BGRP

(Border Gateway Reservation Protocol)

A Tree-Based Aggregation Protocol for Inter-Domain Reservations

Ping Pan
Bell Labs + Columbia

Ellen Hahne
Bell Labs

Henning Schulzrinne
Columbia

Outline

- Resource Reservation
  - Applications
  - Architectures
  - Challenges
- Protocol Scaling Issues
- BGRP Protocol
  - Major Messages
  - Performance
- Conclusions & Future Work

Resource Reservation

- Applications
- Old Architecture: Int Serv + RSVP
- Challenges
- New Architecture: Diff Serv + BGRP

Reservation Applications

- Real-Time QoS
  - Voice over IP
  - Video
- Virtual Private Networks
- Differentiated Services
  - Better than Best Effort
- Traffic Engineering
  - Offload congested routes
  - Integration of ATM, Optical & IP (MPLS)
  - Inter-Domain Agreements

Reservation Architectures

- Old Solution: Int Serv + RSVP
  - End-to-end
  - Per-flow
- Challenges
  - Data Forwarding Costs
  - Protocol Overhead
  - Inter-Domain Administration
- New Solution: Diff Serv + BGRP
  - Aggregated
  - Scalable
  - Manageable

Two Scaling Challenges

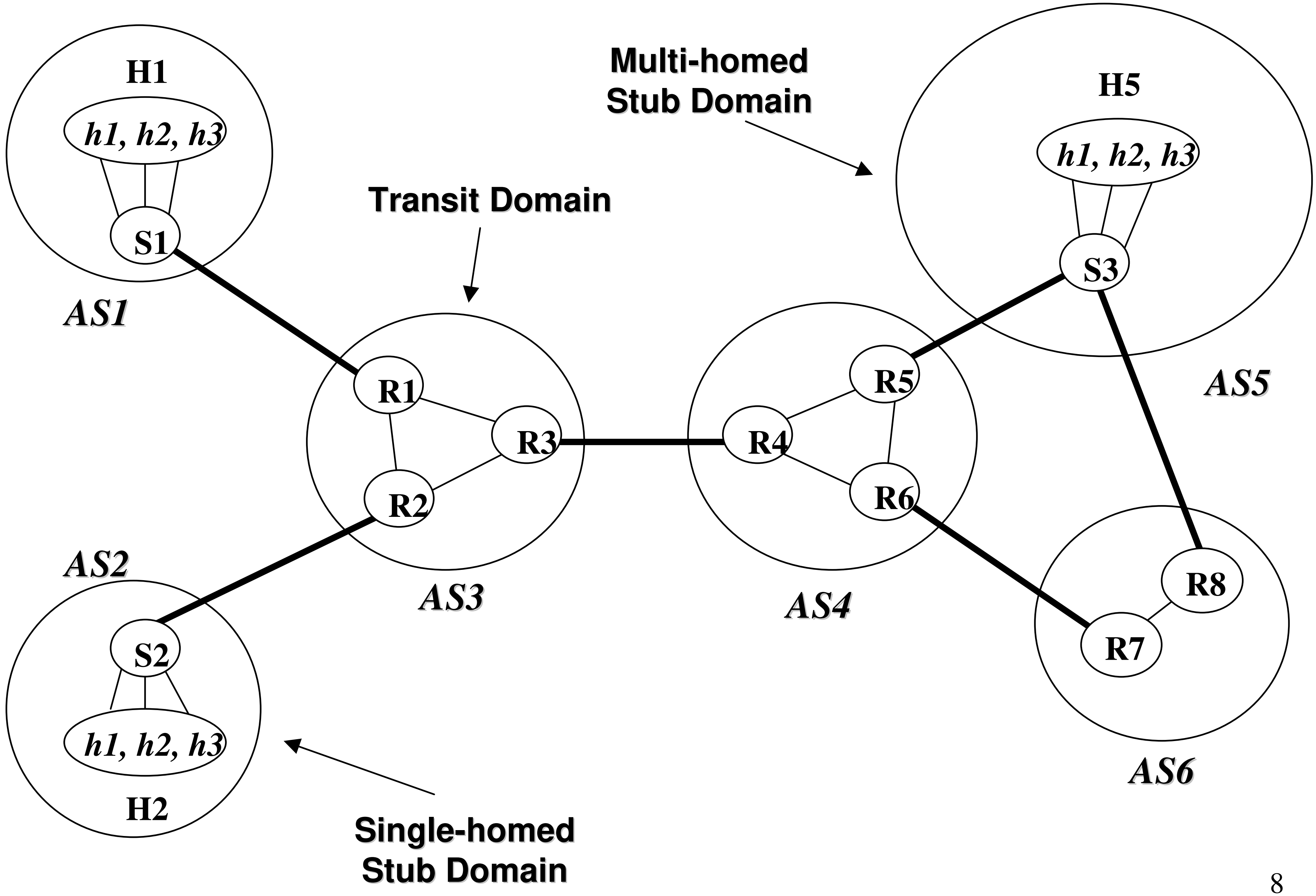
- Data Forwarding Costs
  - Int Serv: per micro-flow
  - Diff Serv: ~32 AF/EF Code Points
  - Solves that problem !

