Integrated Bilevel Optimization for Bus Route Design and Frequency Setting

January 9, 2026

Contents

Notation and Variables

Sets and Indices

N	Set of nodes
A	Set of directed links $a = (u, v)$
K	Set of OD pairs $w = (o, d)$
H	Set of user segments h
M_h	Feasible modes for segment h
m	Mode index, $m \in \{D, X, B, R, W, O\}$
R	Set of candidate bus routes
J	Set of headway/frequency options
K_w^m	Candidate paths for OD w and mode m
k	Path index, $k \in K_w^m$
$\{\phi_r\}$	Flow/capacity ratios for Beckmann/BPR breakpoints

Parameters

t_a^0 [h]	Free-flow travel time on link a
C_a	Capacity of link a
α	BPR parameter (default 0.15)
β	BPR parameter (default 4)
$\bar{v}_a$	Exogenous background flow on link a (non-decision)
D_{odh}	Demand from o to d for segment h
D_w	Total demand for OD w: $\sum_h D_{odh}$
VOT_h	Value of time for segment h
$\mathrm{VOT}_{\mathrm{bg}}$	Background value of time
$\hat{C}_m^{\text{op-env}}$	Per-hour env/operational cost for mode $m \in \{D, X, \mathrm{bg}\}$
FC_r	Fixed cost to activate bus route r
C_{fleet}	Per-bus fleet cost (purchase/maintenance)
H_j [h]	Headway for option $j \in J$
C_B	Bus capacity per vehicle
Ψ_{odh}^m	Fixed utility for exogenous modes $m \in \{R, W, O\}$
Cost_{od}^m	Monetary trip cost for $m \in \{D, X\}$
κ_m	Cost coefficient for mode m
θ	SUE entropy weight (dispersion)
μ	MNL entropy weight (scale)

Upper-Level Decision Variables

$x_r \in \{0, 1\}$	Activate bus route r
$w_{rj} \in \{0, 1\}$	Select headway option j for route r (at most one)
$n_r \in \mathbb{Z}_+$	Number of buses allocated to route r

Lower-Level Primal Variables

$v_a \geq 0$	Total flow on link a
$f_k \geq 0$	Path flow for path k (defined over K_w^m)
$q_{odh}^m \geq 0$	Demand assigned to mode m for OD w=(o, d) and segment h

Lower-Level Dual Variables

ρ_a	Dual for link-flow definition and Beckmann cuts
λ_{od}^m	Dual for path-to-mode conservation (per OD and mode)
γ_{od}	Dual for OD demand conservation

Auxiliary Variables and Linearization Cuts

$\tau_a \geq 0$	Piecewise-linear approximation of Beckmann integral on link a
$\varphi_k \geq 0$	Tangent-based approximation for path entropy $f_k \ln f_k$
$\xi_{odh}^m \geq 0$	Tangent-based approximation for mode entropy $q \ln q$
$\tilde{t}_a \geq 0$	Piecewise-linear BPR travel time for background cost (upper level)
$\zeta_k^{\text{f-time}} \geq 0$	McCormick variable for $f_k \cdot t_k$ (auto D/X)
$\zeta_k^{\text{f-TT}} \geq 0$	McCormick variable for bus in-vehicle time
$\zeta_k^{\text{f-WT}} \geq 0$	McCormick variable for bus waiting time
$\alpha_{a,r} \geq 0$	Beckmann dual-cut weights (sum to 1 for each a)
$\beta_{k,j} \geq 0$	Path-entropy dual-cut weights (sum to $1/\theta$)
$\eta_{odh,j}^m \geq 0$	Mode-entropy dual-cut weights (sum to $1/\mu$)

Derived Quantities

$t_a(v)$	BPR link travel time $t_a^0\big[1 + \alpha(v/C_a)^\beta\big]$
TT_k	In-vehicle time for bus path k
WT_k	Waiting time under chosen headway
$\hat{t}_k$	Path travel time expression for auto paths (sum over links)

1 Problem Structure

1.1 Upper Level (Leader): Route Authority

The transit authority chooses which routes to operate and at what frequency to **minimize total system cost**.

Decision Variables:

$$x_r \in \{0,1\}, \quad w_{rk} \in \{0,1\}, \quad n_r \in \mathbb{Z}^+ \tag{1}$$

where x_r indicates route activation, w_{rk} selects frequency class k for route r, and n_r is the number of buses allocated to route r.

Objective:

$$\min_{x,w,n} Z_{\text{upper}} = Z_{\text{op}}(x,n) + Z_{\text{user}}(x,w,n,q^*) + Z_{\text{bg}}(v^*) \tag{2}$$

where q^* and v^* are optimal user responses to decisions (x, w, n).

Critical Insight: The operator now makes two independent decisions:

1. **Route Design:** x_r (activate/deactivate) and w_{rk} (select frequency)

2. **Fleet Allocation:** n_r (allocate discrete number of buses)

The objective function depends on n_r (fleet size), *not* on w_{rk} (frequency selection). This reflects the operational reality that bus procurement costs are determined by vehicle numbers, not scheduling frequency.

1.2 Lower Level (Follower): User Behavior

Given operator decisions (x, w), users choose routes and modes to minimize their own generalized cost. The model distinguishes two fundamentally different path choice mechanisms plus fixed-utility modes:

extbfAuto Mode (D/X): Users freely select among pre-generated k-shortest paths (e.g., 3-5 paths per OD pair) based on stochastic user equilibrium (SUE). All paths are always available.

extbfBus Mode (B): Each bus path corresponds to a fixed transit route r. Users can only choose routes that are *activated by the upper level* $(x_r = 1)$. If a route is not selected $(x_r = 0)$, its flow is constrained to zero.

extbfExogenous Modes (R/W/O): These modes have no path-level choice. Their utilities are fixed per (o, d, h, m) via precomputed travel time and monetary cost parameters, and demand q^m_{odh} is determined directly by mode choice without path flows.

This distinction is enforced by constraints:

$$f_k^{w,B} \le M_w \cdot x_r, \quad \forall k \in K_w^B \tag{3}$$

enabling the upper level to optimize the transit network design while users respond via equilibrium on the available network.

1.3 Key Modeling Innovation

The formulation achieves a critical distinction:

- **Auto users:** Experience route choice as in standard SUE (MNL path choice among all candidates)
- **Bus users:** Experience route choice only among *operator-activated routes*, coupling upper-level design with lower-level equilibrium
- **Exogenous modes (R/W/O):** No path choice; they enter mode choice with fixed utilities built from exogenous time/cost inputs

Both are unified under the same mathematical framework (SUE+MNL via entropy regularization), but bus path availability is explicitly controlled by binary decisions x_r, creating the essential bi-level structure.

