan (1997), and Gamache et al. (1998).

The crew pairing, bidline, and rostering problems can all be cast as *set partitioning problems*. While simple in form, the set partitioning model is powerful in this context. To illustrate, consider the crew pairing problem.

The crew pairing problem has one binary decision variable for each possible pairing, and its objective is to minimize the cost of the selected pairings such that for each flight, exactly one pairing containing that flight is chosen. By defining variables as *sequences of flight legs*, only feasible pairings are considered and explicit formulation of complicated work rules is unnecessary. Moreover, nonlinear pairing costs can be computed offline and captured as a *single value*. The crew pairing model, is then a set partitioning problem, with a linear objective function and binary decision variables. The one drawback to this modeling approach is that there are potentially many billions of possible crew pairings, especially in hub-and-spoke flight networks with numerous aircraft at hubs at the same time, allowing many possible crew connections.

### 2.4.1. Solving Crew Scheduling Problems

Because of the immense size of airline crew scheduling problems, i.e., thousands of constraints and billions of variables, researchers initially focused on heuristic methods to achieve solutions. Many of these heuristics generate solutions by considering only a relatively small number of pairings. Examples of this strategy can be found in Arabeyre et al. (1969), Anbil et al. (1991), Gershkoff (1989), and Hoffman and Padberg (1993).

While producing improved solutions compared to those generated earlier, these heuristics still have a drawback; namely, the quality of their solutions cannot be quantified relative to an optimal solution. With advances in optimization theory and computing, however, researchers have built enhanced crew scheduling capabilities and designed optimization-based approaches capable of generating provably optimal or near-optimal crew pairing solutions.

To generate crew solutions with known optimality bounds, even when variable enumeration is not practical, many researchers have turned to *branch-and-price*. Branch-and-price, surveyed in Barnhart et al. (1998b), is a smart enumeration technique in which linear programming bounds are generated at each node of a branch-and-bound tree using *column generation*. Column generation allows very large LPs to be solved without explicitly enumerating all of the variables, or *columns*. Rather than directly solving the master problem (the problem with all possible columns), a restricted master problem containing only a subset of the original columns is solved. The dual solution to the restricted master problem is used to identify columns with negative reduced cost. These columns are added to the restricted master problem; the restricted master problem is then resolved, and the process is repeated until no negative reduced cost columns can be found. At that point, the algorithm terminates with an optimal solution to the master problem.

In the case of the crew pairing problem, negative reduced cost columns are identified without explicitly considering each pairing. Instead, all columns are implicitly considered by solving a pricing problem, often formulated as a multilabel, shortest path problem in which some paths correspond to pairings. By exploiting dominance, shortest paths—and hence minimum reduced cost pairings—are identified without examining all paths, or pairings. For a detailed exposition of multilabel, shortest path problems, see Desrochers and Soumis (1988).

A particular challenge in branch-and-price algorithms is that a *standard* branching rule based on
variable dichotomy is often difficult to implement. Instead, a branch-on-follow-ons branching rule is typically used (Ryan and Foster 1981). This rule is based on pairs of flights, with one flight immediately following the other. One branching decision requires schedules containing either flight to contain both flights and the other decision disallows this. For applications of this branching strategy to airline crew scheduling, see Lavoie et al. (1988), Desrosiers et al. (1991), Ryan (1992), Gamache and Soumis (1993), Barnhart et al. (1994), Vance et al. (1997), and Desaulniers et al. (1998).

Even with the effective branch-on-follow-ons rule, the number of branching decisions to prove optimality is typically excessive for crew-scheduling problems. To remedy this, researchers have embedded heuristics within the branch-and-price optimization process. An example is the variable-fixing approach described in Marsten (1994) in which variables with fractional values close to one are sequentially fixed to one. Marsten shows that by reducing the number of LPs solved in generating integer solutions, improved solutions can be generated with significantly less computing time and memory. Klabjan et al. (2001) integrate the heuristic consideration of subsets of columns with an optimization-based pricing approach. They produce improved solutions, compared to those achieved with branch-and-price, by initially developing hundreds of millions of columns and applying random selection and pricing techniques to prune the number of columns ultimately considered.

2.4.2. Impacts and Challenges. Most major airlines use optimization tools to partially or fully generate their crew schedules, with significant economic impacts. Already in the 1960s, airlines were solving crew pairing problems to reduce costs and automate their burdensome manual planning process (Arabeyre et al. 1969). With advances in optimization and computing, crew pairing solvers became increasingly sophisticated and applicable. Using advanced heuristic techniques, Anbil et al. (1991) achieved cost savings at American Airlines of $20 million dollars annually, representing an increase in crew utilization of 1.5%.

Motivated by the potential for even further cost savings, researchers focused on the development of optimization-based approaches in the decades spanning the 1980s and 1990s. Barnhart et al. (2003) reported that using a branch-and-price approach, one major U.S. airline was able to produce near-optimal crew solutions costing $50 million per year less than the solutions generated by a state-of-the-art heuristic.

In addition to these economic benefits, crew scheduling optimization can be a useful tool in contract negotiations, helping airlines to quantify the impacts of proposed changes in cost structures, benefits, and work rules. The economic impact of each proposed change can be evaluated by changing selected inputs to the crew scheduling model, and rerunning the optimization procedure.

While significant, the effects of optimization on crew scheduling are nonetheless limited by the sequential schedule planning process. With the flight schedule, fleet assignment, and aircraft routing decisions fixed, the range of crew scheduling possibilities is limited. To mitigate the myopic effects of sequential solutions, researchers have developed extended models that incorporate crew considerations into some of the other subproblems solved. For example, Clarke et al. (1996a) and Barnhart et al. (1998c) extend fleet assignment models to account for some of the downstream effects on crews. The fleet assignment model in the sequential solution process is then replaced with one of these extended models, resulting in an increase in fleet assignment costs that is offset by a greater reduction in crew costs.