- Control Protocol Overhead
  - RSVP: $O(N^2)$, N = # hosts
  - BGRP: $O(N)$, N = something much smaller
  - Much more to say about this !

Protocol Scaling Issues

- Network Structure
- Network Size
- How much Aggregation?
- How to Aggregate?

Network Structure: Multiple Domains (AS)



Current Network Size

- 10^8 (60,000,000) Hosts
- 10^5 (60,000) Networks
- 10^4 (6,000) Domains

Traffic Trace

(90-sec trace, 3 million IP packet headers, at MAE-West, June 1, 1999)

Granularity	# Sources	# Destinations	# Source-Destination Pairs
Application	(Addr + Port) 143,243	(Addr + Port + Proto) 208,559	(5-tuple) 339,245
Host	56,935	40,538	131,009
Network	13,917	20,887	79,786
Domain	2,244	2,891	20,857

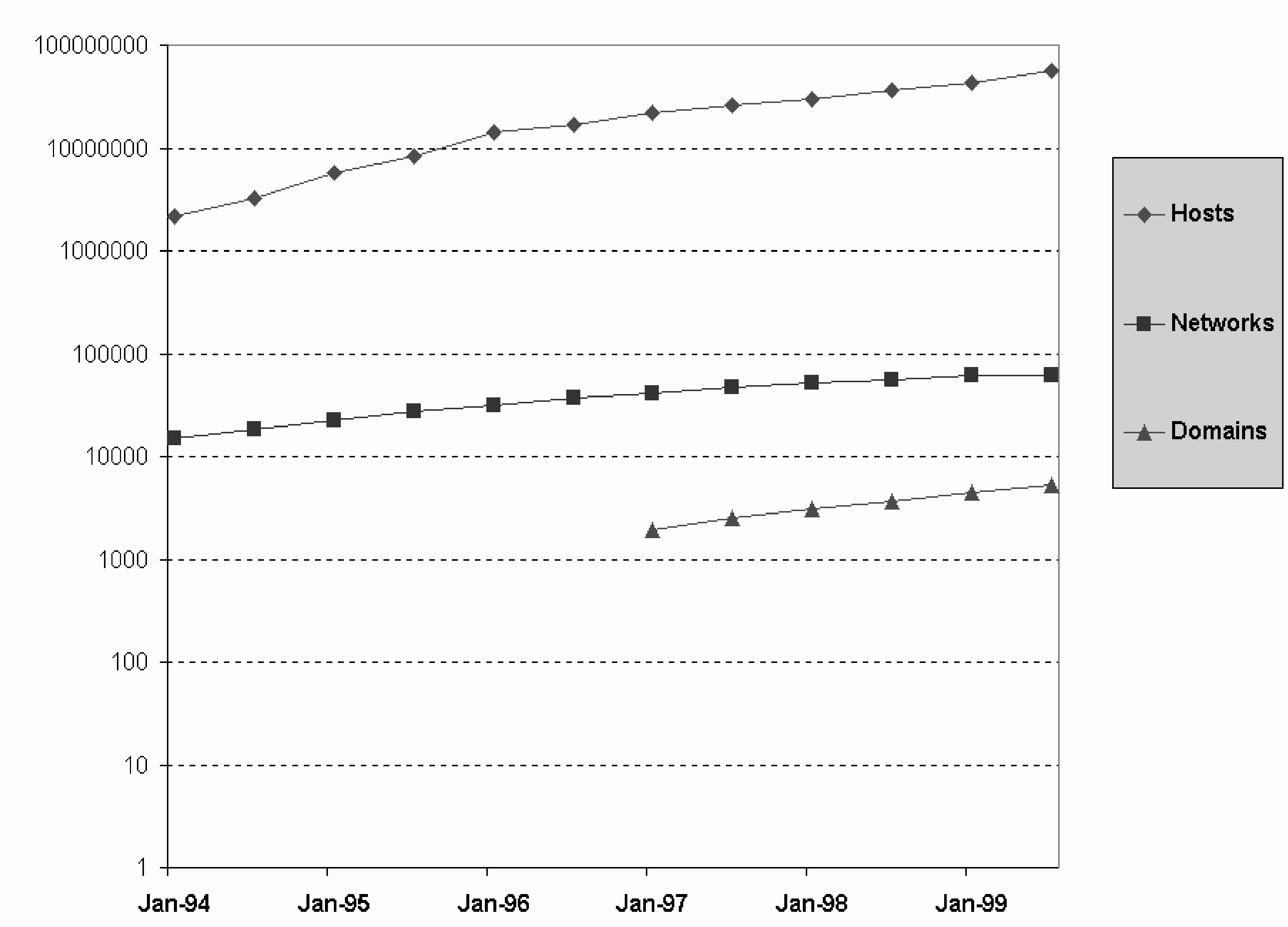
Traffic Trace

- Over 1-month span (May 1999) at MAE-West:
  - 4,908 Source AS seen
  - 5,001 Destination AS seen
  - 7,900,362 Source-Destination pairs seen!