2 Upper Level Objective: Total System Cost

The upper-level objective consists of three major components representing different stakeholders' costs. We minimize the total system cost:

$$Z_{\text{upper}} = Z_{\text{sys-op}} + Z_{\text{user}} + Z_{\text{bg}} \tag{4}$$

2.1 Component 1: System Operator Cost $Z_{\text{sys-op}}$

The operator (transit authority) incurs costs from operating bus routes and environmental/operational costs from private vehicle usage.

2.1.1 Bus Operating Costs

$$Z^{\text{bus}}_{\text{sys-op}} = \sum_{r \in R} \text{FC}_r \cdot x_r + \sum_{r \in R} C_{\text{fleet}} \cdot n_r \tag{5}$$

where:

- FC_r is the fixed cost for activating route r (e.g., driver depot allocation, route supervision)
- C_{fleet} is the per-vehicle cost (e.g., bus purchase, maintenance, insurance per year)
- n_r is the number of buses allocated to route r (integer decision variable)

Key Change: Operating cost now depends on *fleet size* (n_r), not *frequency selection* (w_{rk}). This aligns with real transit operations where:

1. Purchasing buses is a capital/maintenance expense (proportional to n_r)
2. Scheduling frequency (w_{rk}) affects service quality but not directly the vehicle procurement cost
3. The constraint $n_r \times C_B \geq$ (total demand on route r) and $n_r \times \bar{H}_j \geq T_r \times w_{rj}$ ensures fleet is sufficient for chosen frequency and capacity

2.1.2 Auto Operational/Environmental Costs (Linearized)

Using McCormick auxiliary variables $\zeta_k^{\text{f-time}}$ to linearize the product of flow and travel time:

$$Z_{\text{sys-op}}^{\text{auto}} = \sum_{m\in\{D,X\}} \sum_{w\in W} \sum_{k\in K_w^m} \hat{C}_m^{\text{op-env}} \cdot \zeta_k^{\text{f-time}} \tag{6}$$

where $\zeta_k^{\text{f-time}}$ approximates $f_k \cdot t_k$ (path flow times travel time).

Important Note on Flow Variables: The path flow variable f_k represents the *total flow* on path k aggregated across all user segments $h \in H$. Therefore, $\zeta_k = f_k \times t_k$ already captures the full flow-weighted travel time, and we do **not** multiply by the external demand parameter D_w^h (which would cause double-counting since f_k is itself an endogenous flow variable determined by the SUE equilibrium).

2.2 Component 2: User Cost Z_{user} (Linearized)

User costs include both time costs (monetized via VOT) and monetary costs (tolls, fares, operating costs).

$$\begin{aligned} Z_{\text{user}} = \sum_{w\in W} \Bigg[& \\ & \underbrace{\overline{\text{VOT}}_w \sum_{k\in K_w^D \cup K_w^X} \zeta_k^{\text{f-time}}}_{\text{Auto Time Cost}} + \underbrace{\sum_{k\in K_w^D \cup K_w^X} f_k \cdot C_w^{m(k)}}_{\text{Auto Monetary Cost}} \\ & + \underbrace{\overline{\text{VOT}}_w \sum_{k\in K_w^B} \left(\zeta_k^{\text{f-TT}} + \zeta_k^{\text{f-WT}}\right)}_{\text{Bus Time Cost (Linearized)}} + \underbrace{\sum_{k\in K_w^B} f_k \cdot \text{Fare}_B}_{\text{Bus Fare (Monetary)}} \Bigg] \quad (7) \\ & + \sum_{(o,d,h)\in\mathcal{K}\times H} \sum_{m\in\{R,W\}} \left(-\ \text{VOT}_h \cdot \Psi_{odh}^m\right) \cdot q_{odh}^m \quad (8) \end{aligned}$$

where:

- $\overline{\text{VOT}}_w = \sum_{h\in H} \omega_{wh} \cdot \text{VOT}_h$ is the segment-weighted value of time for OD pair w
- $\omega_{wh} = D_w^h / \sum_{h'\in H} D_w^{h'}$ is the demand-share weight for segment h on OD pair w

- C_w^m is the monetary cost per trip for mode m on OD pair w (includes fuel, tolls)
- Fare_B: unified per-trip bus fare (applied to all bus paths)
- $\zeta_k^{\text{f-time}}$: Linearizes $f_k \times$ (travel time) for auto modes
- $\zeta_k^{\text{f-TT}}$: Linearizes $f_k \times$ (in-vehicle time) for bus
- $\zeta_k^{\text{f-WT}}$: Linearizes $f_k \times$ (waiting time) for bus

Utility Composition (for MNL/SUE) All disutilities/utility terms follow the convention that more negative is worse. Coefficients come from `code/model_parameters.py` unless overridden by data.

- Time coefficients $\beta_{TT,m}$: e.g., $\beta_{TT,D} = -3.0$, $\beta_{TT,X} = -3.5$, $\beta_{TT,B} = -0.6$, $\beta_{TT,R} = -0.5$, $\beta_{TT,W} = -2.0$ (units: utility per hour).
- Waiting time coefficient $\beta_{WT} = -1.0$; cost coefficient $\beta_{TC} = -0.2$.
- Mode constants $\beta_{0,m}$: $\beta_{0,D} = 0$, $\beta_{0,X} = -0.1$, $\beta_{0,B} = 0.8$, $\beta_{0,R} = -0.2$, $\beta_{0,W} = -1.5$, $\beta_{0,O} = -2.0$.
- Auto monetary cost C_w^m: distance-based cost (fuel//) from `Cost_od_auto[(o,d,m)]`; applied positively in Z_{user} as per-trip cost × flow.
- Bus fare Fare_B: unified per-trip fare (currently 5.0) applied to all bus paths.
- Exogenous modes (R,W,O): fixed utility $\Psi_{odh}^m = \beta_{TT,m}\,\text{Time}_{od}^m + \beta_{TC}\,\text{Cost}_{od}^m + \beta_{0,m}$; user cost contribution is $-\text{VOT}_h \cdot \Psi_{odh}^m$.
- Value of time (VOT): segment-specific VOT_h; weighted OD VOT $\overline{\text{VOT}}_w = \sum_h \omega_{wh}\,\text{VOT}_h$ with $\omega_{wh} = D_w^h / \sum_{h'} D_w^{h'}$.