2.5. Ongoing and Future Challenges
Notwithstanding the substantial progress made in solving aircraft and crew scheduling problems in recent decades, significant work remains. Research opportunities, beyond those identified for the schedule design, fleet assignment, aircraft routing, and crew scheduling problems described above, include:

Schedule planning: (1) Integrating decisions involving schedule design and aircraft and crew routing and scheduling; (2) expanding schedule planning models to include pricing and revenue management decisions; (3) assessing the systemwide cost and service impacts of paradigm shifts in airline scheduling,
such as depeaking flight schedules by spreading out aircraft arrival and departure banks; and (4) developing tools, such as the stochastic model and airline operations simulator described in Rosenberger et al. (2002), to evaluate airline plans and recovery policies in a stochastic operating environment.

Robust planning: Optimized solutions are rarely executed as planned. Crew sickness, mechanical failures, and adverse weather result in necessary changes to the plan, often leading to significantly increased costs. Moreover, a finely tuned, optimized solution achieves increased utilization through the removal of slack, providing crews with less time to connect between flights and aircraft with less time on the ground between flying. Less slack time, although economical in theory, can translate in practice into less robustness and increased costs. To address these issues, researchers have begun to investigate a different planning paradigm, one that considers unplanned, disruptive events and attempts to minimize realized, and not planned, costs. Current research, such as that of Ageeva (2000), Rosenberger et al. (2001a), Schaefer et al. (2001), and Chebalov and Klabjan (2001), attempts to achieve robustness by isolating causes of disruption and/or downstream effects. In Lan (2003), robust aircraft routes are generated to place slack judiciously, that is, where it is needed to minimize the disruptive effects of delays on passengers.

Operations recovery: Given a plan and one or more disruptions, the goal of operations recovery is to return the airline to its plan quickly (often requiring decisions within minutes) and cost effectively. Detailed descriptions of recovery models and algorithms are included in Jarrah et al. (1993), Clarke (1997), Thengvall et al. (2000), Rosenberger et al. (2001b), Stojkovic et al. (2002), Bratu (2003), and Yu et al. (2003). Although these problems typically involve multiple airports and many of the airline’s resources, most of the work to date focuses on recovering a particular resource, either aircraft, crews, or passengers. For example, in their Edelman Award-winning work, Yu et al. (2003) report that Continental Airlines saved approximately $40 million in 2001 using their crew recovery system. The challenge moving forward is to enhance current capabilities by considering integrated passenger, crew, and aircraft recovery decisions, while rapidly determining low-cost solutions.

3. Airline Revenue Management
Even with an optimized fleet assignment and schedule of operations, some flight departures will have empty seats while others will experience more passenger demand than capacity. In an effort to better match the demand for each flight with its capacity and to increase total revenues, airlines practice differential pricing by offering a variety of fare products at different price levels for the same flight. Revenue management is the practice of determining the number of seats on each flight to be made available at each fare level, limiting low-fare seats and protecting seats for later-booking, higher-fare passengers. Given that the operating costs of a scheduled flight departure are in large part fixed in the very short run, the goal of revenue management is to fill each flight with the maximum possible revenue to maximize operating profit.

This section provides a brief review of the role of operations research in the development of airline revenue management (RM) models, with an emphasis on the works that have most influenced the state of the practice in the industry. A much more comprehensive survey of OR literature dealing with revenue management and related problems can be found in McGill and van Ryzin (1999). In addition, Weatherford and Bodily (1992) developed a categorization of “perishable asset revenue management” problems, of which the airline revenue management problem is the best-known example.

This review begins with an introduction to the functions of typical airline RM systems, followed by descriptions of the types of OR models employed to perform three of the principal techniques of revenue management—overbooking, fare class mix, and O-D control. Therefore, the focus of this discussion is on the seat inventory control component of airline revenue maximization, as virtually all airline RM systems assume that the fare structure is determined exogenously by a separate airline pricing function.

3.1. The Evolution of Airline RM Systems
The sheer size and complexity of the airline seat inventory control problem led to the development
of computerized RM systems. A medium-sized airline might operate 1,000 flight legs per day, using 10 booking (fare) classes in its reservations system, and accepting bookings up to 330 days prior to each departure. At any point in time, this airline’s seat inventory includes over three million booking limits, which can change with each booking that is accepted.

Airline RM systems have evolved in both their computer database and mathematical modeling capabilities over the past 20 years. The first RM systems, developed in the early 1980s, were designed to collect and store data extracts from computer reservations systems (CRS). By the mid-1980s, several RM systems offered additional monitoring capabilities that allowed actual flight bookings to be tracked dynamically, relative to an expected or "threshold" booking curve for the flight. By the late 1980s, the more advanced airlines began to implement RM systems that could perform forecasting and optimization by booking class for each future flight leg departure, in addition to having all of the database and booking monitoring capabilities of previous systems. It was into this “third generation” of RM systems that OR models, some of which had been developed a decade or more earlier, began to be integrated.

The major components of a typical third-generation RM system are illustrated in Figure 2. Historical booking data for the same flight leg and day of the week are combined with actual booking information for each future flight departure to generate a forecast of total demand by booking class for that departure. These forecasts, together with estimates of the revenue value of each booking class, are then fed into an optimization model that calculates the recommended booking limits for the flight departure in question. At the same time, the demand forecasts are fed into an overbooking model, which also makes use of historical information about passenger no-show rates to calculate an optimal overbooking level for the future flight departure. Both the booking class limits and overbooking levels calculated by the mathematical models are then presented as recommendations to the RM analyst.