How many Reservations? (How much aggregation?)

- 1 reserv'n per source-dest'n pair?
  - 10^{16} host pairs
  - 10^{10} network pairs
  - 10^{8} domain pairs
- 1 reserv'n per source OR 1 reserv'n per dest'n?
  - 10^{8} hosts
  - 10^{5} networks
  - 10^{4} domains
- Router capacity: 10^{4} < # Reserv'ns < 10^{6}
- Conclusion: 1 reserv'n per Network or Domain for each Diff Serv traffic class

Network Growth (1994-1999)



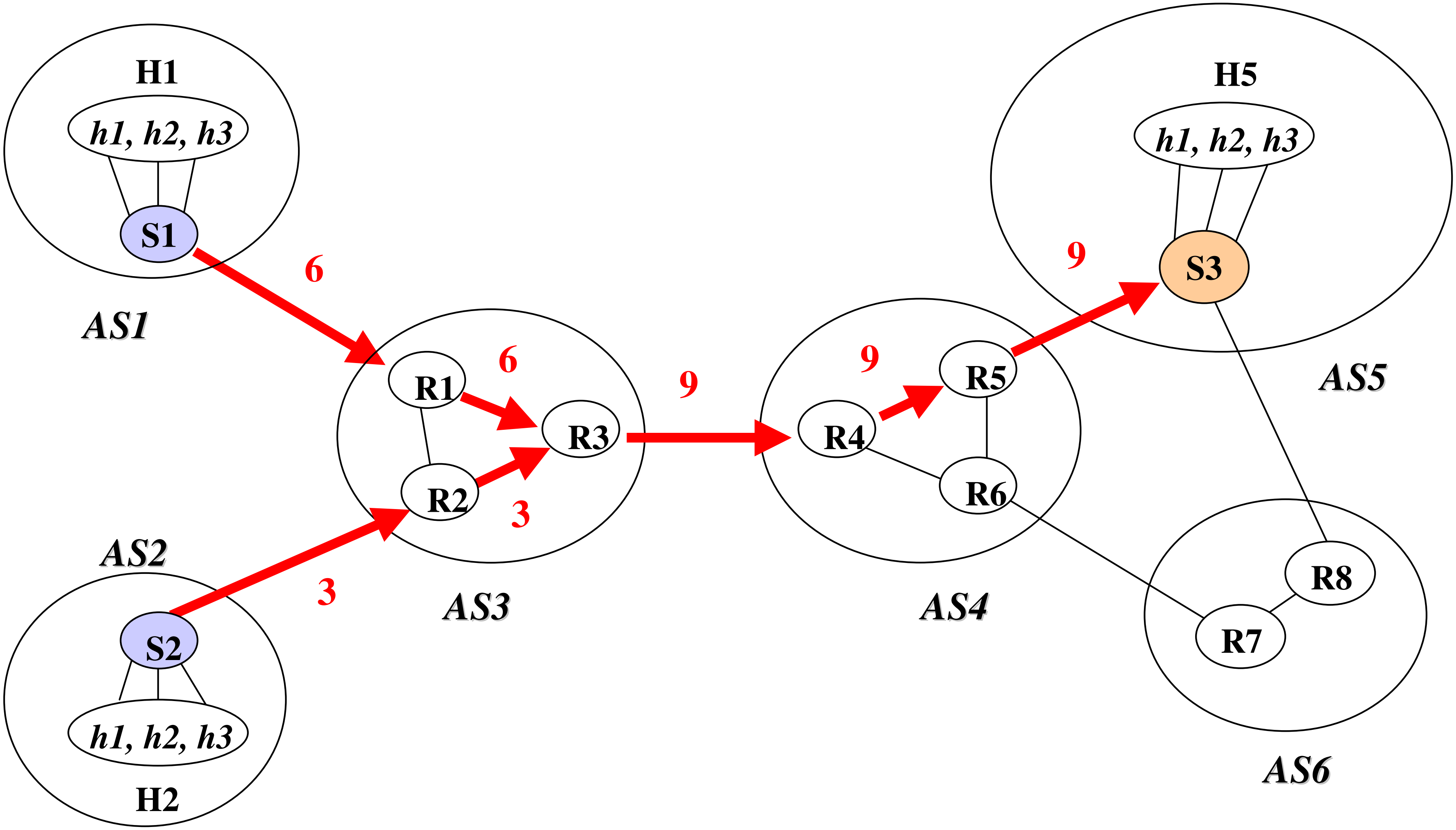
Growth Rates

- Graph has a Log Scale
- H (# Hosts) : Exponential growth
- D (# Domains) : Exponential growth
- Moore's Law can barely keep up!
- Overhead of control protocols?
  - $O(H)$ or $O(D)$, May be OK
  - $O(H^2)$ or $O(D^2)$, Not OK !