Utility Expressions (mode-specific) Using the sign convention "more negative = worse", the systematic utility for each mode is:

$$\textbf{Auto } (m \in \{D, X\}): \quad U_{w,h}^m = \beta_{TT,m}\,\hat{t}_w^m + \beta_{TC}\,\text{Cost}_w^m + \beta_{0,m} \tag{9}$$

$$\textbf{Bus } (m = B): \quad U_{w,h}^B = \beta_{TT,B}\,\text{TT}_w^B + \beta_{WT}\,\text{WT}_w^B + \beta_{TC}\,\text{Fare}_B + \beta_{0,B} \tag{10}$$

$$\textbf{Exogenous } (m \in \{R, W\}): \quad U_{w,h}^m = \beta_{TT,m}\,\text{Time}_w^m + \beta_{TC}\,\text{Cost}_w^m + \beta_{0,m} \tag{11}$$

$$\textbf{Other } (m = O): \quad U_{w,h}^O = \beta_{0,O} \tag{12}$$

Where:

- $\hat{t}_w^m$: path travel time for auto mode (congestion-dependent, represented via $\zeta^{\text{f-time}}/f$ in the model), aggregated over links.
- Cost_w^m: distance-based per-trip monetary cost for $m \in \{D, X\}$ (fuel//)
- TT_w^B: in-vehicle time on activated bus route/path; WT_w^B: waiting time from chosen headway.
- Fare_B: unified bus fare (currently 5.0), applied once per boarding.
- $\text{Time}_w^m, \text{Cost}_w^m$: exogenous times/costs for Rail/Walk, precomputed in data pipeline.

extbfUser cost link: In the objective, user cost is Z_{user} = VOT-weighted time terms + monetary terms. For exogenous modes, user cost contribution is $-\text{VOT}_h \cdot U^m_{w,h}$ (since utility is negative disutility).

extbfAuto Mode Cost Structure:

- **Drive Alone (D):** \$2.0/km (fuel + vehicle depreciation + maintenance + insurance)
- **Taxi/Ride-sharing (X):** \$3.0 base fare + \$2.5/km

extbfExogenous Mode Utility (R/W/O):

- $\Psi^m_{odh} = \beta_{TT,m} \cdot \text{Time}^m_{od} + \beta_{TC} \cdot \text{Cost}^m_{od} + \beta_{0,m}$ for $m \in \{R, W\}$ (time/cost precomputed per OD)
- $\Psi^O_{odh} = \beta_{0,O}$ (other mode uses only the mode-specific constant)
- These utilities enter Z_{user} through the term $-\sum q^m_{odh} \Psi^m_{odh}$ without any path-level flows

Critical Design Note: Since f_k aggregates flows across all user segments and $\zeta_k = f_k \times t_k$ already contains this aggregated flow information, we use a *weighted average VOT* rather than summing over segments with demand multipliers. This prevents double-counting (external demand D^h_w should not multiply endogenous flow variables f_k).

2.3 Component 3: Background Traffic Cost Z_{bg} (Piecewise-Linear BPR)

Background traffic cost now uses the same piecewise-linear BPR approximation as the lower level (breakpoints $\{\phi_r\}$):

$$Z_{\text{bg}} = \sum_{a \in A} \bar{v}_a \cdot \tilde{t}_a \cdot (\text{VOT}_{\text{bg}} + \hat{C}^{\text{op-env}}_{\text{bg}}) \tag{13}$$

with supporting hyperplanes for $t_a(v_a) = t^0_a \big[1 + \alpha (v_a / C_a)^\beta\big]$:

$$ildet_a \geq t_a(v^r_a) + t'_a(v^r_a)\,(v_a - v^r_a), \quad v^r_a = \phi_r C_a,\ \alpha = 0.15,\ \beta = 4. \tag{14}$$

This ensures the upper-level background cost aligns with the lower-level congestion representation.

2.4 Summary: Linearization and McCormick Technique

Key Design Principle: To maintain MILP structure (required by Gurobi), all nonlinear terms in the upper-level objective are linearized:

- **Congestion effects** use piecewise-linear approximations: Beckmann integral in the lower level and BPR travel time in the upper-level Z_{bg} with the same breakpoints
- **Upper-level objective** uses linear/McCormick approximations for bilinear products
- **Bilinear products** $f_k \times t_k$ are replaced by McCormick variables ζ with linearization constraints

2.5 McCormick Linearization Constraints

To handle bilinear terms $f_k \times t_k$ in the objective function, we introduce McCormick auxiliary variables and corresponding linearization constraints.

2.5.1 Auxiliary Variables

Define three types of McCormick auxiliary variables:

- $\zeta_k^{\text{f-time}}$: Linearizes $f_k \times t_k$ for auto modes ($m \in \{D, X\}$)
- $\zeta_k^{\text{f-TT}}$: Linearizes $f_k \times \text{TT}_k$ (in-vehicle time) for bus mode ($m = B$)
- $\zeta_k^{\text{f-WT}}$: Linearizes $f_k \times \text{WT}_k$ (waiting time) for bus mode ($m = B$)

2.5.2 Constraint Formulations

For each path k, we apply the standard McCormick envelope linearization with four inequalities per bilinear product. Let $f_k \in [f_L, f_U]$ and $t_k \in [t_L, t_U]$ denote the bounds on flow and time variables.

(1) Auto Mode Travel Time Constraints:

For $k \in K_w^D \cup K_w^X$ (drive-alone or taxi paths), we linearize $f_k \times t_k$ using auxiliary variable $\zeta_k^{\text{f-time}}$ with:

$$\zeta_k^{\text{f-time}} \geq t_{k,L} \cdot f_k \qquad \forall k \in K^{\text{auto}} \tag{15}$$
$$\zeta_k^{\text{f-time}} \geq t_k + t_{k,U} \cdot (f_k - f_{k,U}) \qquad \forall k \in K^{\text{auto}} \tag{16}$$
$$\zeta_k^{\text{f-time}} \leq t_{k,U} \cdot f_k \qquad \forall k \in K^{\text{auto}} \tag{17}$$
$$\zeta_k^{\text{f-time}} \leq t_k + t_{k,L} \cdot (f_k - f_{k,U}) \qquad \forall k \in K^{\text{auto}} \tag{18}$$

where:

- $t_k^0 = \sum_{a \in p_k} t_a^0$ is the free-flow travel time (used as $t_{k,L}$)
- $t_{k,U}$ is the upper bound on travel time (free-flow time + maximum congestion delay)
- $f_{k,L} = 0$ (flow non-negativity)
- $f_{k,U} = \sum_{w,h} D_w^h$ (maximum possible flow bounded by total demand)

(2) Bus In-Vehicle Time Constraints:

For $k \in K_w^B$ (bus paths), we linearize $f_k \times \text{TT}_k$ using auxiliary variable $\zeta_k^{\text{f-TT}}$ with:

$$\zeta_k^{\text{f-TT}} \geq \text{TT}_{k,L} \cdot f_k \qquad \forall k \in K^B \tag{19}$$
$$\zeta_k^{\text{f-TT}} \geq \text{TT}_k + \text{TT}_{k,U} \cdot (f_k - f_{k,U}) \qquad \forall k \in K^B \tag{20}$$
$$\zeta_k^{\text{f-TT}} \leq \text{TT}_{k,U} \cdot f_k \qquad \forall k \in K^B \tag{21}$$
$$\zeta_k^{\text{f-TT}} \leq \text{TT}_k + \text{TT}_{k,L} \cdot (f_k - f_{k,U}) \qquad \forall k \in K^B \tag{22}$$

where TT_k^0 is the free-flow in-vehicle time on bus path k (used as $\text{TT}_{k,L}$), and $\text{TT}_{k,U}$ includes congestion.

