Cornell University
Operations Research and
Information Engineering

Vehicle Mix in EMS Systems

Shane G. Henderson

Joint work with:

Mix: Kenneth C. Chong, Mark E. Lewis

Bound: Matt Maxwell, Eric Ni, Chaoxu Tong, Susan Hunter, Huseyin Topaloglu

Thanks to:

NSF CMMI 0758441, Optima Corporation, Toronto EMS, Ambulance Victoria, Armann Ingolfsson, Andrew Mason

http://people.orie.cornell.edu/~shane

Outline

- Part 1: All-ALS or Tiered (Mixed) Fleet?
- Part 2 if time: Bounding Performance

ALS Only or Tiered?

- ALS: advanced life support (paramedics)
- BLS: basic life support (EMTs)
- Should the ambulance fleet be all-ALS or a mix?
  - All-ALS: e.g., Ornato et al (1990), Wilson et al . (1992)
  - Mix: e.g., Braun (1990), Clawson (1989), Slovis et al. (1985), Stout et al. (2000)
- In NL, what if have nurse shortage?

ALS Only or Tiered?

- All-ALS
  – Never sends a BLS ambulance to a call that needs ALS
  – Can potentially triage more quickly
- Tiered:
  – Many calls don't require paramedics
  – ALS is more expensive, so mixed fleets can be larger – shorter response times
  – Hiring and training paramedics can be hard
- Which is better?

Modeling Structure

- Decision variables:
  - n_a, n_b = Number of ALS, BLS
- Constraints:
  - B = annual operating budget
  - c_a (c_b) = annual cost per ALS (BLS)
  - $c_a n_a + c_b n_b \leq B$
- Enumerate over n_a to get optimal sol.
- Objective function?

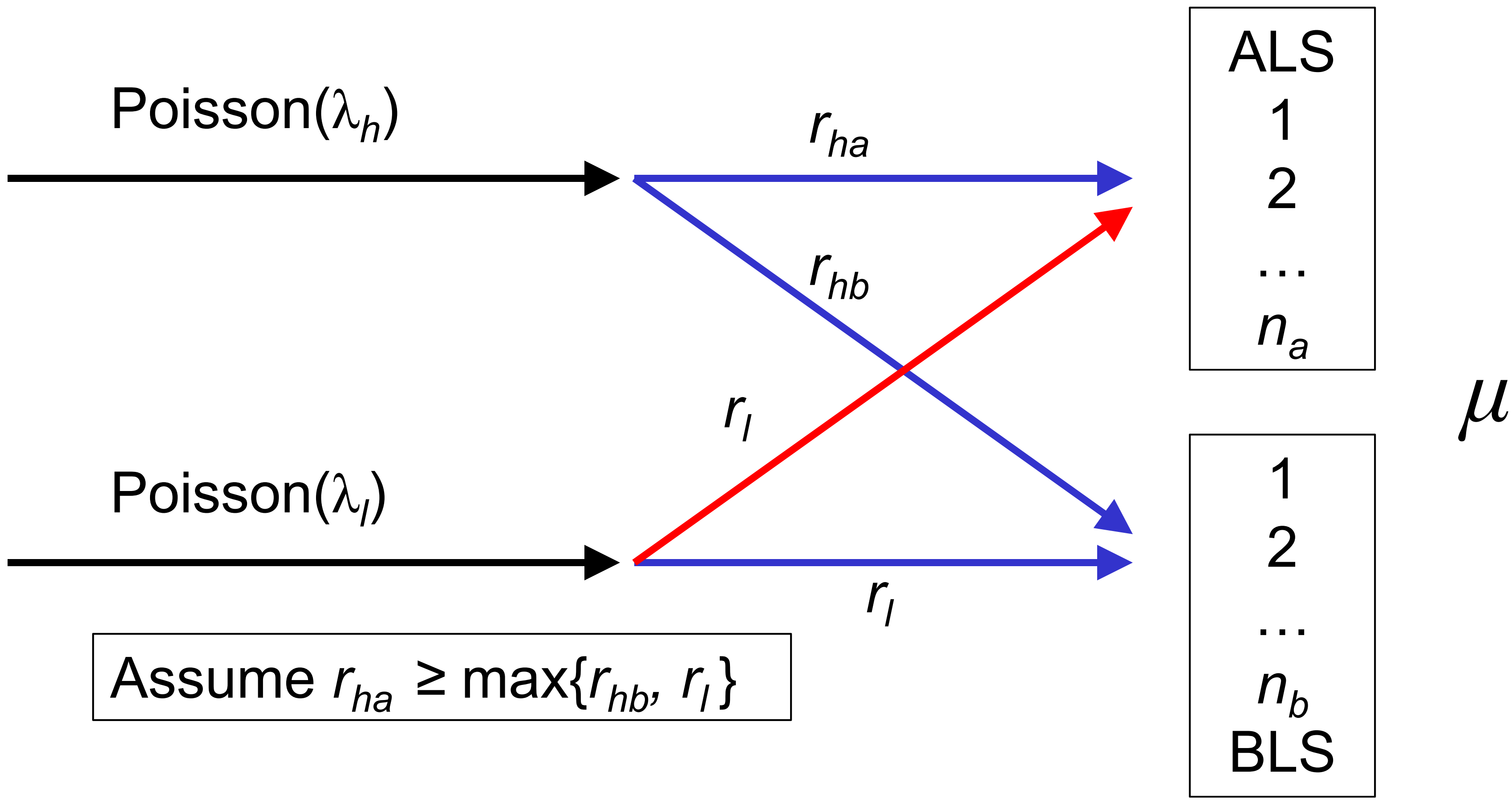
Objective Function

- For each (n_a, n_b), simulate to get "performance?"
  - Using what dispatching policy?
  - With what deployment across the city?
  - Using redeployment?
- Two Models:
  - Optimal dispatching (MDP)
  - Optimal deployment (IP)

MDP for Dispatching

- Two classes of server (ALS, BLS), two classes of call (high and low)
- Instead of P(respond in *x* minutes)
  Maximize E(reward)
- (Not the first to use MDPs for EMS)
  - E.g., Jarvis (1975), Berman (1981), Zhang (2012), McLay & Mayorga (2012)

The MDP


Poisson(λ_h)
Poisson(λ_l)
r_{ha}
r_{hb}
r_l
r_l
ALS
1
2
…
n_a
1
2
…
n_b
BLS
μ
Assume $r_{ha} \geq \max\{r_{hb}, r_l\}$