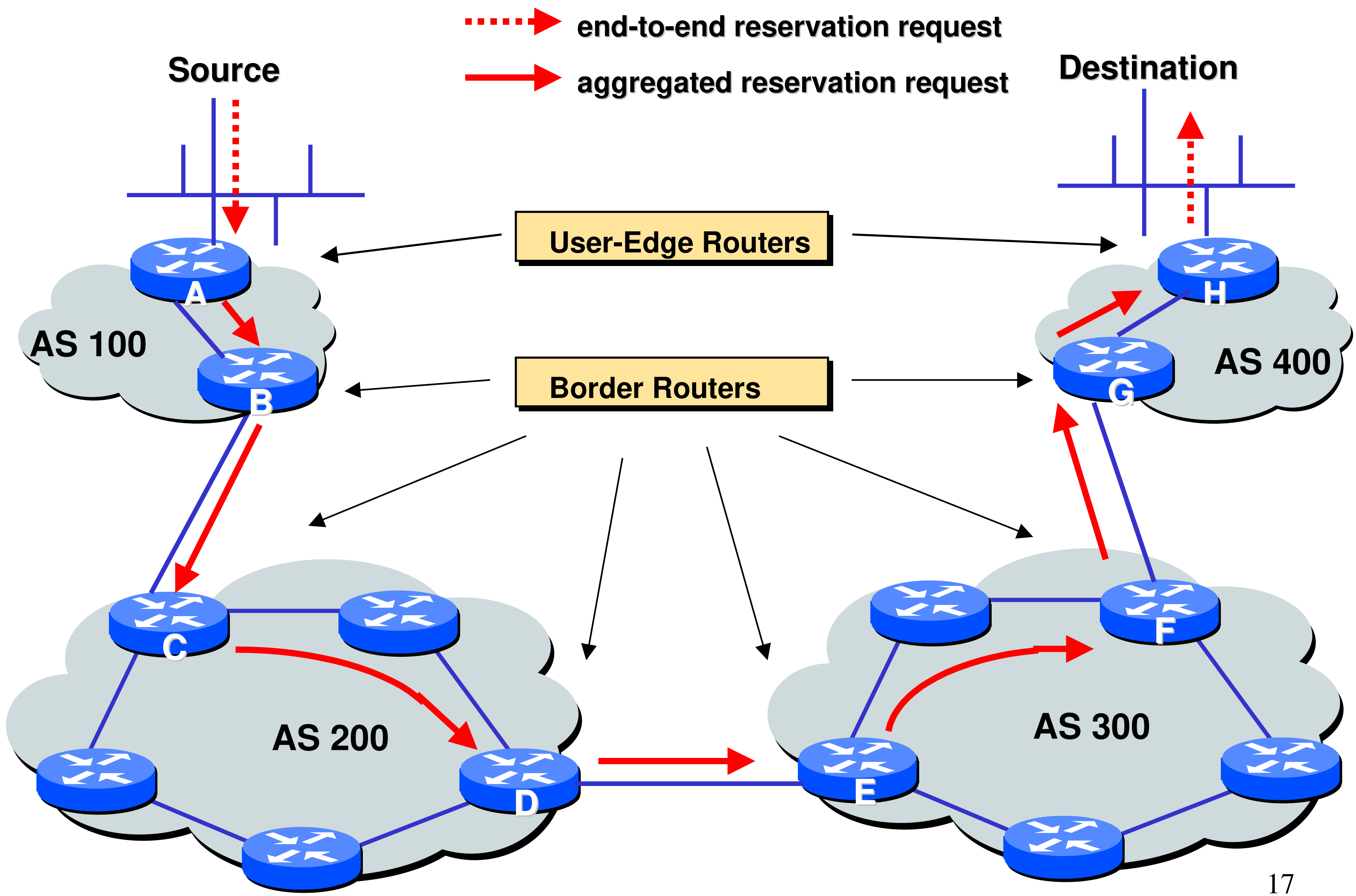
How to Aggregate?

- Combine Reserv'ns from *all* Sources
  to *1* Dest'n for *1* Diff Serv class
- Data & Reserv'ns take *BGP* route to Dest'n
- BGP routes form *Sink Tree* rooted at
  Dest'n domain (no load balancing)
- Aggregated Reserv'ns form *Sink Tree*
- Where 2 branches meet, *Sum* Reserv'ns

A Sink Tree rooted at S3



How to handle end-user reservation?