(3) Bus Waiting Time Constraints:

For $k \in K_w^B$ (bus paths), we linearize $f_k \times \text{WT}_k$ using auxiliary variable $\zeta_k^{\text{f-WT}}$ with:

$$\zeta_k^{\text{f-WT}} \geq 0 \qquad \forall k \in K^B \tag{23}$$
$$\zeta_k^{\text{f-WT}} \geq \text{WT}_k + \text{WT}_{k,U} \cdot (f_k - f_{k,U}) \qquad \forall k \in K^B \tag{24}$$
$$\zeta_k^{\text{f-WT}} \leq \text{WT}_{k,U} \cdot f_k \qquad \forall k \in K^B \tag{25}$$
$$\zeta_k^{\text{f-WT}} \leq \text{WT}_k \qquad \forall k \in K^B \tag{26}$$

where:

- $\mathrm{WT}_k = 0.5 \times \sum_{j \in J_{r(k)}} H_j \cdot w_{r(k),j}$ is the average waiting time
- $r(k)$ is the route corresponding to path k
- H_j is the headway for frequency option j
- $w_{r,j} \in \{0,1\}$ is the binary frequency selection variable
- $\mathrm{WT}_{k,L} = 0$ (minimum waiting time when highest frequency is selected)
- $\mathrm{WT}_{k,U} = 0.5 \times \max_j H_j$ (maximum average waiting time)

2.5.3 Linearization Strategy and Envelope Tightness

McCormick Envelope: The four-inequality system forms a tight linear relaxation (convex hull) of the bilinear region $\{(f,t,\zeta) : \zeta = f \cdot t, f \in [f_L, f_U], t \in [t_L, t_U]\}$. This is the tightest possible linear approximation.

Handling Congestion: To avoid introducing quadratic terms $f_k \times v_a$ (which would violate MILP structure), the time variable t_k in the McCormick constraints is treated as:

- **Lower bound** $t_{k,L}$: Free-flow time t_k^0 (constant, independent of congestion)
- **Upper bound** $t_{k,U}$: Free-flow time plus maximum expected congestion delay
- **Actual value** t_k: Can vary between bounds (as a function of link flows from lower level)

Implementation Detail: Since t_k is not an explicit decision variable (it depends implicitly on link flows v_a through the lower-level problem), we use a *conservative outer approximation* in constraints (2) and (4):

- In constraint (2): Replace t_k with its lower bound $t_{k,L}$, giving:

$$\zeta_k^{\text{f-time}} \geq f_{k,U} \cdot t_{k,L} + t_{k,U} \cdot f_k - f_{k,U} \cdot t_{k,U} \tag{27}$$

- In constraint (4): Replace t_k with its upper bound $t_{k,U}$, giving:

$$\zeta_k^{\text{f-time}} \leq f_{k,U} \cdot t_{k,U} + t_{k,L} \cdot f_k - f_{k,U} \cdot t_{k,L} \tag{28}$$

This creates a valid (though slightly looser) outer approximation that maintains MILP linearity while capturing the essential bounds.

The full congestion effects are captured exactly in the lower-level problem through:

- The Beckmann integral term: $\sum_{a \in A} \int_0^{v_a} t_a(\xi) d\xi$
- Piecewise linear approximation: $\tau_a \geq t_a(v_a^r)(v_a - v_a^r) + B_a(v_a^r)$ for multiple breakpoints r

This design ensures:

1. The upper-level objective remains linear (required for MILP)
2. Congestion effects are exactly modeled in the lower-level equilibrium
3. The McCormick envelope provides valid (conservative) bounds on flow-time products
4. The strong duality equality links upper and lower levels, ensuring consistency

2.5.4 Variable Bounds

All McCormick auxiliary variables are non-negative:

$$\zeta_k^{\text{f-time}}, \zeta_k^{\text{f-TT}}, \zeta_k^{\text{f-WT}} \geq 0, \quad \forall k \tag{29}$$

Upper bounds are implicitly determined by:

- Flow bounds: $f_k \leq \sum_{w,m} D_w$ (total demand)
- Time bounds: t_k^0 are fixed constants from network data
- Headway bounds: $\sum_j H_j \cdot w_{r,j} \leq \max_j H_j$ (largest headway option)

Decision Flow:

1. Operator simultaneously chooses:
   - Routes to activate: $x_r \in \{0, 1\}$
   - Service frequency for each route: $w_{rk} \in \{0, 1\}$ (selects 1 of $|J_r|$ frequency options)
   - Fleet allocation: $n_r \in \mathbb{Z}^+$ (number of buses for each route)
2. Fleet decisions must satisfy:
   - *Capacity constraint:* $n_r \times C_B \geq$ total passenger demand on route r
   - *Activation constraint:* $n_r \leq n_{\max} \times x_r$ (cannot deploy buses to inactive routes)
3. Users respond via lower-level equilibrium, determining $q_w^m, f_k^{w,m}, v_a$
4. Link flows v_a determine congestion (captured in lower level via Beckmann, approximated linearly in upper objective)
5. All three cost components depend on both operator decisions and user responses
6. Optimizer seeks the (x, w, n) that minimizes total system cost: $\min Z_{\text{op}}(x, n) + Z_{\text{user}}(x, w, n) + Z_{\text{bg}}(v)$

2.6 Lower Level (Follower): User Behavior

Given operator decisions (x, w), users choose routes and modes via a convex optimization problem that simultaneously achieves:

1. **Stochastic User Equilibrium (SUE)** for route choice
2. **Multinomial Logit (MNL)** for mode choice

The lower level is formulated as a convex optimization problem, and its KKT conditions are mathematically equivalent to SUE+MNL.