The demand forecasts and booking limits are reviewed by the RM systems at regular intervals during the flight booking process, as often as daily in some cases. Should unexpected booking activity occur, the system reforecasts demand and reoptimizes its booking limit recommendations. A substantial proportion of the revenue gain attributable to fare mix optimization comes from this dynamic revision of booking limits.

Most large and medium-sized airlines throughout the world have implemented third-generation RM systems, and the benefits of such systems have been well documented. Effective use of techniques for overbooking and fare class mix alone have been estimated to generate revenue increases of as much as 4%-6% compared to situations in which no seat inventory control tools were applied (Belobaba 1992b, Smith et al. 1992).

3.2 Overbooking Models

Airlines have been accepting reservations in excess of aircraft capacity for more than two decades, in an effort to reduce the revenue losses associated with no-shows. With the development of RM systems, overbooking was incorporated into the seat inventory control functions of these systems. The objective of the flight overbooking component of revenue management is to determine the maximum number of bookings to accept for any given future flight departure, trading off the risks and costs of denied boardings against the potential revenue loss from unsold or spoiled seats.

Operations research work on the airline overbooking problem can be traced to the 1960s, well before
overbooking became integrated into RM systems (and before RM systems were even required). Notable early published works include those of Simon (1968) and Vickrey (1972), as well as articles by Rothstein (1971, 1985). Rothstein’s 1985 article is a survey of previous OR literature dealing with the airline overbooking problem, and includes a discussion of the customer service impacts of inaccurate overbooking and the role of government in regulating denied boarding penalties.

Statistical overbooking models typically represent no-show rates as Gaussian random variables. The objective is to find the maximum authorized number of bookings (or “AU”) that will keep denied boardings below some airline-specified target level with a desired level of confidence. These models provide airline managers with flexibility in determining their own overbooking policy, for example, by increasing denied boarding tolerance or reducing statistical confidence.

An extension of the statistical overbooking approach is the cost-based overbooking model, which explicitly accounts for the actual costs associated with denied boardings and with empty seats (“spoilage”). The objective is to find the optimal overbooking level that minimizes the total combined costs of denied boardings and spoilage by performing a search over a reasonable range of AU values. The cost-based overbooking model is the current state of the practice at many airlines. However, this approach represents a static formulation of the overbooking problem, in that the dynamics of passenger bookings, cancellations, and no-shows are not explicitly accounted for in determining an overbooking level.

The OR literature contains many additional works on the airline overbooking problem, some of which propose dynamic programming (DP) formulations. Rothstein’s (1968) Ph.D. thesis was the first to describe such a DP approach, while Alstrup et al. (1986) extended the DP formulation to a two-class, joint overbooking and fare class mix problem. More recently, Chatwin (1996) as well as Feng and Xiao (2001) have proposed DP-based approaches that allow for the incorporation of time-dependent no-show and cancellation rates associated with multiple fare classes. In practice, few airlines have implemented such complex DP formulations because of the difficulties of providing adequate and accurate inputs in the form of booking and cancellation rates by the time-period before departure.

The economic motivation for airline overbooking is substantial. In the United States, domestic airline no-show rates average 15%–25% of final predeparture bookings. Given that most airlines struggle to attain a 5% operating margin (revenues over costs), the loss of 15%–25% of potential revenues on fully booked flights (which would occur without overbooking) can represent a major negative impact on profits. As part of a revenue management system, effective overbooking has been shown to generate as much revenue gain as optimal fare class seat allocation, described below.

3.3. Fare Class Mix
The second major technique of airline revenue management is the determination of the revenue-maximizing mix of seats available to each booking (fare) class on each future flight-leg departure. Virtually all airline RM systems were developed with the capability to optimize fare class mix as their primary objective. As introduced earlier, RM systems forecast the expected demand for each fare class on each future flight-leg departure by applying statistical models to historical booking data for the same fare class on previous departures of the same flight. The forecasting of expected booking demands for future flight departures has been addressed in many OR papers, including Littlewood (1972), L’Heureux (1986), Lee (1990), and Curry (1994).

These demand forecasts are then used as inputs to a seat allocation optimization model, which determines booking limits to be applied to each of the booking classes on the flight departure in question. The vast majority of airline reservations systems now have inventory structures based on serial nesting of booking classes, as shown in Figure 3. Seats are not “allocated” to partitioned classes, but are instead “protected” for higher fare classes and nested booking limits are applied to the lower fare classes. All available seats in the shared inventory are made available to the highest booking class, in the (unlikely) event that the entire capacity of the aircraft can be filled with demand for the highest-priced fare product. This ensures that the
airline would never turn down a high fare request, as long as there are any seats still available for the flight.

OR work on “seat allocation” problems for two fare classes (full-fare and discount fare) can be traced to Littlewood (1972). Littlewood’s (1972) rule for protecting full-fare seats was extended to multiple nested fare classes by Belobaba (1987, 1989), who wrote the first doctoral dissertation on airline pricing and revenue management. The Expected Marginal Seat Revenue (EMSR) approach for setting RM booking limits was then refined to become the “EMSRb” model (Belobaba 1992a). These models were the first to recognize that the optimality conditions for traditional resource (seat) allocation problems did not result in the maximum expected revenues in a multiple class, nested booking limit environment. Both models are based on heuristic decision rules for nested booking classes and have become the most commonly used seat inventory control models in airline RM systems. Optimal formulations of the multiple nested class problem have also been published by Curry (1990), Brumelle et al. (1990), and Wollmer (1992).