Additional Assumptions

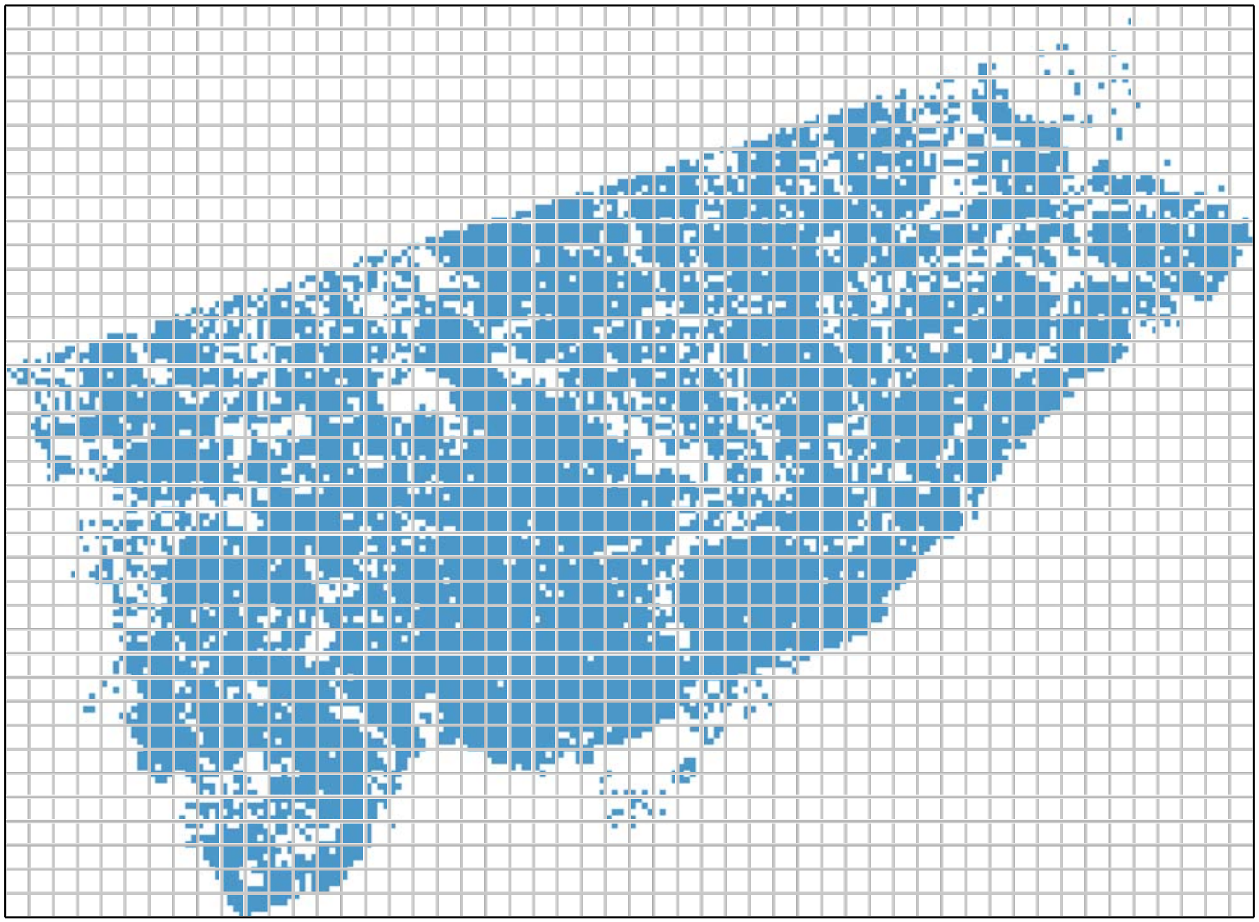
- High priority must get response if any ambulance is available of either type
- BLS ambulances can treat high-priority
  - Can also handle “delayed till ALS is free”
- Rates are constant in time
- No queueing
  - Redirected to allied service
- Service rates are the same for all combinations
  - Easily relaxed numerically

Dispatching Policy

- State space {0, 1, …, n_a} x {0, 1, …, n_b}
- State (i, j): i ALS and j BLS are busy
- Only decision: Respond to low priority call with ALS if all BLS are busy?
- Maximize long-run average reward
- We have structural results, but for this work numerical results are of interest

Data from Toronto EMS

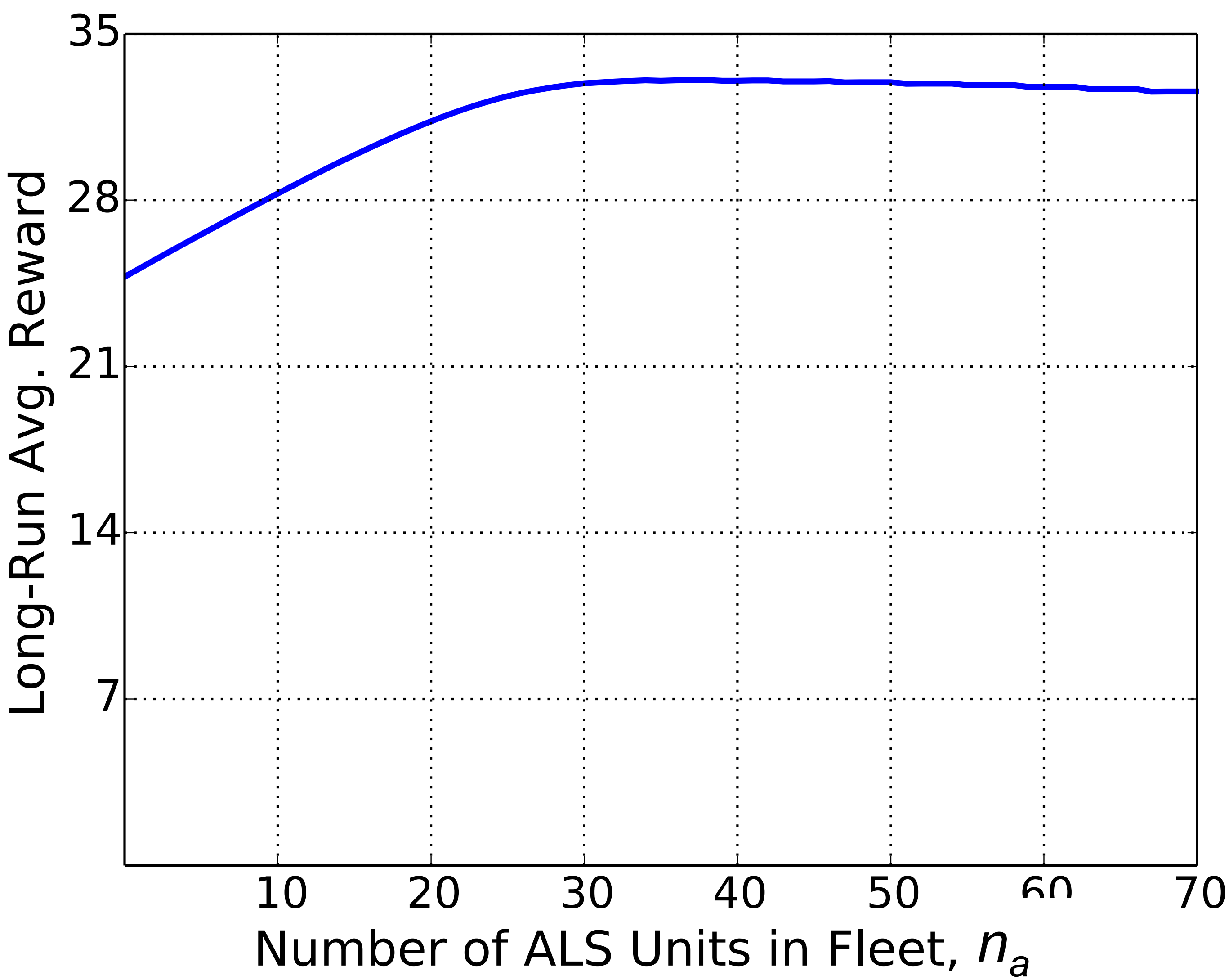
- 371,903 records from 1/1/07 – 31/12/08