3 Lower Level Problem: User Equilibrium with Mode Choice

3.1 Decision Variables

Let $f_k^{w,m}$ denote path flow and q_w^m denote mode demand.

$$f_k^{w,m} \geq 0 \quad \text{(path flow for mode } m \text{ on OD pair } w) \tag{30}$$

$$q_w^m \geq 0 \quad \text{(total demand for mode } m \text{ on OD pair } w) \tag{31}$$

$$v_a \geq 0 \quad \text{(link flow on link } a) \tag{32}$$

3.2 Objective Function: Convex Formulation (Utility Units)

$$\begin{aligned}\min_{f,q,v} \quad Z_{\text{Lower}} = \quad & \underbrace{\sum_m \beta_{TT,m} \sum_{a\in A} \int_0^{v_a^m} t_a(x)dx}_{\text{(1) Beckmann Link Cost (utility units, mode-weighted)}} \\ & + \underbrace{\frac{1}{\theta}\sum_{w,m,k} f_k^{w,m}(\ln f_k^{w,m} - 1)}_{\text{(2) SUE Entropy}} + \underbrace{\frac{1}{\mu}\sum_{w,m} q_w^m(\ln q_w^m - 1)}_{\text{(3) MNL Entropy}} - \underbrace{\sum_{w,m} q_w^m \Psi_w^m}_{\text{(4) Fixed Utility}}\end{aligned} \tag{33}$$

.

3.2.1 Component (1): Beckmann Link Cost (Mode-Specific Utility Units)

$$\sum_m \beta_{TT,m} \int_0^{v_a^m} t_a(\xi)d\xi \tag{34}$$

where v_a^m is the flow of mode m on link a, and $t_a(v_a) = t_a^0[1 + \alpha(v_a/C_a)^\beta]$ is the BPR travel time function (in hours). Multiplying by $\beta_{TT,m} < 0$ converts physical time to utility units (negative = cost). The mode-specific approach ensures:

- **Accurate user valuation:** Driving users with $\beta_{TT,D} = -3.0$ are much more sensitive to congestion than bus users with $\beta_{TT,B} = -0.6$, which is reflected in the objective function weights.
- **Strict convexity:** The sum of mode-weighted integrals remains strictly convex, guaranteeing unique equilibrium.
- **Dimensional consistency:** All mode-specific path costs are in utils, enabling direct combination with mode-level utilities in MNL.

3.2.2 Component (2): SUE Entropy for Route Choice

$$\frac{1}{\theta}\sum_{w,m,k} f_k^{w,m}(\ln f_k^{w,m} - 1) \tag{35}$$

The parameter θ is the SUE dispersion parameter. This entropy term induces logit-based path choice, recovering SUE from KKT conditions.

3.2.3 Component (3): MNL Entropy for Mode Choice

$$\frac{1}{\mu}\sum_{w,m} q_w^m(\ln q_w^m - 1) \tag{36}$$

The parameter μ is the MNL scale parameter. This entropy term induces logit-based mode choice, recovering MNL from KKT conditions.

3.2.4 Component (4): Fixed Utility Component

$$-\sum_{w,m} q_w^m \Psi_w^m \tag{37}$$

where Ψ_w^m is the fixed (non-time) utility of mode m, e.g., mode-specific constants, comfort attributes, etc.

3.3 Constraints

3.3.1 Flow Conservation (Path to Mode)

$$\sum_{k \in K_w^m} f_k^{w,m} = q_w^m, \quad \forall w \in W,\ m \in \{D, X, B\} \tag{38}$$

Total path flow for path-based modes equals their mode demand. Exogenous modes $m \in \{R, W, O\}$ have no path sets; their demand q_w^m is decided directly in mode choice. **Dual variable:** λ_w^m

3.3.2 Demand Conservation (Mode to OD)

$$\sum_{m \in \{D,X,B,R,W,O\}} q_w^m = D_w, \quad \forall w \in W \tag{39}$$

All demand must be allocated across the full mode set. **Dual variable:** γ_w

3.3.3 Link Flow Definition

$$v_a = \sum_{w \in W} \sum_{m \in \{D,X,B\}} \sum_{k \in K_w^m} f_k^{w,m} \delta_{ak}^{w,m}, \quad \forall a \in A \tag{40}$$

Link flow is the sum of all path flows traversing that link (exogenous modes do not generate link flows). **Dual variable:** ρ_a

3.3.4 Non-negativity

$$f_k^{w,m} \geq 0, \quad q_w^m \geq 0, \quad v_a \geq 0 \tag{41}$$

3.3.5 Path Availability Constraints (Distinguishing Auto and Bus Path Choice)

Critical Design Distinction: The model differentiates between auto and bus path choice mechanisms through upper-level controlled availability constraints.

Auto Mode Path Choice (Free Choice): For auto modes ($m \in \{D, X\}$), users have *unrestricted access* to all pre-generated k-shortest paths for each OD pair. These paths form a candidate set (e.g., 3-5 paths per OD pair), and users distribute flows according to SUE equilibrium based on congestion-dependent travel times:

$$f_k^{w,D} \geq 0, \quad f_k^{w,X} \geq 0, \quad \forall k \in K_w^D \cup K_w^X \tag{42}$$

All paths are always available; no activation constraint exists.

Bus Mode Path Choice (Upper-Level Controlled): For bus mode ($m = B$), each path k corresponds to a *fixed transit route* r with a predetermined physical alignment. The upper-level decision variable $x_r \in \{0, 1\}$ controls whether route r is activated:

$$f_k^{w,B} \leq M_w \cdot x_r, \quad \forall k \in K_w^B,\ r = \text{route}(k) \tag{43}$$

where M_w is a sufficiently large constant (e.g., total OD demand D_w), and route(k) maps path k to its corresponding route r.

Interpretation:

- If $x_r = 0$ (route not activated): $f_k^{w,B} = 0$ (no flow on this path)
- If $x_r = 1$ (route activated): $f_k^{w,B}$ can take positive values determined by SUE equilibrium

Additionally, bus mode demand is only allowed when at least one bus route serves the OD pair:

$$q_w^B \leq M_w \cdot \sum_{r \in R_w} x_r, \quad \forall w \in W \tag{44}$$

where R_w is the set of routes serving OD pair w. This ensures users cannot choose bus mode if no routes are available.

Exogenous Modes (Fixed Utility): Modes $m \in \{R, W, O\}$ have no path sets or availability constraints. Their demand q_w^m is chosen directly in mode choice based on fixed utilities Ψ_{odh}^m computed from exogenous travel time and monetary cost parameters; they do not contribute link flows.