The general premise of these fare class mix models is as follows. Given the forecast demand for each booking class, expressed in terms of a mean and standard deviation, along with its associated average fare, the expected marginal revenue of each incremental seat on a flight leg can be determined. It is equal to the average fare of the booking class under consideration multiplied by the probability that demand will materialize for that incremental seat. The optimal protection level for a higher-class seat is equal to the number of seats with an expected marginal seat revenue greater than or equal to the average fare in the next lower class. Because higher fare classes have access to unused lower class seats in a nested booking class inventory structure, the problem is to find seat protection levels for higher classes, and booking limits on lower classes. Simulations and actual airline experience have indicated that use of these decision models for establishing the fare class mix of seats on each flight departure can increase total airline revenues by 2%–4% (Belobaba 1989).

Recent research on the single-leg fare class mix problem has been rather limited, as both researchers and practitioners have focused on network optimization techniques for airline revenue management instead, as described below. The most recent published works involve extensions that include the application of DP methods and the joint optimization of fare class booking limits and overbooking levels (for example, Zhao and Zheng 2001), as well as the joint optimization of the airline pricing and fare class mix problems, as described in §3.5.

3.4. Network Revenue Management: Origin-Destination Control

Network RM (or O-D Control) represents a major step beyond the fare class mix capabilities of most third-generation RM systems, and is currently being pursued by the largest and most advanced airlines in the world. As its name implies, O-D control gives the airline the capability to manage its seat inventory by the revenue value of the passenger’s origin-destination itinerary on the airline’s network, not simply according to the fare class requested on a single flight leg. Because a leg-based RM system cannot distinguish among itineraries in the same fare class, optimizing fare class mix on each flight leg individually will not ensure that total network revenues are being maximized. This is especially true for the large connecting hub networks operated by many airlines, in which a substantial proportion of passenger itineraries involve multiple flight legs and a connection at the hub.

As a first step in the implementation of O-D control strategies, American Airlines in the 1980s developed
what has come to be known as “virtual nesting” (Smith et al. 1992), an assignment process that maps each itinerary/fare type to a hidden or “virtual” value class within the airline’s own reservations system. For an itinerary/fare assigned to a given virtual class, the seat availability on a given flight leg is established by the booking limit for that virtual class. Initial implementations of virtual nesting were based on total itinerary fare values, giving preference to longer-haul connecting passengers with higher total fares. The shortcoming of this “greedy” approach was that it did not address the need under certain circumstances to give preference to two “local” passengers instead of taking the connecting passenger.

OR models have been instrumental in the subsequent development of O-D control strategies that allowed airlines to manage seat inventories on the basis of the network revenue value of each itinerary/fare combination. Network revenue value can be defined as the total itinerary fare (ticket value) minus the revenue displacement that might occur on connecting flight legs if the passenger’s request for a multiple-leg itinerary is accepted. For example, the network revenue value of an $800 total fare on the first leg of a connecting itinerary must be reduced by $300 if acceptance of this passenger would result in displacement of $300 in total network revenue on the second flight leg. The displacement cost associated with the last (lowest valued) available seat on a given flight leg is its marginal value to the total network revenue—the amount by which the total network revenue would decrease if we were to “remove” (or sell) one seat on the flight leg. If the flight leg consistently has empty seats, the marginal network value of its last seat would be zero. In the more realistic situation where the flight leg has a significant probability of being full and it carries both local and connecting passengers, determining the impact on total network revenues of removing a seat can require substantially more complicated analysis, including network optimization models.

Various network optimization and heuristic algorithms have been applied to the problem of determining the network revenue value contribution of each O-D itinerary and fare product (or ODIF). As described in Belobaba (1998), several large airlines make use of leg-based EMSR values to approximate the network revenue value associated with a given connecting itinerary/fare combination. Because these approaches do not make use of any formal network optimization models, they are referred to as “leg-based heuristic” models for estimating network revenue displacement.

The majority of published works devoted to the O-D seat inventory control problem have proposed mathematical programming approaches to determine optimal seat allocations for every ODIF over a network of connecting flight legs (see, for example, Glover et al. 1982, Dror et al. 1988, and Curry 1990). Williamson’s (1992) doctoral thesis provides a comprehensive review of the literature relevant to the application of mathematical programming and network flow models to the O-D seat inventory control problem. More recently, there have been new works proposing more sophisticated network models. For example, Bratu (1998) developed a “probabilistic network convergence” algorithm for estimating the marginal network value of the last available seat on each leg in an airline network. Talluri and van Ryzin (1999a) propose a randomized linear programming method, while stochastic dynamic programming is also being pursued as an alternative algorithm for determining the marginal network value of a range of remaining seats on each flight leg to account for changes to the marginal value as bookings are accepted.

Given an estimate of down-line revenue displacement it is possible to map ODIFs to virtual classes based on their estimated network revenue value. Displacement adjusted virtual nesting (DAVN) increases availability to connecting passengers, while adjustment for down-line displacement of revenues ensures that two local passengers (with a higher total revenue) will receive preference when two connecting flights are expected to be heavily booked.

The estimated marginal network revenue of the lowest-valued available seat on a flight leg can also be used in a much simpler inventory control mechanism based on bid price controls. At the time of an ODIF request for seat availability, the ODIF fare is evaluated against the sum of the leg bid prices over the itinerary being requested. As was the case for estimates of revenue displacement associated with connecting passengers, leg bid prices can similarly be
calculated either with network optimization tools or leg-based heuristics such as EMSR approaches. Talluri and van Ryzin (1999b) have done much work on network bid price control and identify the conditions under which this approach provides revenue optimality, while de Boer et al. (2002) compare the performance of deterministic and stochastic network formulations for O-D control.

Until recently, relatively few airlines had implemented network optimization models for dynamic calculation of displacement costs and/or bid prices for O-D control. Because most reservations systems and, in turn, third-generation RM systems were developed on the basis of leg/fare class data, most airlines did not have access to the detailed historical ODIF booking data required by network optimization models. Use of large-scale network optimization models also raised technical and computational issues related to the solution times and frequency of reoptimization. However, with the development of airline databases designed to capture detailed ODIF historical data, along with advances in both solution algorithms and computational speeds, network revenue management has been implemented by over a dozen airlines in different parts of the world.