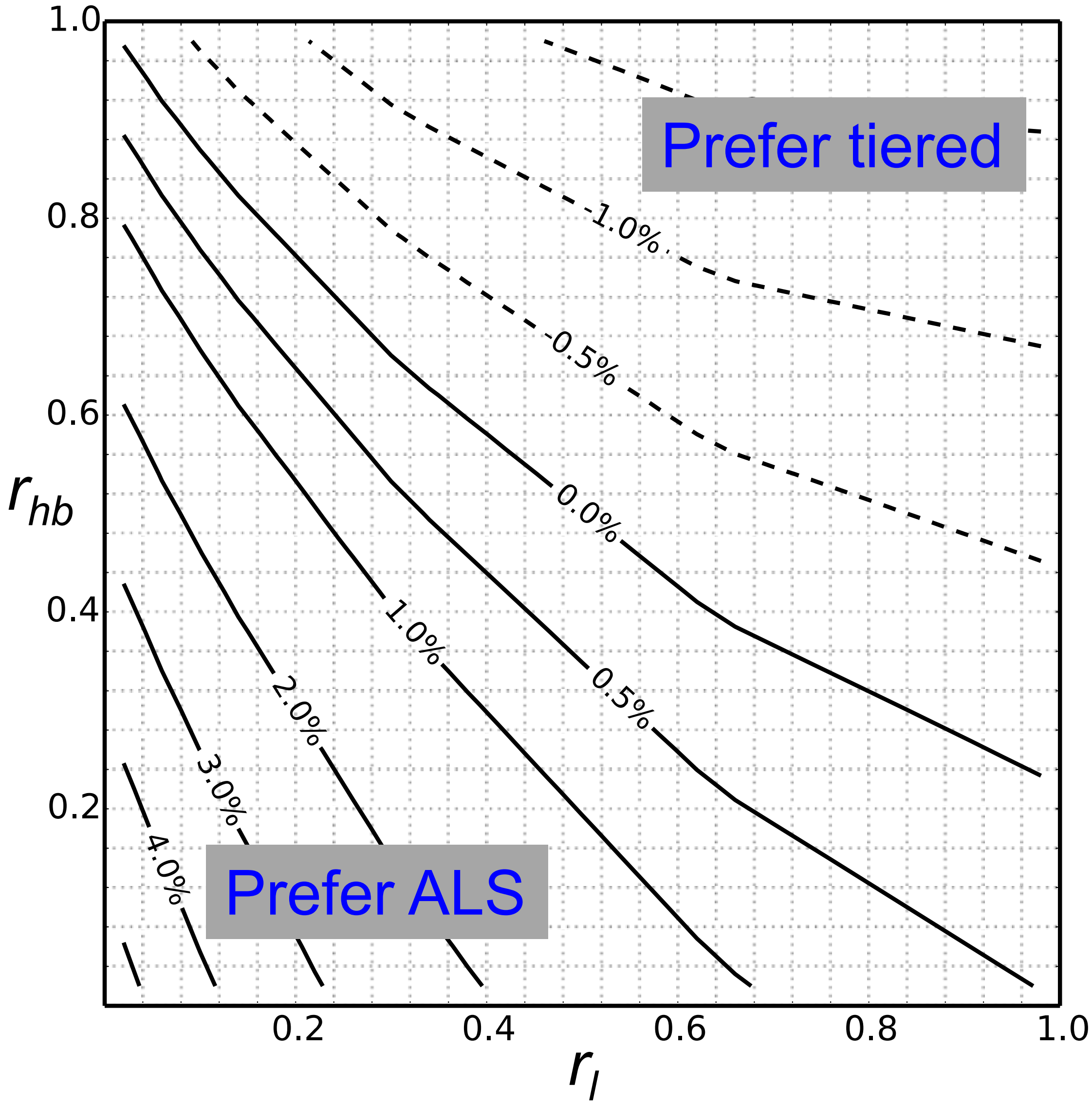
Input Parameters

- Rates: $\lambda_h = 8$, $\lambda_l = 13$, $\mu = ¾$ per hr (mean service time 80 min)
- Rewards $r_{ha} = 1$, $r_{hb} = 0.5$, $r_l = 0.6$
- Costs $c_a = 1.25$, $c_b = 1$, $b = 87.5$
- Vehicle mixes we evaluate:
  - $\{(n_a, n_b): n_a \leq 70, n_b = \text{max possible}\}$
  - {(0, 87), (1, 86), (2, 85), (3, 83), …}

Results


35
28
21
14
7
10
20
30
40
50
60
70
Long-Run Avg. Reward
Number of ALS Units in Fleet, n_a