Implementation of Two-Level Choice: This design achieves a hierarchical decision structure:

1. **Upper Level:** Operator optimizes *which bus routes to activate* (x_r) and *headway choices* (w_{rk})
2. **Lower Level:** Users respond via:
   - **Mode choice (MNL):** Select between D/X/B/R/W/O based on expected costs
   - **Path choice (SUE):**
     - Auto users: choose freely among all k-shortest paths (congestion-aware)
     - Bus users: choose only among *activated routes* ($x_r = 1$), subject to capacity and waiting time
     - Exogenous modes (R/W/O): no path choice; demand is determined directly from mode utilities

4 Proof of Equivalence: KKT Conditions Recover SUE and MNL

4.1 Step 1: Lagrangian Formulation

Introduce Lagrange multipliers λ_w^m, γ_w, ρ_a for the three constraint sets. The Lagrangian is:

$$\mathcal{L} = Z_{\text{Lower}} + \sum_{w,m} \lambda_w^m \left(q_w^m - \sum_k f_k^{w,m} \right) + \sum_w \gamma_w \left(D_w - \sum_m q_w^m \right) + \sum_a \rho_a \left(v_a - \sum_{w,m,k} f_k^{w,m} \delta_{ak}^{w,m} \right) \tag{45}$$

4.2 Step 2: KKT Condition for Path Flow (Route Choice SUE)

Taking $\frac{\partial \mathcal{L}}{\partial f_k^{w,m}} = 0$:

$$\beta_{TT,m} \sum_{a \in A} t_a(v_a) \delta_{ak}^{w,m} + \frac{1}{\theta} \ln f_k^{w,m} - \lambda_w^m = 0 \tag{46}$$

Let $C_k^{w,m} = \beta_{TT,m} \sum_a t_a(v_a) \delta_{ak}^{w,m}$ be the path cost in utility units (travel time × mode-specific marginal utility). Rearranging:

$$\ln f_k^{w,m} = \theta(\lambda_w^m - C_k^{w,m}) \tag{47}$$

$$f_k^{w,m} = \exp(\theta \lambda_w^m) \exp(-\theta C_k^{w,m}) \tag{48}$$

Using flow conservation $\sum_k f_k^{w,m} = q_w^m$:

$$q_w^m = \exp(\theta \lambda_w^m) \sum_k \exp(-\theta C_k^{w,m}) \tag{49}$$

For auto modes ($m \in \{D, X\}$), $C_k^{w,m}$ does not depend on f_k, so the usual logit form holds:

$$P_k^{w,m} = \frac{\exp(-\theta C_k^{w,m})}{\sum_{l \in K_w^m} \exp(-\theta C_l^{w,m})}, \quad m \in \{D, X\}. \tag{50}$$

For bus ($m = B$), the path cost is also independent of f_k (waiting time is mode-level constant in Ψ_w^B, not path-dependent), so the logit form applies identically:

$$P_k^{w,B} = \frac{\exp(-\theta C_k^{w,B})}{\sum_{l \in K_w^B} \exp(-\theta C_l^{w,B})}. \tag{51}$$

This is the logit path choice model—exactly SUE for all modes including bus!

4.3 Step 3: KKT Condition for Mode Demand (Mode Choice MNL)

Taking $\frac{\partial \mathcal{L}}{\partial q_w^m} = 0$:

$$\frac{1}{\mu} \ln q_w^m - \Psi_w^m + \lambda_w^m - \gamma_w = 0 \tag{52}$$

Let $U_w^m = \Psi_w^m - \lambda_w^m$ be the systematic utility (fixed utility minus inclusive value of paths).

Fixed utility minus logsum equals utility. Rearranging the stationarity condition gives

$$\ln q_w^m = \mu\big(\gamma_w + \Psi_w^m - \lambda_w^m\big) = \mu\big(\gamma_w + U_w^m\big) \tag{53}$$

so the MNL choice probability uses $U_w^m = \Psi_w^m - \lambda_w^m$. Here Ψ_w^m is path-independent (for bus, $\beta_{TC}\,\mathrm{Fare}_B + \beta_{0,B}$), while λ_w^m is the logsum over path costs and therefore already contains path time/wait components (for bus, TT, WT). Thus "fixed utility $-\ \lambda$" recovers the full systematic utility used in mode choice.

Rearranging:

$$\ln q_w^m = \mu(\gamma_w - \lambda_w^m + \Psi_w^m) = \mu(\gamma_w + U_w^m) \tag{54}$$

$$q_w^m = \exp(\mu \gamma_w) \exp(\mu U_w^m) \tag{55}$$

Using demand conservation $\sum_m q_w^m = D_w$:

$$D_w = \exp(\mu \gamma_w) \sum_m \exp(\mu U_w^m) \tag{56}$$

Therefore:

$$P(m|w) = \frac{q_w^m}{D_w} = \frac{\exp(\mu U_w^m)}{\sum_{n \in M} \exp(\mu U_w^n)} \tag{57}$$

Detailed derivation: Why $U_w^m = \Psi_w^m - \lambda_w^m$ satisfies utility With mode-specific scaling, all path costs are now in utility units:

$$C_k^{w,m} = \beta_{TT,m} \sum_a t_a(v_a) \delta_{ak}^{w,m} \quad \text{[utils]} \tag{58}$$

For all modes, $C_k^{w,m}$ is exogenous to the flow $f_k^{w,m}$ (depends only on link congestion v_a). Therefore, all modes exhibit standard logit path choice:

$$\lambda_w^m = -\frac{1}{\theta} \ln \sum_k \exp(-\theta C_k^{w,m}) \quad \text{[utils]} \tag{59}$$

This is the **inclusive value (logsum)** for mode m, also in utility units. The systematic utility for mode choice is then:

$$U_w^m = \Psi_w^m - \lambda_w^m = \underbrace{\beta_0^m + \beta_{TC}\,\text{Cost}_m + \beta_{WT}\,\text{WT}_m}_{\Psi_w^m\text{:fixed utility [utils]}} + \underbrace{\frac{1}{\theta} \ln \sum_k \exp(-\theta C_k^{w,m})}_{\text{logsum [utils]}} \tag{60}$$

5 Single-Level MILP Reformulation

To solve the original bilevel problem, we convert the lower-level optimization into constraints via strong duality and outer approximation.

5.1 Step 1: Piecewise Linear Approximation

The lower-level objective contains nonlinear convex terms. We approximate them using linear outer approximations.