The benefits of leg-based revenue management and incremental benefit of O-D controls over leg-based fare class controls have been estimated by several researchers through simulation. For example, Williamson (1992) developed a network revenue management simulation approach that allowed different schemes for optimization and control of seat inventories to be tested. An even more realistic approach to simulating the impacts of different RM schemes in a full-scale, competitive airline network environment is that of the passenger origin-destination simulator (PODS). Developed originally by researchers at Boeing (Hopperstad 1997), PODS has been enhanced to realistically simulate large networks in which competing airlines generate RM forecasts and set seat inventory controls based on “historical” (i.e., previously simulated) data. At the same time, the simulated passengers in PODS choose among alternative airlines, fares, restrictions, schedules, and seat inventory availability as established by each airline’s own RM system. PODS has been used to simulate the competitive impacts of RM (Belobaba and Wilson 1997), as well as the benefits of improved forecasting models and the impacts of RM on airline alliances.

The simulations cited here along with others performed by academics and airlines have provided consistent estimates of the potential for revenue gains of 1%–2% from advanced network revenue management methods, above and beyond the 4%–6% gains realized from conventional leg-based fare class control. The potential to realize even 1% in additional revenue through network RM is substantial enough that many of the world’s largest airlines have implemented or are in the process of developing their O-D control capabilities. For a large airline with annual revenues of $5 to $10 billion or more, successful implementation of a network RM system can lead to total revenue increases of $50 to $100 million per year.

3.5. Future Challenges

Development of OR models for the “next generation” of airline revenue management is currently an extremely popular topic among academics and practitioners alike. The most obvious next steps in the further enhancement of airline revenue management systems is to integrate the pricing and seat inventory control decisions currently being made with different decision support tools and, at many airlines, in different parts of the organization. Clearly, the ability to relax the traditional RM assumption that fare structures are given and fixed has the potential to further increase the revenue gains of RM. Joint pricing and inventory optimization requires the incorporation of passenger choice and demand elasticity models, and promising OR work in this direction has been published by Weatherford (1997), Gallego and van Ryzin (1997), and Cote et al. (2003), among others. In a recent Ph.D. dissertation de Boer (2003) examines this problem in a network context and presents a variety of other modeling advances and insights.

Looking ahead, it is apparent that information about the utilization of seat inventories and the response of passenger demand to different pricing strategies can and should provide useful feedback to fleet assignment and even scheduling of airline flight departure times. The integration of airline pricing and seat inventory decisions with those of the
scheduling and fleet assignment functions is therefore considered to be the “ultimate” challenge for airline operations research. Development of more responsive and even dynamic decision support systems that take into account expected passenger choice behavior, as well as the expected response of competitors in terms of price, schedule, and capacity in determining a profit-maximizing strategy for the airline, will only become more important with growing competitive pressures in the airline industry.

4. Applications to Aviation Infrastructure

The infrastructure of the global aviation system consists of two principal elements, airports and air traffic management (ATM) systems. Airports can be further subdivided into airside facilities (runways, taxiways, aprons, aircraft stands) and landside facilities (passenger and cargo buildings, curbside), while ATM systems are now viewed as being comprised of a tactical subsystem—air traffic control (ATC)—and a strategic one—air traffic flow management (ATFM). The design, development, and operation of all these facilities and systems has attracted extensive interest on the part of operations researchers, usually in response to ongoing developments in the field. For example, much of the fundamental work on airside capacity was performed during the 1960s and early 1970s, the time when it was first realized that runways constituted an important bottleneck of the air transport system. Overall, the body of work on aviation infrastructure has led to insights and models that have proved of critical importance in practice and have, in some cases, been adopted by airport and ATM service providers on a global scale. Because of space limitations, this section briefly reviews OR applications in airport airside operations and air traffic flow management—only two of the four major areas identified above. Surveys of OR models for the analysis of passenger terminal operations can be found in Tosic (1992) and de Neufville and Odoni (2003). Of the many OR-related topics addressed by research on air traffic control, the widely investigated subject of detecting and resolving potential “conflicts” between airborne aircraft is reviewed well in Kuchar and Yang (2000). Various other analytical and simulation models on several different aspects of ATC are covered in Odoni et al. (1997).

4.1. Airside Operations

The runway complexes of major airports are among the scarcest resources of today’s international air transport system and, barring a drastic change in the landing and takeoff requirements of commercial aircraft, will continue to be so in the foreseeable future. New runways are very expensive to build, require great expanses of land, and most importantly have environmental and other impacts that necessitate long and complicated approval processes with uncertain outcomes. It is not surprising therefore that one of the most “mature” areas of transportation science deals with the modeling of runway operations and, more generally, airside operations. The products of this work include both analytical (“mathematical”) models and simulation tools.

4.1.1. Analytical Capacity and Delay Models.

Analytical models preceded viable simulation tools by about 20 years. In a landmark paper, Blumstein (1959) defined the capacity of a runway as the expected number of movements (landings and takeoffs) that can be performed per unit of time—typically one hour—in the presence of continuous demand and without violating air traffic control separation requirements. He also presented a model for computing the capacity of single runways used for arrivals only, for departures only, and for strings of arrivals followed by strings of departures. Subsequent generalizations included the possibility of inserting departures between successive arrivals, possibly by increasing (“stretching”) the separation between arrivals (Hockaday and Kanafani 1972) and the treatment of some of the parameters of Blumstein’s (1959) models as random variables, instead of constants (Odoni 1972).