Robustness: Rewards



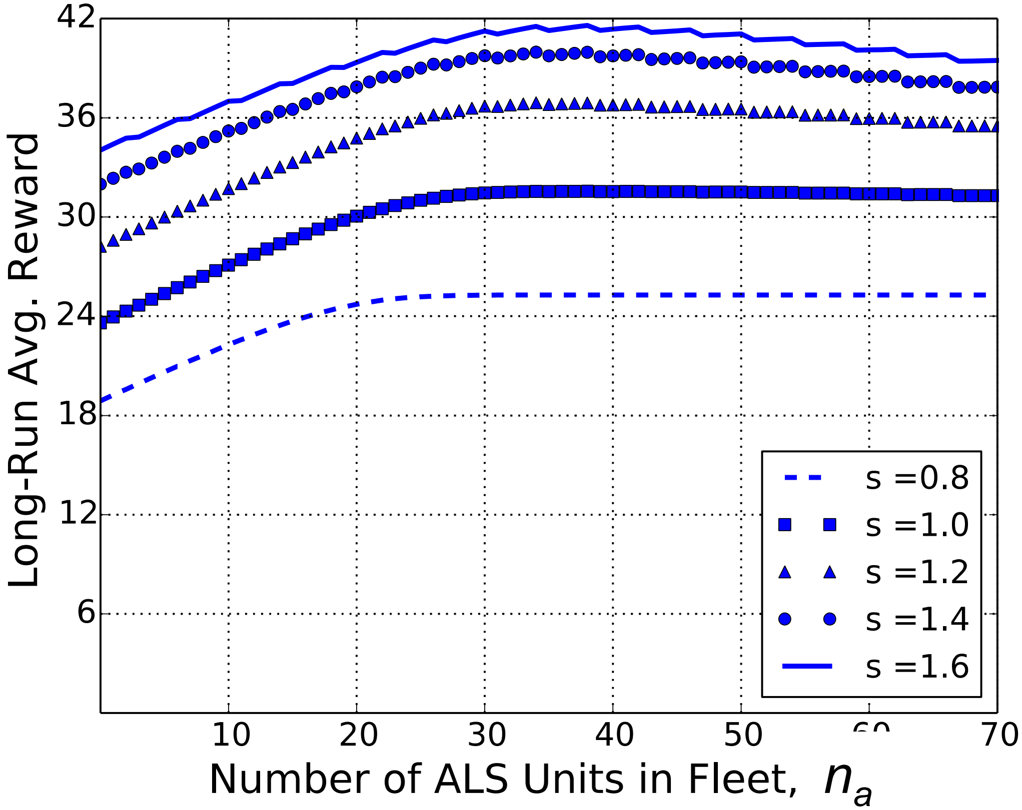
All ALS fleet (70, 0) versus tiered system (27, 53)

$r_{ha} = 1$

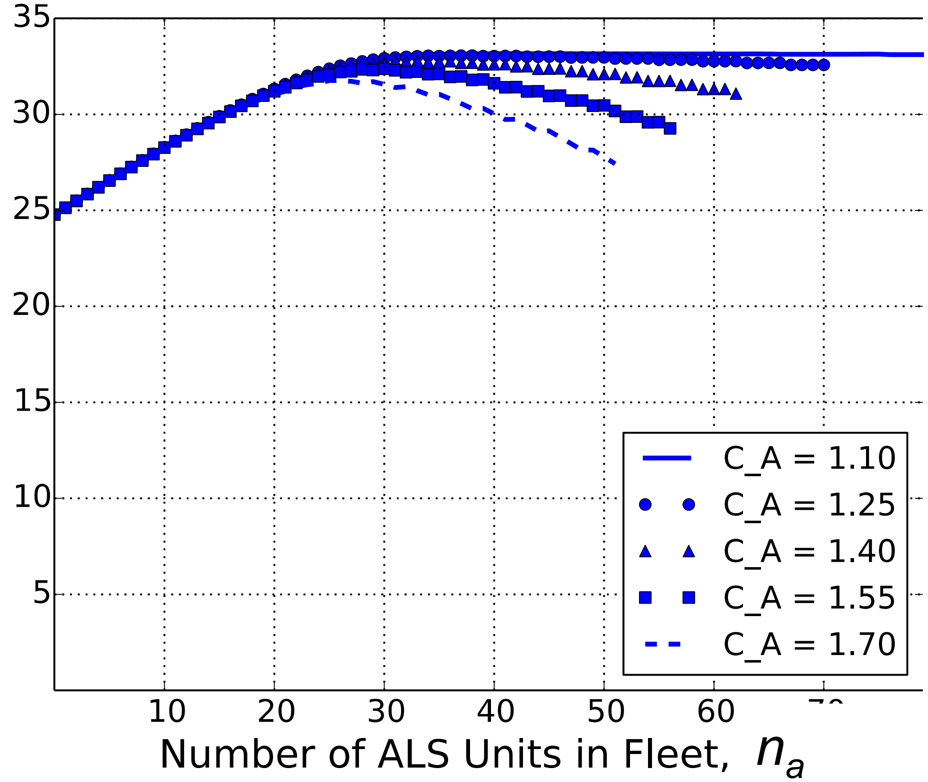
Performance of tiered fleets relative to ALS is fairly insensitive to reward values

Robustness II

Scale arrival rates



Change cost of ALS



Criticism

- The MDP ignores geography!
  - (To allow dispatching complexity)
- Allows complete pooling of fleet
- Do the conclusions change if we take account of geography?
- To take account of geography (deployment), need to simplify dispatch

Integer Programming

- Road network is a graph (N, E)
- Arrival rates at node i : λ_i^h, λ_i^l
- Call response
  - T = response-time threshold (9min – call handling, turnout = 7min or so)
  - t_{ij} = travel time between nodes i and j
  - C_i = Neighbourhood of $i = \{j: t_{ij} \leq T\}$

Integer Programming

- Model related to MEXCLP (Daskin)
  - Busy probabilities p_a, p_b for each amb
  - Ambulances independently busy
- No call queueing
- Fraction of low priority calls receiving ALS response because all BLS are busy = q (approximated from MDP)

Decision Vars and Objective

- x_i^a, x_i^b = # ALS, BLS at Node i
- y_{iab} = 1 if Node *i* covered by *a* ALS and *b* BLS exactly, 0 otherwise
- When y_{iab} = 1, collect reward at rate

$$\lambda_i^h\, r(h, a, b) + \lambda_i^l\, r(l, a, b),$$

where

$$r(h, a, b) = r_{ha}\,(1\text{-}p_a{}^a) + r_{hb}\, p_a{}^a\,(1 - p_b{}^b)$$

$$r(l, a, b) = r_l\,(1\text{-}p_b{}^b + p_b{}^b\,(1 - p_a{}^a)\; q)$$

Integer Program

$$\max \quad \sum_{i\in N}\sum_{a=0}^{n_a}\sum_{b=0}^{n_b}(\lambda_i^h r(h,a,b) + \lambda_i^l r(l,a,b))y_{iab}$$

$$\text{s.t.} \quad \sum_{i\in N} x_i^a \le n_a$$

$$\sum_{i\in N} x_i^b \le n_b$$

$$\sum_{a=0}^{n_a} a \sum_{b=0}^{n_b} y_{iab} \le \sum_{j\in C_i} x_j^a \qquad \forall i \in N$$

$$\sum_{b=0}^{n_b} b \sum_{a=0}^{n_a} y_{iab} \le \sum_{j\in C_i} x_j^b \qquad \forall i \in N$$