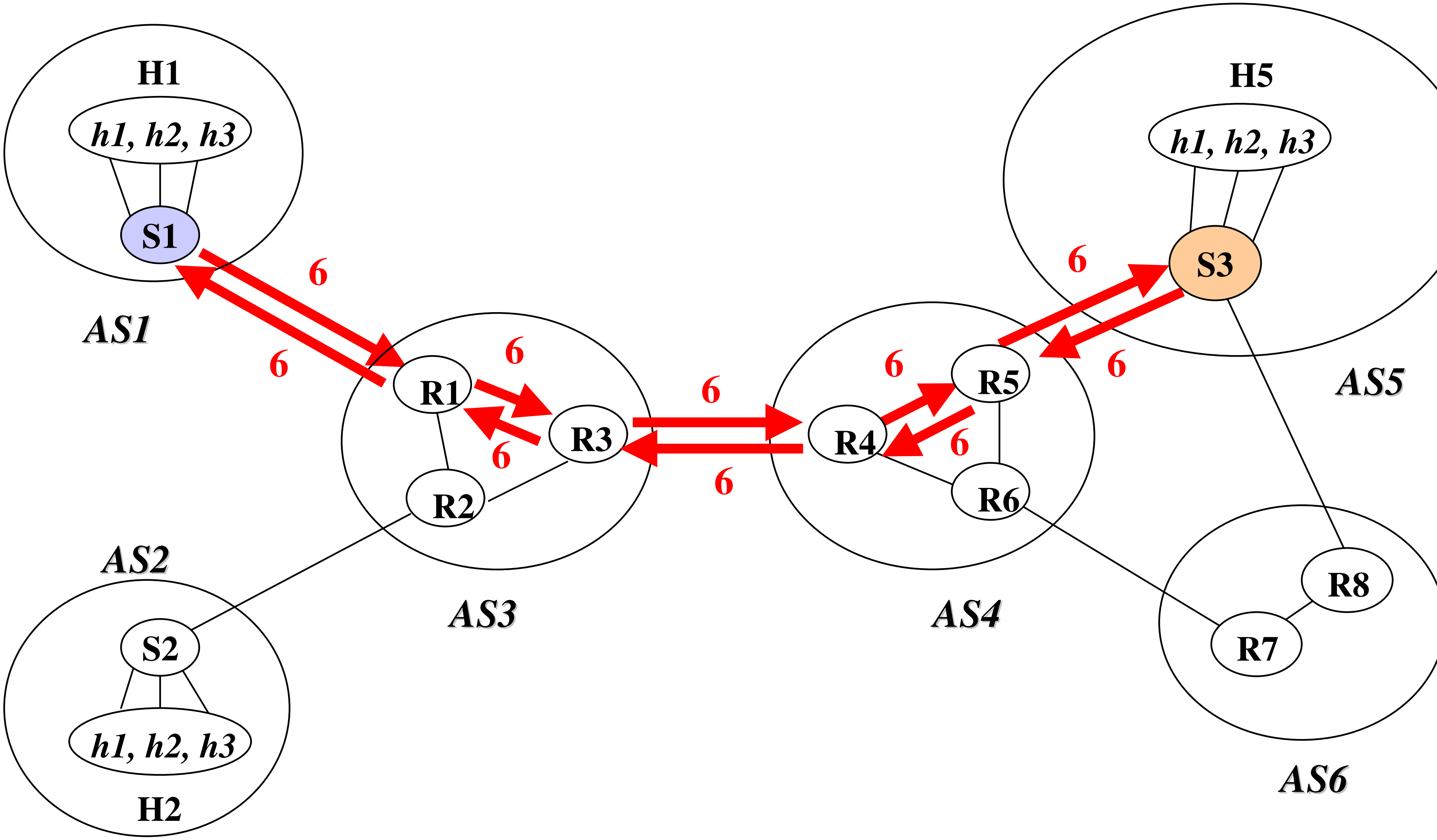
BGRP Protocol

- Basic Operation
- Comparison with RSVP
- Enhancements
- Performance Evaluation

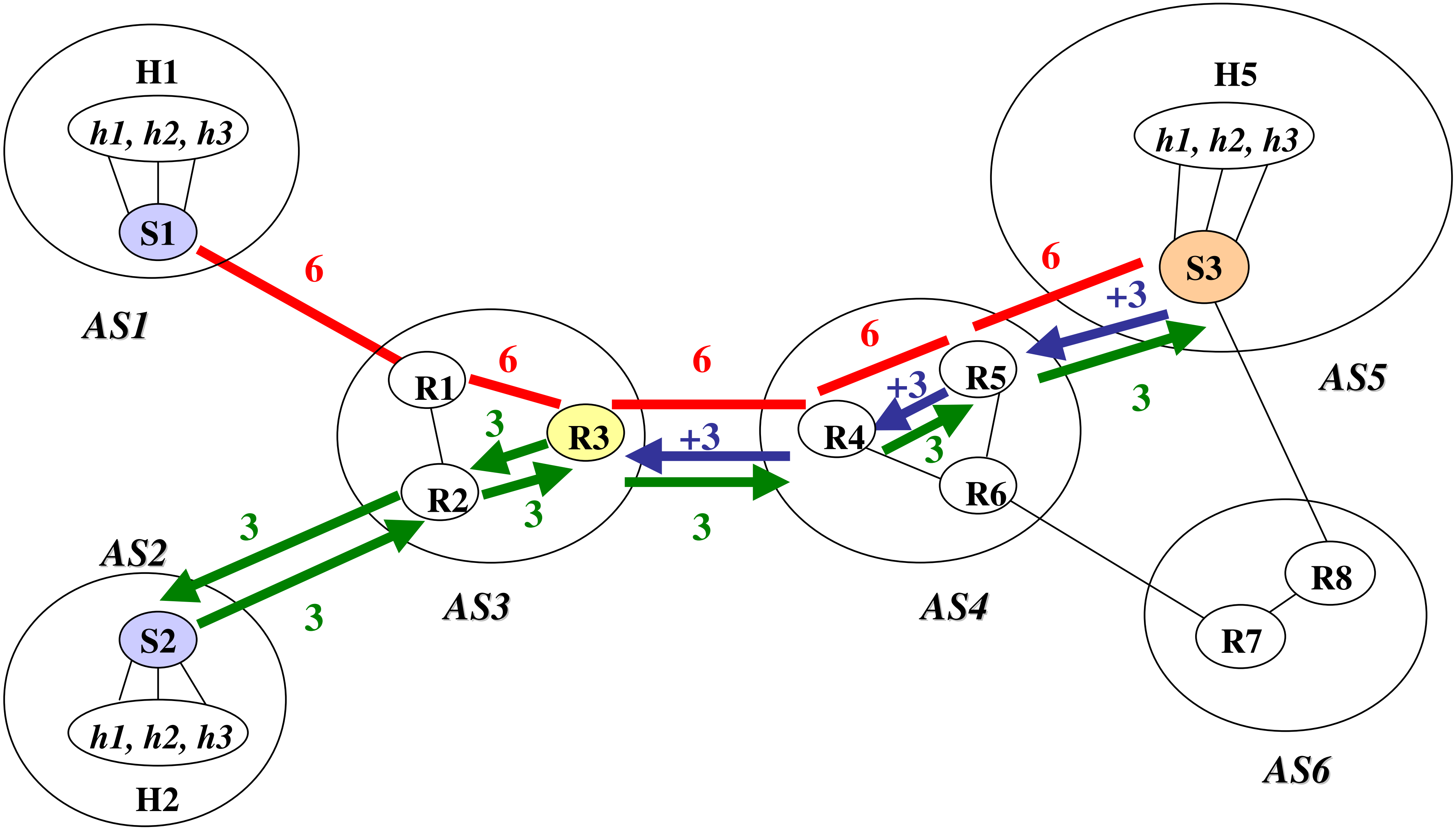
BGRP Basics

- Inter-Domain only
- Runs between Border Routers
- Follows BGP Routes
- Reserves for Unicast Flows
- Aggregates Reserv'ns into Sink Trees
- Delivers its Messages Reliably
- 3 Major Messages
  - *Probe:* source to dest'n; reserv'n path discovery
  - *Graft:* dest'n to source; reserv'n establishm't & aggreg'n
  - *Refresh:* adjacent routers; reserv'n maintenance