Extensions to cases involving two or more simultaneously operating runways were also developed at an early stage—see, e.g., Swedish (1981). The complexity of multirunway models depends greatly on the extent to which operations on different runways inter-
The queueing behavior of an airport with
Airports, at the time among the world’s busiest—that
dates drawn from New York’s Kennedy and LaGuardia
Koopman (1972) argued—and showed through exam-
port delays analytically. In another landmark paper,
cal approaches to the problem of computing air-
frequently when weather conditions are less than optimal.
“worst-case” and “best-case” estimates, respectively.
be bounded by the characteristics of the
Airports vary strongly over the course of a typical
demand rates, as well as in the capacity to the really interesting cases. The
zero one or two runways. Such models have proved
airports planning, as well as in
in assessing the impacts of proposed procedural or tech-
Another topic of intensive study has been the esti-
through the use of queueing models, of the
delays caused by the lack of sufficient runway capac-
This is a problem that poses a serious challenge to
operations researchers: The closed-form results devel-
the voluminous literature of classical steady-
queueing theory are largely nonapplicable—at
least when it comes to the really interesting cases. The
reason is that airport queues are, in general, strongly
nonstationary. The demand rates and, in changing
weather conditions, the service rates at most major
airports vary strongly over the course of a typical
day. Moreover, the demand rates may exceed capacity
(\( \rho > 1 \)), possibly for extended periods of time, most
often when weather conditions are less than optimal.

This has motivated the development of numerical
approaches to the problem of computing airport
delays analytically. In another landmark paper,
Koopman (1972) argued—and showed through exam-
plots drawn from New York’s Kennedy and LaGuardia
Airports, at the time among the world’s busiest—that
the queueing behavior of an airport with \( k \) “runway
equivalents” (i.e., \( k \) nearly independent servers) can
be bounded by the characteristics of the
\( M(t)/M(t)/k \) and the \( M(t)/D(t)/k \) queueing models, each providing
“worst-case” and “best-case” estimates, respectively.

Note that this allows for dynamic changes in the ser-
dvice rates, as well as in the demand rates.

Extending the work of Koopman (1972), the
\( M(t)/E_\lambda(t)/k \) system was proposed by Kivestu (1976)
as a model that could be used to directly com-
pute approximate queueing statistics for airports—
rather than separately solving the \( M(t)/M(t)/k \) and
\( M(t)/D(t)/k \) models and then somehow interpolating
their results. (Note that negative exponential service
times (\( M \)) and constant service times (\( D \)) are sim-
ply special cases of the Erlang (\( E_k \)) family, with \( k = 1 \)
and \( k = \infty \), respectively.) Kivestu (1976) noted that \( k \)
should be determined from the relationship
\( E[S]/\sigma_S = \sqrt{k} \), where \( E[S] \) and \( \sigma_S \) denote the expected value
and the standard deviation of the service times and
can be estimated from field data. He also devel-
oped a powerful numerical approximation scheme
that computes the (time varying) state probabilities
for the \( M(t)/E_\lambda(t)/k \) system efficiently. Malone (1995)
has demonstrated the accuracy and practicality of
Kivestu’s (1976) approach and developed additional
efficient approximation methods, well suited to the
analysis of dynamic airfield queues. Fan and Odoni
(2002) provide a description of the application of
Kivestu’s (1976) model to a study of the gridlock con-
ditions that prevailed at LaGuardia Airport in 2000

Additional (numerical) analytical models for com-
puting airport delays have been developed over the
last few years. Peterson et al. (1995) and Daniel
(1995) describe two different models for computing
delays at hub airports, which are characterized by
sharp “banks” or “waves” of arrivals and departures.
Hansen (2002) has used a deterministic model, based
on the notion of cumulative diagrams, to compute
delay externalities at Los Angeles International Air-
port. Finally, Long et al. (1999) and Malone (1995)
present two dynamic queueing network models and
their application to the study of congestion in the
National Airspace System. Ingolfsson et al. (2002)
offer a comprehensive survey and comparison of sev-
eral alternative approaches to the analysis of nonsta-
tionary queueing systems.

Many of the best features of some of the analyti-
cal capacity and delay models just described have
been integrated recently in a number of new software
packages (Long et al. 1999, Stamatopoulos et al. 2003, EUROCONTROL 2001) that perform both capacity and delay analyses and, in the instance of the last two references, include models of aprons and aircraft stands. After some necessary enhancements and an adequate amount of testing in a variety of airport environments, packages of this type may provide airport planners within a few years with an easy-to-use and very fast set of tools for the study of a host of airside issues.

### 4.1.2. Airside Simulations

General-purpose simulation models of airside operations first became viable in the early 1980s and have been vested with increasingly sophisticated features since then. Three models currently dominate this field internationally: SIMMOD, The Airport Machine, and the Total Airport and Airspace Modeler (TAAM). A report by Odoni et al. (1997) contains detailed reviews (somewhat out-of-date by now) of these and several other airport and airspace simulation models and assesses the strengths and weaknesses of each. At their current state of development (and in the hands of expert users), they can be powerful tools in studying detailed airside design issues, such as figuring out the best way to remove an airside bottleneck or estimating the amount by which the capacity of an airport is reduced due to the crossing of active runways by taxiing aircraft.

Unfortunately, these models are frequently misused in practice, at great cost to the client organization. This happens when they are applied to the study of “macroscopic” issues that can have only approximate answers because of the uncertainty inherent in the input data. An example is a question that often confronts airport operators: When will airside delays reach a level that will require a major expansion of an airport’s capacity (e.g., through the construction of a new runway)? Questions of this type, often requiring a look far into the future, are best answered through the approximate analytical models surveyed earlier, which permit easy exploration of a large number of alternative scenarios and hypotheses. Detailed simulation models, by contrast, cannot cope well with the massive uncertainty involved because they require inputs that are difficult to produce (e.g., a detailed schedule of aircraft movements at the airport for a typical day 10 or 15 years hence) and lack credibility under the circumstances.