5.1.1 Beckmann Term Linearization

Replace $\sum_m \beta_{TT,m} \int_0^{v_a^m} t_a(\xi)d\xi$ with auxiliary variable τ_a (in utility units) bounded by:

$$\tau_a \geq \sum_m \beta_{TT,m} \left[t_a(v_a^{m,r}) \cdot (v_a^m - v_a^{m,r}) + B_a(v_a^{m,r}) \right], \quad \forall a, m, r \tag{61}$$

where $\{v_a^{m,r}\}$ are pre-selected breakpoints for mode m on link a, and $B_a(v_a^{m,r}) = \int_0^{v_a^{m,r}} t_a(\xi)d\xi$ is the integral at breakpoint.

5.1.2 Entropy Term Linearization

Replace $\sum_k f_k(\ln f_k - 1)$ with auxiliary variable φ_k bounded by:

$$\varphi_k \geq (1 + \ln \hat{f}_j) f_k - \hat{f}_j, \quad \forall k, j \tag{62}$$

where $\{\hat{f}_j\}$ are pre-selected breakpoints.

Similarly for q_w^m with variable ξ_w^m.

5.2 Step 2: Strong Duality

The linearized lower-level problem becomes a standard LP. By strong duality, we can replace the lower-level optimization with:

1. **Primal Feasibility:** Linearized constraints
2. **Dual Feasibility:** Dual constraints derived from Lagrangian
3. **Strong Duality Equation:** Primal objective = Dual objective

5.2.1 Dual Problem

For each link a, path k, OD w, mode m, introduce dual variables:

- ρ_a: dual for link flow definition
- λ_w^m: dual for flow conservation
- γ_w: dual for demand conservation
- $\alpha_{a,r}, \beta_{k,j}, \eta_{w,m,j} \geq 0$: duals for approximation constraints

Dual Objective:

$$\max Z_{\text{Dual}} = \sum_w D_w \gamma_w + \sum_{a,r} K_{a,r} \alpha_{a,r} + \sum_{k,j} K_{k,j}^{\varphi} \beta_{k,j} + \sum_{w,m,j} K_{w,m,j}^{\xi} \eta_{w,m,j} \tag{63}$$

where K terms are the intercepts from the linear approximations.
Dual Constraints (KKT stationarity conditions):
For link flow v_a:

$$\rho_a - \sum_r S_{a,r} \alpha_{a,r} \leq 0 \tag{64}$$

For path flow $f_k^{w,m}$:

$$\lambda_w^m - \sum_a \delta_{ak} \rho_a - \sum_j S_{k,j}^{\varphi} \beta_{k,j} \leq 0 \tag{65}$$

For mode demand q_w^m:

$$-\lambda_w^m + \gamma_w - \sum_j S_{w,m,j}^{\xi} \eta_{w,m,j} \leq -\Psi_w^m \tag{66}$$

For approximation variables:

$$\sum_r \alpha_{a,r} = 1 \tag{67}$$

$$\sum_j \beta_{k,j} = \frac{1}{\theta} \tag{68}$$

$$\sum_j \eta_{w,m,j} = \frac{1}{\mu} \tag{69}$$

Strong Duality Equality:

$$\sum_a \tau_a + \frac{1}{\theta} \sum_k \varphi_k + \frac{1}{\mu} \sum_{w,m} \xi_w^m - \sum_{w,m} q_w^m \Psi_w^m = \sum_w D_w \gamma_w + \sum_{a,m,r} K_{a,m,r} \alpha_{a,m,r} + \dots \tag{70}$$

where the left-hand side is the primal objective (in utility units, with Beckmann term implicitly weighted by mode-specific $\beta_{TT,m}$ through the τ_a variables) and the right-hand side is the dual objective.

5.3 Step 3: Big-M Linearization for Operator Decisions

To couple operator decisions (x, w) with user responses (f, q, v), add Big-M constraints:

Path Activation (Big-M 1):

$$f_k^{w,m} \leq M \cdot x_{r(k)}, \quad \forall k \tag{71}$$

If route r is inactive ($x_r = 0$), no flow on its paths.

Mode Availability (Big-M 2):

$$q_w^B \leq M \cdot \sum_{r \in R_w} x_r, \quad \forall w \tag{72}$$

Bus demand is allowed only if at least one bus route for OD w is activated.

Upper Level Constraints:

Constraint 1: Headway Selection (Single Choice)

$$\sum_k w_{rk} = x_r, \quad \forall r \tag{73}$$

extbfInterpretation:

- If a route is activated ($x_r = 1$), exactly one headway option must be selected ($\sum_k w_{rk} = 1$)
- If a route is not activated ($x_r = 0$), no headway can be chosen ($\sum_k w_{rk} = 0$)
- This enforces: *open routes must have an explicit frequency; closed routes have none*

extbfExample: three headway options $k \in \{1, 2, 3\}$ (e.g., 5/10/15 minutes):

- $x_r = 1, w_{r1} = 1, w_{r2} = 0, w_{r3} = 0 \rightarrow$ route is open with 5-minute headway
- $x_r = 0, w_{r1} = 0, w_{r2} = 0, w_{r3} = 0 \rightarrow$ route is closed, no frequency

Constraint 2: Bus Fleet Size and Capacity Constraint **Design choice (no transfers):** each bus path is one physical route (point-to-point, no transfers). Path k maps one-to-one to route $r(k)$. The system must decide the *number of buses* (n_r) deployed on each route to handle user demand.

Fleet capacity constraint:

$$n_r \times C_B \geq \sum_{w \in W} \sum_{k \in K_w^B : r(k) = r} f_k^{w,B} \quad \forall r \in R \tag{74}$$

Fleet activation constraint:

$$n_r \leq n_{\max} \times x_r \quad \forall r \in R \tag{75}$$

Symbols (reusing existing notation):

- R: candidate bus routes; J_r: headway options for route r
- K_w^B: bus paths for OD w (each path corresponds to one route $r(k)$)
- $f_k^{w,B}$: path flow for bus mode
- C_B: capacity per bus (pax/veh)

- H_j: headway option j (minutes)
- $w_{rj} \in \{0,1\}$: route r selects headway j; $x_r \in \{0,1\}$: route activation
- $n_r \in \mathbb{Z}_+$: **(NEW)** number of buses deployed on route r
- $n_{\max}$: upper bound on fleet size per route

Meaning of (74)**:**

Total passenger demand on route r cannot exceed total vehicle capacity. The sum of all bus path flows using route r must satisfy:

$$\text{Route } r \text{ flow} = \sum_{w \in W} \sum_{k \in K_w^B : r(k)=r} f_k^{w,B} \leq n_r \times C_B$$

This is the primary constraint: fleet size times per-vehicle capacity ensures all users are seated.

Meaning of (75)**:**

Buses only exist when the route is active: if $x_r = 0$, then $n_r = 0$.