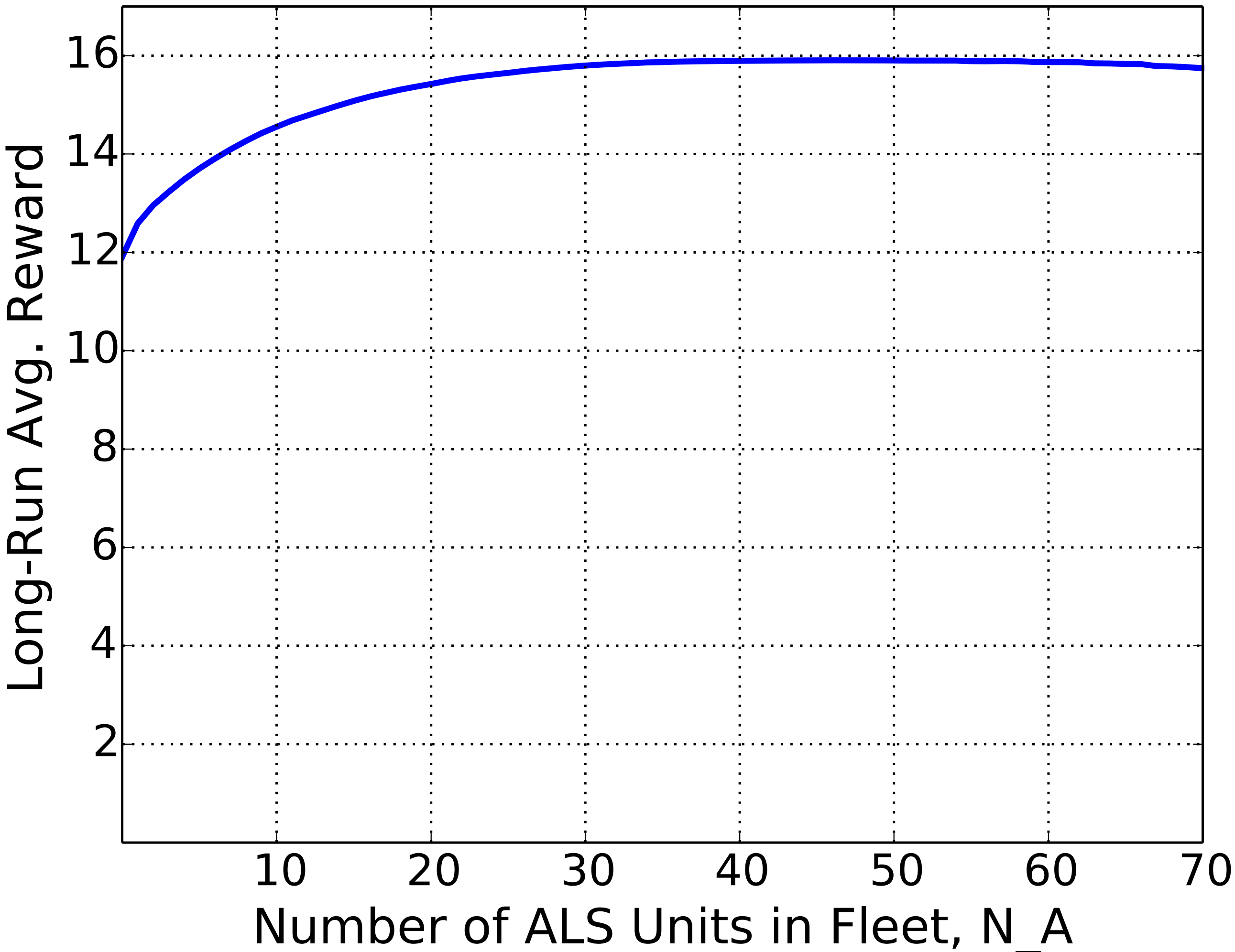
$$\sum_{a=0}^{n_a}\sum_{b=0}^{n_b} y_{iab} \le 1 \qquad \forall i \in N$$

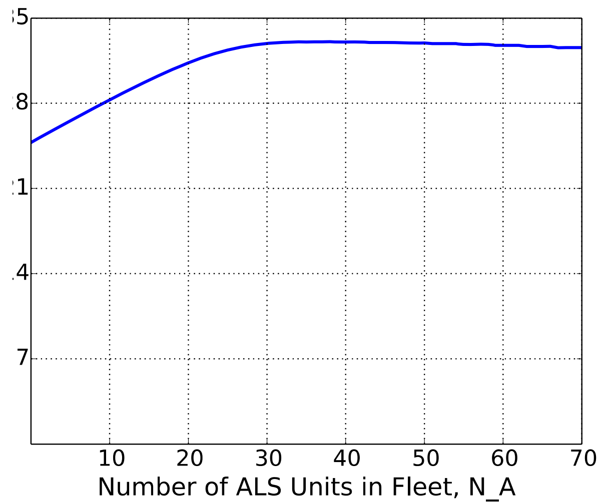
$$x_i^a,\ x_i^b,\ y_{iab} \in \mathbb{Z}_+$$

Getting Integer Solutions

- Hard to solve IP, so use randomized rounding (Williamson & Shmoys)
  - Solve LP relaxation
  - Interpret x's as expected number of ambulances at that location, y's similarly
  - Repeat:
    - Generate consistent random deployment
    - Compute objective function
- Optimality gap almost always << 1%

Results (52 x 38 nodes)





(Robustness is similar to MDP)

Mixed Fleets?

- A wide range of tiered fleets can perform comparably (or outperform) an all-ALS fleet
- So can base the decision on other factors
  - History/politics
  - Paramedic (or in NL, nurse) availability
  - Maintaining skills of paramedics
- Provided that you dispatch/deploy well