#### 4.1.3. Optimizing Airside Operations

The airside models discussed so far are descriptive in nature. Their objective is to help users understand and predict the operational characteristics of the various airside facilities under different operating scenarios. A considerable amount of OR work with an optimization focus also exists, much of it concerned with the effective use of runway systems.

The capacity of a runway is largely determined by the separation requirements specified by the providers of ATM services (e.g., the FAA in the United States). For any pair of consecutive runway operations these requirements depend on the type of aircraft involved. For example, in the United States, when an arriving “heavy” (H) aircraft—defined as one with a maximum takeoff weight (MTOW) greater than 255,000 lbs—is immediately followed by an arriving “small” (S) aircraft (MTOW < 41,000 lbs), the required separation between them, at the instant when H is about to touch down on the runway, is 6 nautical miles (∼10.9 km). This is because “heavy” aircraft (wide-body jets) may generate severe wake turbulence, which may be hazardous to other aircraft behind it. By contrast, when an aircraft of type S is followed by one of type H, the required separation is 2.5 nautical miles (∼4.5 km). Note that given a number $n$ of aircraft, all waiting to land on a runway, the problem of determining the sequence of landings, such that the time when the last aircraft lands is minimized, is a Hamiltonian path problem with $n$ points.

However, this is only a static version of a problem which in truth is a dynamic one: Over time the pool of aircraft available to land changes, as some aircraft reach the runway while new aircraft join the arrivals queue. Moreover, minimizing the “latest landing time” (or maximizing “throughput”) should not necessarily be the objective of optimal sequencing. Many alternative objective functions, such as minimizing the average waiting time per passenger, are just as reasonable. A further complication is that the very idea of “sequencing” runs counter to the traditional adherence of ATM systems to a first-come, first-served (FCFS) discipline. Deviations from FCFS raise concerns among some airside users about the
possibility of systematic discrimination against certain classes of aircraft operators (e.g., general aviation) when it comes to runway access. In a dynamic environment, this may even result in a compromise of safety, if some aircraft are indefinitely relegated to the end of the queue as new aircraft show up to land.

These observations have led many investigators to study the runway-sequencing problem with the objective of increasing operating efficiency while ensuring that all airport users are treated equitably. Dear (1976) and Dear and Sherif (1991) developed the concept of constrained position shifting (CPS), i.e., of a limit in the number of positions by which an aircraft can deviate from its FCFS position in a queue. For instance, an aircraft in the 16th position in a FCFS queue would have to land in one of the positions 14–18 if the specified maximum position shift (MPS) is 2. Through many numerical examples and for several reasonable objective functions, Dear (1976) showed that by setting MPS to a small number, such as two or three, one can obtain most of the benefits of an unconstrained optimized system (e.g., 60%–80% of the potential improvements). This finding motivated several researchers (e.g., Psaraftis 1980, Venkatakrishnan et al. 1992, Bianco et al. 2001) to investigate a number of increasingly complex and realistic versions of the sequencing problem. Two advanced terminal airspace automation systems, CTAS and COMPAS, that have been implemented in the United States and in Germany, respectively, incorporate sequencing algorithms based on CPS (Erzberger 1995). However, this feature of CTAS and of COMPAS has not been activated, primarily because of concerns about a potential increase in controller workload.

Gilbo (1993) and Hall (1999) have gone beyond the sequencing of arrivals only by considering how available capacity can best be allocated in a dynamic way between landings and takeoffs to account for the distinct peaking patterns in the arrival and departure streams at airports over the course of a day. Pujet et al. (1999) have further examined the issue of optimizing the number of aircraft taxiing out during periods of congestion, based on the empirical observation that departure rates at major airports seem to decrease when the number of active aircraft on the taxiway system exceeds a certain airport-specific threshold. Although still at the theoretical stage, some of these promising ideas will eventually find their way into practice.

4.2. Air Traffic Flow Management

The most advanced OR work on aviation infrastructure to date is undoubtedly associated with air traffic flow management (ATFM). ATFM took on major importance in the United States and Europe during the 1980s, when rapid traffic growth made it necessary to adopt a more strategic perspective on ATM. Rather than addressing congestion through local measures (e.g., by holding arriving aircraft in the airspace near delay-prone airports) the goal of ATFM is to prevent local system overloading by dynamically adjusting the flows of aircraft on a national or regional basis. It develops flow plans that attempt to dynamically match traffic demand with available capacity over longer time horizons, typically extending from 3–12 hours in the future. The prototypical application of ATFM is in ground holding, i.e., in intentionally delaying an aircraft’s takeoff for a specified amount of time to avoid airborne delays and excessive controller workload later on. Other ATFM tactics include rerouting of aircraft and metering (controlling the rate) of traffic flows through specified spatial boundaries in airspace.

An important difference in the nature of the ATFM problem in the United States and in Europe should also be noted. In the United States, ATFM is primarily driven by airport capacity constraints, whereas in Europe en route airspace acts as the principal “bottleneck.” Europe’s Central Flow Management Unit, located in Brussels, currently determines (heuristically) ground delays to ensure that no en route sector capacity constraints are violated. This difference may, however, become moot in the near future due to continuing progress in increasing en route airspace capacity in Europe.

OR model development related to ATFM can be viewed as going through two distinct stages. The first stage involved problem definition and development of large-scale mathematical optimization models of an aggregate scope. Attwool (1977) was the first to cast ATFM issues in mathematical terms, while
Odoni’s (1987) detailed description of the single-airport ground holding problem (GHP) as a dynamic and stochastic optimization problem stimulated much of the subsequent work. Important advances in modeling and solving the GHP are marked by the stochastic programming models of Richetta and Odoni (1993), the extension to a multiairport setting by Vranas et al. (1994), and the inclusion of en route constraints and rerouting options by Bertsimas and Stock (1998). Many other interesting papers on various aspects of optimizing ATFM and GHP appeared in the 1990s. Good reviews of the literature and of computational results can be found in Andreatta et al. (1993) and Hoffman and Ball (2000).