Path activation (unchanged):

$$f_k^{w,B} \leq M_w \cdot x_{r(k)} \quad \forall w \in W,\ k \in K_w^B \tag{76}$$

Path k can carry positive flow only if its route is activated ($x_{r(k)} = 1$).

Comparison with previous frequency-only approach:

Aspect	Frequency-Based (Old)	Fleet-Based (New)
Fleet size variable	Implicit, from H_j	Explicit decision n_r
Capacity model	Per-link, aggregated	Per-route, direct
Headway cost	Fixed per headway	Reflected in fleet size
Cost structure	$\text{FC}_r \times x_r + C_{op,rk} \times w_{rk}$	$\text{FC}_r \times x_r + c_{bus} \times n_r$
Flexibility	Limited (headway choice)	Higher (fleet size optimized)

extbfNumerical illustration: Two routes share link a:

- Route r_1 (path i_1): $C_B = 40$, $H_1 = 10$ min ($w_{r_1,1} = 1$) → capacity 4 pax/min; load 100 pax on the path ($\delta_{i_1,a} = 1$)
- Route r_2 (path i_2): $C_B = 40$, $H_2 = 5$ min ($w_{r_2,1} = 1$) → capacity 8 pax/min; load 150 pax ($\delta_{i_2,a} = 1$)
- Link capacity: $4 + 8 = 12$ pax/min; link load over a 120-min horizon $= (100 + 150)/120 \approx 2.08$ pax/min → constraint satisfied ($2.08 \leq 12$)

If route r_2 switches to $H_2 = 20$ min:

- New capacity: $4 + 2 = 6$ pax/min; still satisfied ($2.08 \leq 6$)

If both routes use $H = 20$ min:

- New capacity: $2 + 2 = 4$ pax/min; close to binding ($2.08 \leq 4$)

6 Final Single-Level MILP

6.1 Complete Formulation

$$\min_{x,w,n,f,q,v,\tau,\varphi,\xi,\lambda,\gamma,\rho,\alpha,\beta,\eta} \quad Z_{\text{op}}(x,w,n) + Z_{\text{user}}(f,q,v) + Z_{\text{bg}}(v) \tag{77}$$

$$\mathit{exts.t.} \quad \text{Upper Level Constraints:} \tag{78}$$

$$\sum_k w_{rk} = x_r, \quad \forall r \tag{79}$$

$$n_r \times C_B \geq \sum_{w\in W} \sum_{k \in K_w^B : r(k)=r} f_k^{w,B}, \quad \forall r \tag{80}$$

$$n_r \leq n_{\max} \times x_r, \quad \forall r \tag{81}$$

$$f_k^{w,B} \leq M_w \cdot x_{r(k)}, \quad \forall w, k \tag{82}$$

$$\tag{83}$$

$$\text{Lower Level Primal (Linearized):} \tag{84}$$

$$\sum_k f_k^{w,m} = q_w^m, \quad \forall w, m \tag{85}$$

$$\sum_m q_w^m = D_w, \quad \forall w \tag{86}$$

$$v_a = \sum_{w,m,k} f_k^{w,m} \delta_{ak}, \quad \forall a \tag{87}$$

$$\tau_a \geq \overline{\beta}_{TT}[t_a(v_a^r)(v_a - v_a^r) + B_a(v_a^r)], \quad \forall a, r \tag{88}$$

$$\varphi_k \geq (1 + \ln \hat{f}_j) f_k - \hat{f}_j, \quad \forall k, j \tag{89}$$

$$\xi_w^m \geq (1 + \ln \hat{q}_j) q_w^m - \hat{q}_j, \quad \forall w, m, j \tag{90}$$

$$\tag{91}$$

$$\text{Lower Level Dual (Stationarity):} \tag{92}$$

$$\rho_a - \sum_r S_{a,r} \alpha_{a,r} \leq 0, \quad \forall a \tag{93}$$

$$\lambda_w^m - \sum_a \delta_{ak} \rho_a - \sum_j S_{k,j} \beta_{k,j} \leq 0, \quad \forall k, w, m \tag{94}$$

$$-\lambda_w^m + \gamma_w - \sum_j S_{w,m,j} \eta_{w,m,j} \leq -\Psi_w^m, \quad \forall w, m \tag{95}$$

$$\sum_r \alpha_{a,r} = 1, \quad \forall a \tag{96}$$

$$\sum_j \beta_{k,j} = \frac{1}{\theta}, \quad \forall k \tag{97}$$

$$\sum_j \eta_{w,m,j} = \frac{1}{\mu}, \quad \forall w, m \tag{98}$$

$$\tag{99}$$

$$\text{Strong Duality Equality:} \tag{100}$$

$$\overline{\beta}_{TT} \sum_a \tau_a + \frac{1}{\theta} \sum_k \varphi_k + \frac{1}{\mu} \sum_{w,m} \xi_w^m - \sum_{w,m} q_w^m \Psi_w^m \tag{101}$$

$$= \sum_w D_w \gamma_w + \sum_{a,m,r} B_a(v_a^{m,r}) \alpha_{a,m,r} + \sum_{k,j} \hat{f}_j \, \beta_{k,j} + \sum_{w,m,j} \hat{q}_j \, \eta_{w,m,j} \tag{102}$$

$$\tag{103}$$

$$\text{Coupling (Big-M constraints):} \tag{104}$$

$$f_k^{w,B} \leq M_w \cdot x_{r(k)}, \quad \forall w, k \in K_w^B \tag{105}$$

$$q_w^B \leq M_w \cdot \sum x_r, \quad \forall w \tag{106}$$

7 Solving the MILP

This single-level MILP can be solved using commercial solvers (Gurobi, CPLEX):

- Binary variables: x_r, w_{rk} (operator decisions)
- Continuous variables: $f_k^{w,m}, q_w^m, v_a, \tau_a, \varphi_k, \xi_w^m, \lambda_w^m, \gamma_w, \rho_a, \alpha_{a,r}, \beta_{k,j}, \eta_{w,m,j}$
- Total: Typically 10^5–10^6 variables and constraints
- Time limit: 1–2 hours depending on network size

8 Summary

1. **Upper Level:** Operator minimizes total system cost via route and frequency decisions
2. **Lower Level:** Users respond via SUE+MNL, formulated as a convex optimization with utility-unit consistency
3. **Equivalence Proof:** KKT conditions of the lower level are mathematically equivalent to SUE and MNL (following standard literature)
4. **Single-Level Conversion:** Via piecewise linear approximation, strong duality, and Big-M linearization
5. **Result:** A single MILP that simultaneously optimizes operator decisions and user equilibrium