Bounding Potential Performance

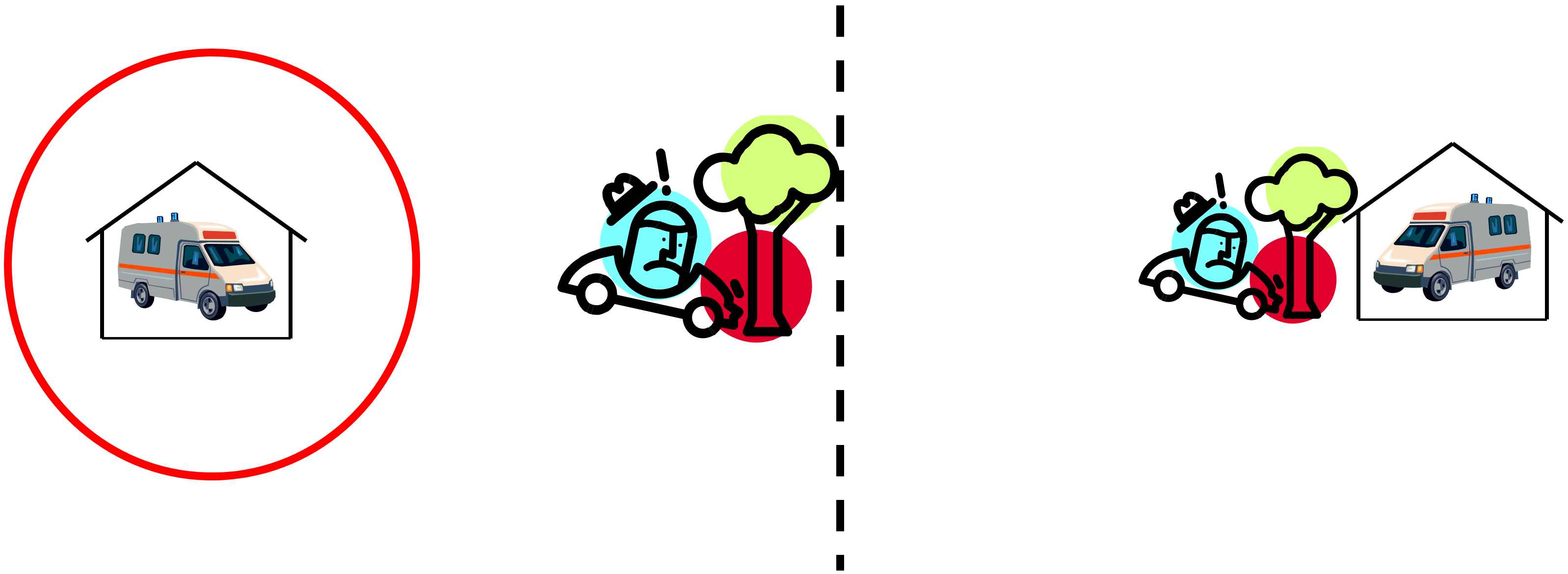
- Can't solve deployment and dispatch at same time, but maybe we could compute bounds and optimize the bounds?
- Can competitor's bid achieve promised performance?
- Can redeployment ensure good performance or do we need to take "other steps?"
- When, as researchers/managers, should we stop looking for improvements?
- The following only works for all-ALS
- Need lower bound on Prob(late call)

A Bound?

Each time a call comes in

- don't look at location yet
- pretend available ambulances are in locations that minimize the fraction of calls outside 9 minutes travel
- Pretend ambulance responds from those locations

Not a Bound

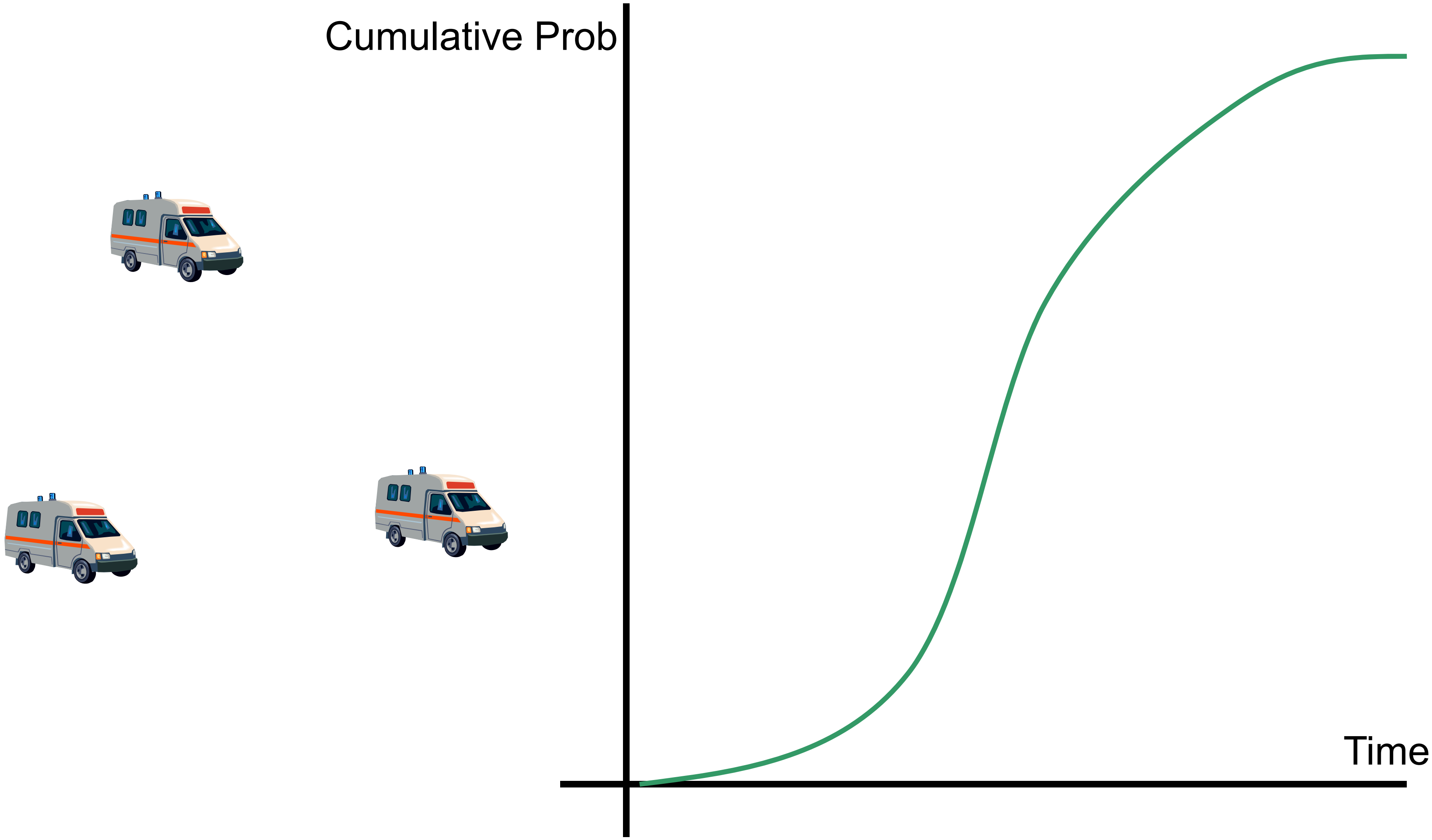


LHS is optimal for next call, but means much more workload. So RHS may be optimal overall