The one common characteristic of the models developed in this first stage is the implicit assumption of a single decision-making authority attempting to optimize a “global” objective function: The providers of ATFM services (e.g., the FAA in the United States, Eurocontrol in Europe) are responsible for the allocation of ground holding delays among individual flights and/or for the rerouting, if necessary, of flights. The objective is to optimize in the aggregate, e.g., by minimizing the overall direct operating costs associated with ground holding and rerouting decisions, summed over all airlines and aircraft. This, however, is an operating philosophy that airlines strongly disagree with. They correctly argue that only individual airlines have the information necessary to make decisions on what is best for their own flights. As an obvious example, airline A, faced with a period of delays at a given congested airport, may assign very high priority to the timely arrival of one of its flights, X, because that flight may be carrying many business-class passengers who will be connecting to other flights or because it carries crews for subsequent flights departing from that airport. The assignment of priority to flight X has, in fact, little to do with direct operating costs of aircraft and is based on the business model of airline A and on information that only A possesses.

In response to such airline concerns, as well as to various complaints about the limitations of the ATFM system during the 1980s and 1990s, the FAA has been engaged for the past 10 years in developing the Collaborative Decision-Making (CDM) Program. After an initial planning period of about five years, CDM was fielded for the first time in 1998 in connection with the FAA’s Ground Delay Programs (GDPs), which go into effect whenever long air traffic delays are anticipated at an airport due to poor weather or other reasons, thus often necessitating ground holding. CDM marks a truly fundamental innovation in the ATM system, possibly the most important one in at least 30 years. The three main elements on which it is based are: (a) a dedicated data communications network (“CDM-net”), which facilitates the continuous exchange of information between the FAA and the airlines (plus any other CDM participants) about the current and near-future states of the ATM system; (b) the use of a common database and a common set of software tools by all CDM participants; and (c) the partial decentralization of decision making. With respect to (c), it is the FAA’s responsibility to forecast the capacity that will be available at each part of the ATM system during the relevant time horizon, as well as to allocate this capacity among the individual airlines and the other ATM system users. And it is the responsibility of each individual airline to decide how it will use its allocated share of capacity at each part of the system. This is a somewhat simplistic description of what, in practice, is a complicated process that employs several types of distributed decision-making techniques, such as rationing by schedule (RBS) and schedule compression—see Wambgsanss (1996) and Vossen et al. (2003) for details.

The driver for the adoption and implementation of the CDM concept was a small, OR-minded team in the FAA, the U.S. Department of Transportation, and especially the Metron Corporation (Chang et al. 2001). The CDM Program has already led to major reductions in delays and missed connections for air travelers and to documented savings of hundreds of millions of dollars in airline operating costs. ATFM-related OR research has concurrently shifted away from large-scale, aggregate optimization models and toward “real-time” decision support tools that assist air traffic managers in the FAA and Airline Operations Centers in taking maximum advantage of the massive, up-to-date information base that CDM has made available. It is important to note, however, that many of the ideas and formulations developed in the
“pre-CDM” models can still be adapted to the CDM environment, often with little modification. For example, one of the most critical problems in the planning of GDPs continues to be the determination of airport acceptance rates (AAR) for several hours into the future and in the presence of uncertainty about airport capacity and air traffic demand. The efficient stochastic integer program developed for this purpose by Ball et al. (2003) can be viewed as a direct descendant of the pre-CDM model proposed by Richetta and Odoni (1993).

ATFM in the CDM era also provides fertile ground for much future research because the scope of potential OR analysis and modeling has expanded greatly. Examples of some topics, along with occasional recent references, include: identifying (as an airline) flights that should be cancelled or delayed (and by how much) in connection with GDPs, recovering (as an airline) from irregular operations (cf. §2) resulting from GDPs or other ATFM interventions, ensuring equity of access to airports and ATM resources (Vossen et al. 2003), collaborative routing of aircraft through congested airspace (Ball et al. 2002), introducing bartering and possibly market-based mechanisms in the allocation of airport slots (Vossen and Ball 2001, Hall 1999), and developing efficient simulation environments for the testing of alternative ATFM strategies.

5. Conclusion
Operations research has been one of the principal contributors to the enormous growth that the air transport sector has experienced during the past 50 years. In the best tradition of OR, the development of models and of solutions has been motivated by issues and problems encountered in practice and has led, in several instances, to insights of a general nature and to important methodological advances in the OR field at large. At this point, OR models and algorithms are diffused throughout the sector and constitute an integral part of the standard practices of airlines, airports, and ATM service providers.

In view of the numerous challenges that it currently faces, it is safe to expect a continuing central role for OR in the air transport sector’s future. As indicated in this paper, there are many promising topics for future research in each of the areas examined. At the most fundamental level, and in general terms, the frontiers can be summarized as follows:

- Relaxing the boundaries between the successive stages of aircraft and crew schedule planning, so that schedule design, fleet assignment, aircraft maintenance routing, and crew scheduling might eventually be performed in an integrated way, rather than solved sequentially as interrelated, but distinct sub-problems.
- Including pricing decisions in revenue management, instead of treating fares and fare classes as fixed, externally specified inputs.
- Developing fast decision support tools that increase the safety and efficiency of air transport operations by taking advantage of the massive, real-time data flows in an increasingly “info-centric” aviation infrastructure.

References


—, —, —, 2001b. Rerouting aircraft for airline recovery. Working paper, Georgia Institute of Technology, Atlanta, GA.