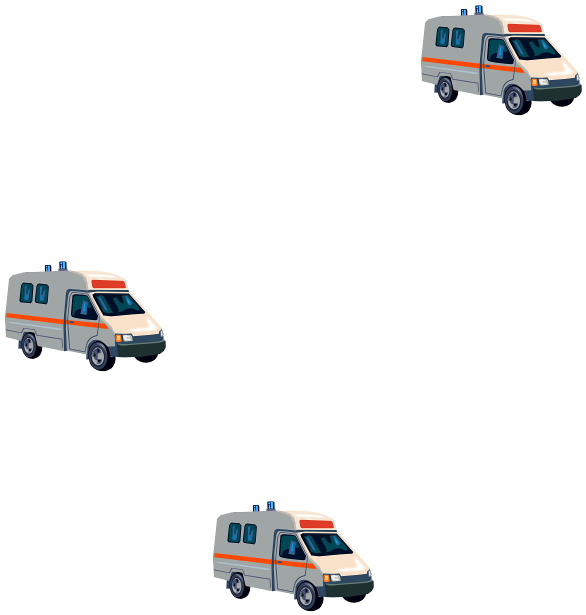
A Lower Bound

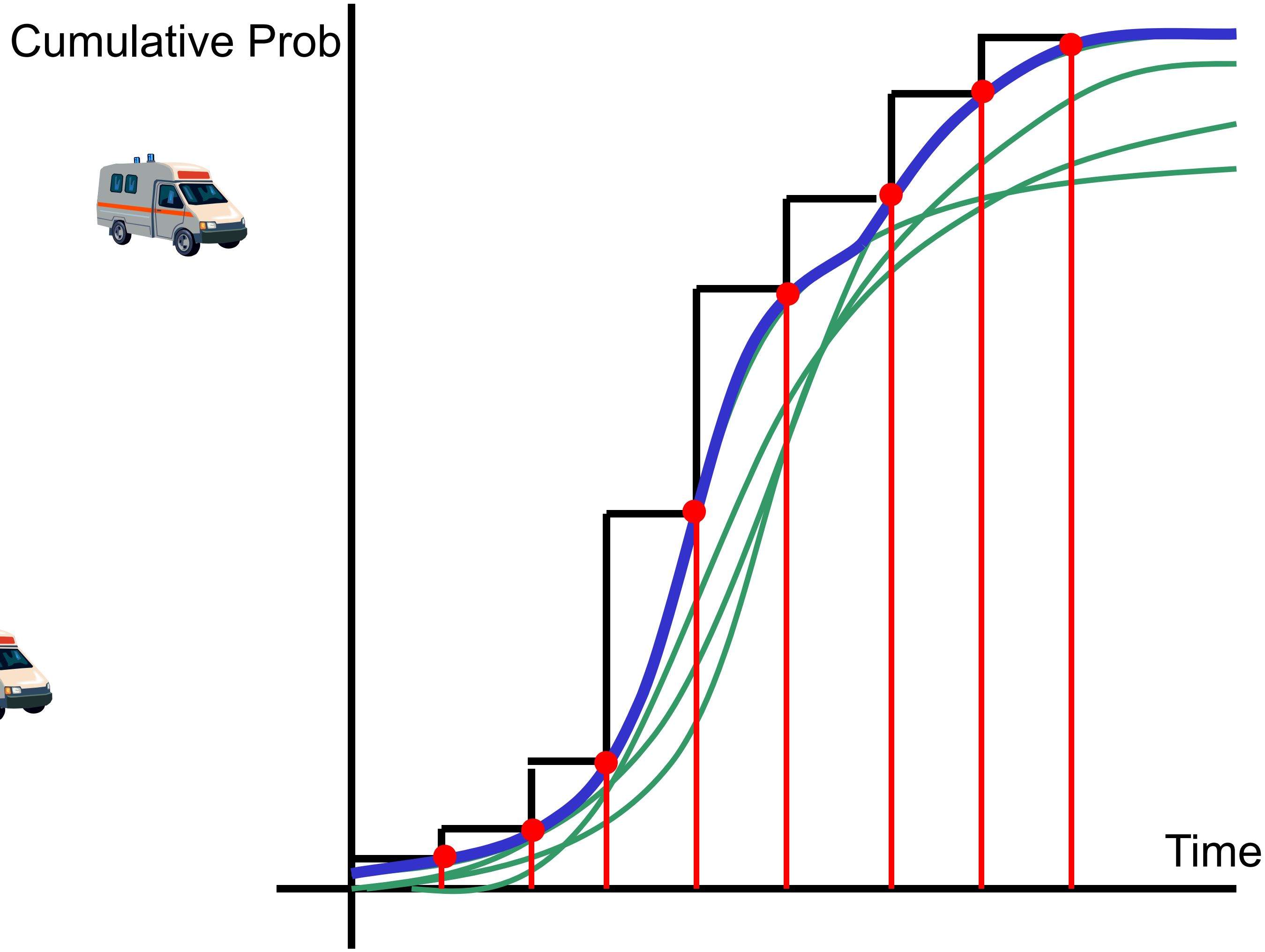
- Whenever call comes in, pretend available ambulances in optimal locations, and compute Prob(reach call)
  - Solve an IP (Church & Revelle ‘74) for each # of available ambulances
- Ensure that always have more ambulances available than in reality. (Coupling)
  - Ambulances are a queueing system
  - Construct a bounding queueing system with “smaller” service times (depend on # free ambs)
  - Simulate bounding queueing system

Stochastic Lower Bound



Stochastic Lower Bound





$$\max \sum_{j=1}^{J} d_j p_j$$

Compute P(service time ≤ τ)

$$\text{s.t.} \sum_{i=1}^{J} x_i \leq m$$

Assign *m* ambulances

$$y_{ij} \leq x_i$$

Respond to *j* from *i* only if *i* has an ambulance

$$\sum_{i=1}^{J} y_{ij} = 1$$

Respond to *j* from *somewhere*

$$p_j = \sum_{i=1}^{J} F_j(\tau - t_{ij}) y_{ij}$$

Compute P(service time$_j$ ≤ τ)

$$x_i \in \{0,1\}, y_{ij} \in \{0,1\}, p_j \in [0,1]$$

Results

For realistic but not real Edmonton model

1. Good static policy: 23.9% late
2. Redeploy: 18.8%
3. Redeploy (extra moves): **16.6%**
4. Lower Bound: **15.1%**

1-2% = 1 amb, #ambs 16

For realistic but not real Melbourne model

1. Good static policy: approx 19% late
2. Redeploy (extra moves): **17.5%**
3. Lower Bound: **11.2%**

#ambs 95

Bounds for Tiered Fleets

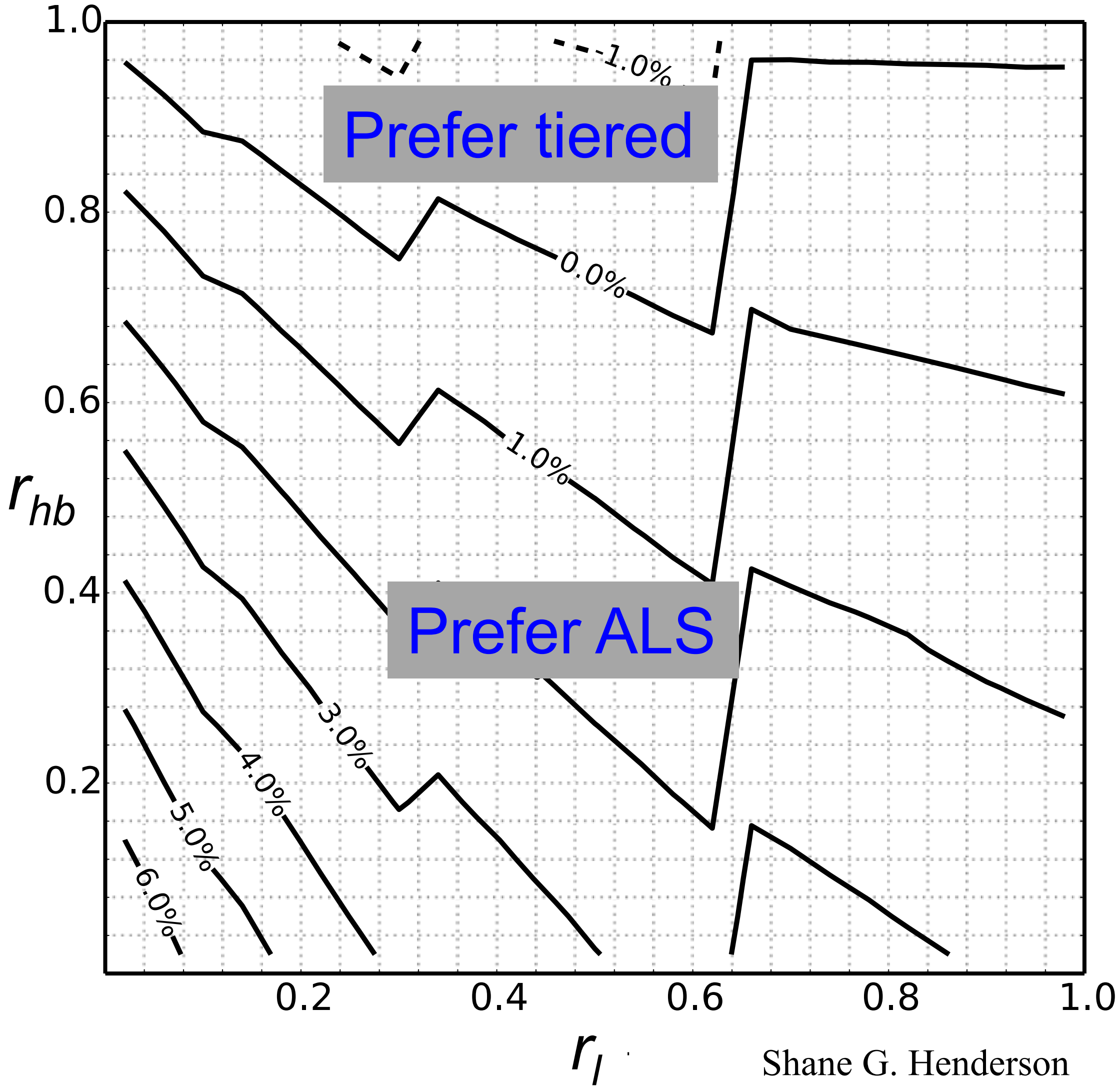
- Bounds on what?
  - Expected long-run reward?
  - Prob(on time with high) s/t bound on low?
  - Expected penalty for late calls? (most tractable)
- Coupling as used here *might* work…
- But Brown, Smith, Sun (2010) seems much more likely, for all-ALS too

Conclusions

- Vehicle mix
  - Tiered fleets just as good as, or better than, all-ALS, provided that fleet has enough ALS
  - Difference is small for well-managed systems
  - Can think about other issues to decide
- Redeployment bound
  - Requires some computation
  - Useful, but hard work to compute
  - *Looking for other ways to compute bounds, and to improve policies, particularly for tiered systems*
- http://people.orie.cornell.edu/~shane

Spares

Robustness: Rewards



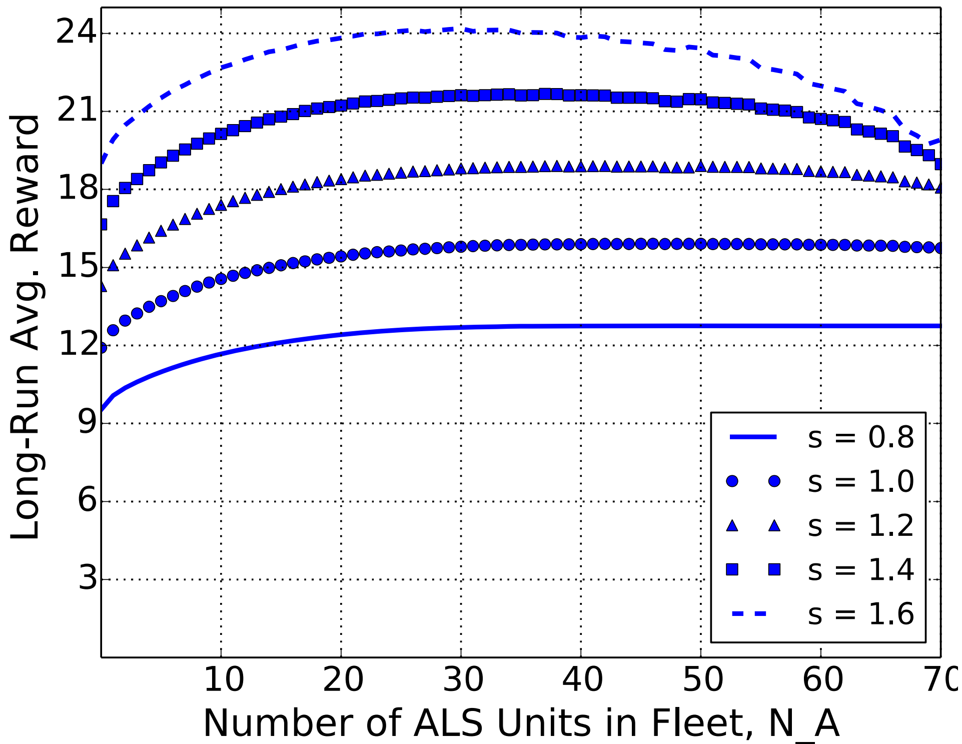
All ALS fleet (70, 0) versus tiered system (27, 53)

r_{ha} = 1

Performance of tiered fleets relative to ALS is fairly insensitive to reward values

Robustness II

Scale arrival rates



Change cost of ALS