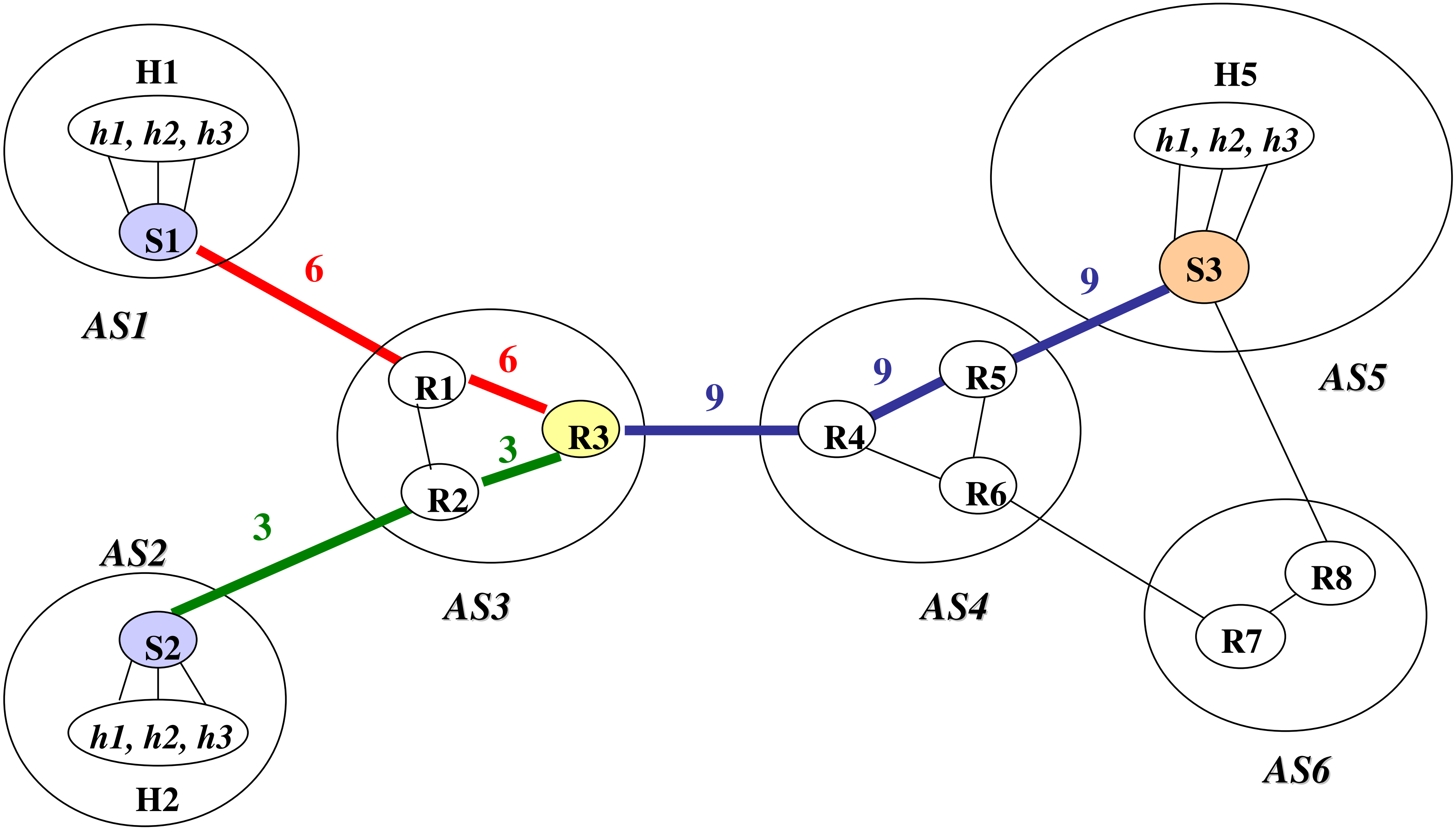
Tree Construction: 1st Branch



Tree Construction: 2nd Branch



Tree Construction: Complete



PROBE Message

- Source (leaf) toward Destination (root)
- Finds reservation path
- Constructs Route Record:
  - Piggybacks Route Record in message
  - Checks for loops
  - Checks resource availability
- Does *not* store path (breadcrumb) state
- Does *not* make reservation

GRAFT Message

- Destination (root) toward Source (leaf)
- Uses path from *PROBE's* Route Record
- Establishes reservations at each hop
- Aggregates reservations into sink tree
- Stores reservation state *per-sink tree*

REFRESH Message

- Sent periodically
- Between adjacent BGRP hops
- Bi-directional
- Updates all reserv'n state in 1 message

Comparison of BGRP vs. RSVP

- Probing:
  - BGRP PROBE vs. RSVP PATH
  - Stateless vs. Stateful [$O(N^2)$]
- Reserving:
  - BGRP GRAFT vs. RSVP RESV
  - State-light [$O(N)$] vs. Stateful [$O(N^2)$]
  - Aggregated vs. Shared
- Refreshing:
  - Explicit vs. Implicit
  - Bundled vs. Unbundled

BGRP Enhancements

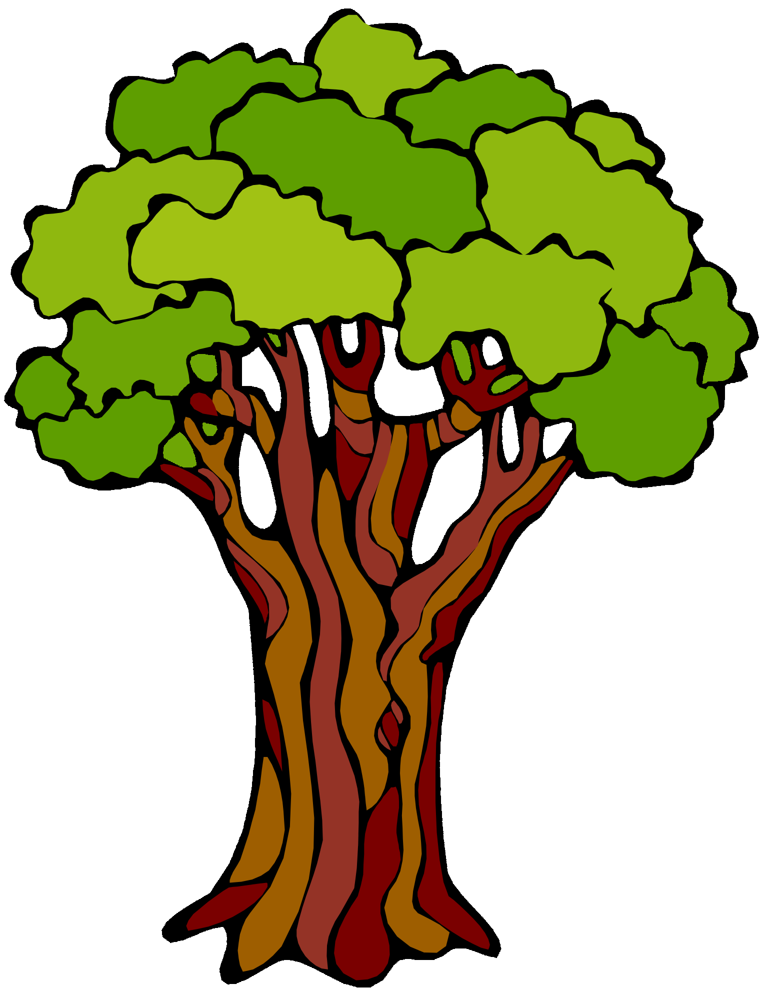
Keeping Our Reservation Tree Beautiful Despite:

- Flapping leaves
- Rushing sap
- Broken branches

Problem: Flapping Leaves

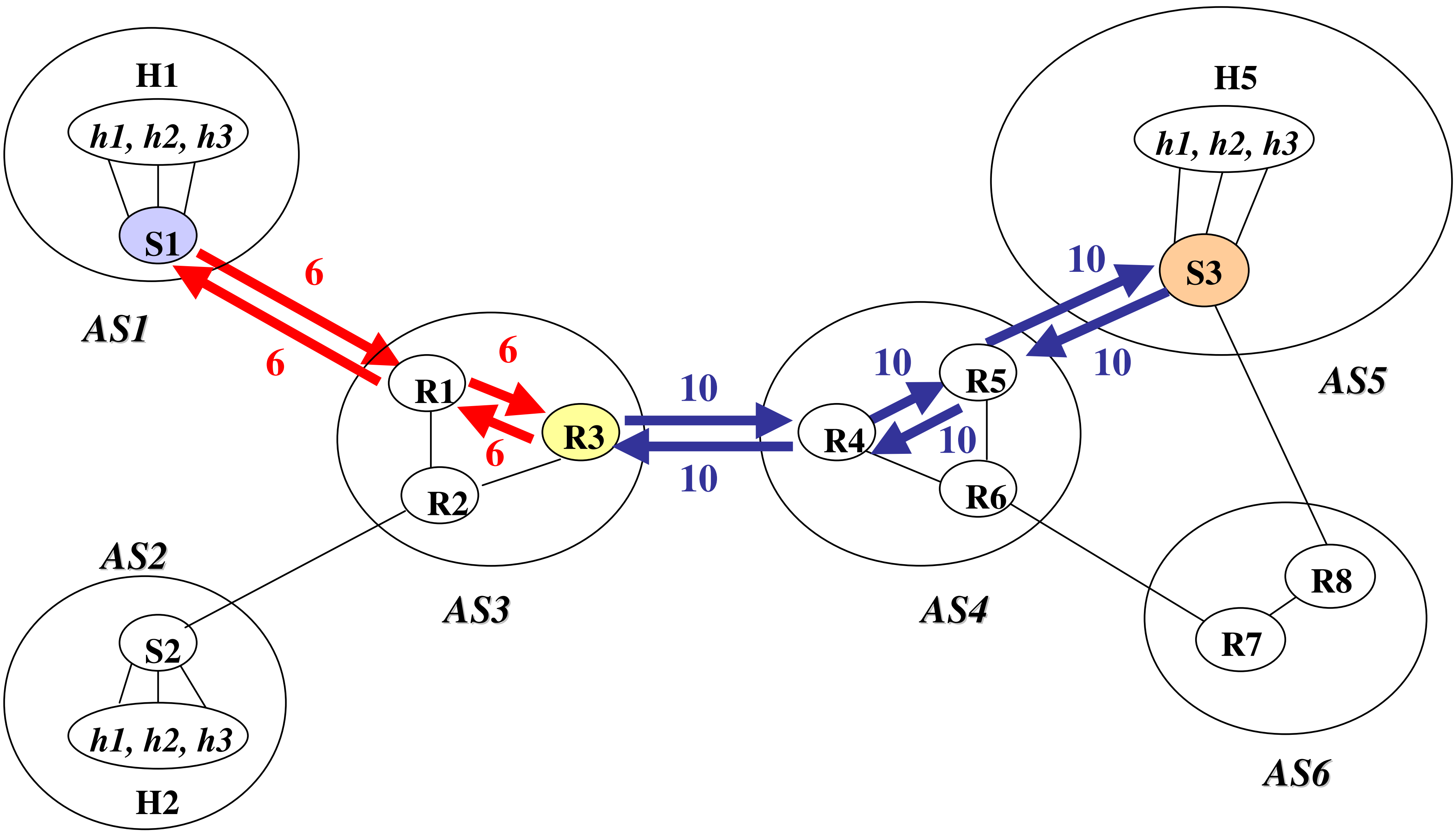
- Over-reservation
- Quantization
- Hysteresis

Problem: Rushing Sap

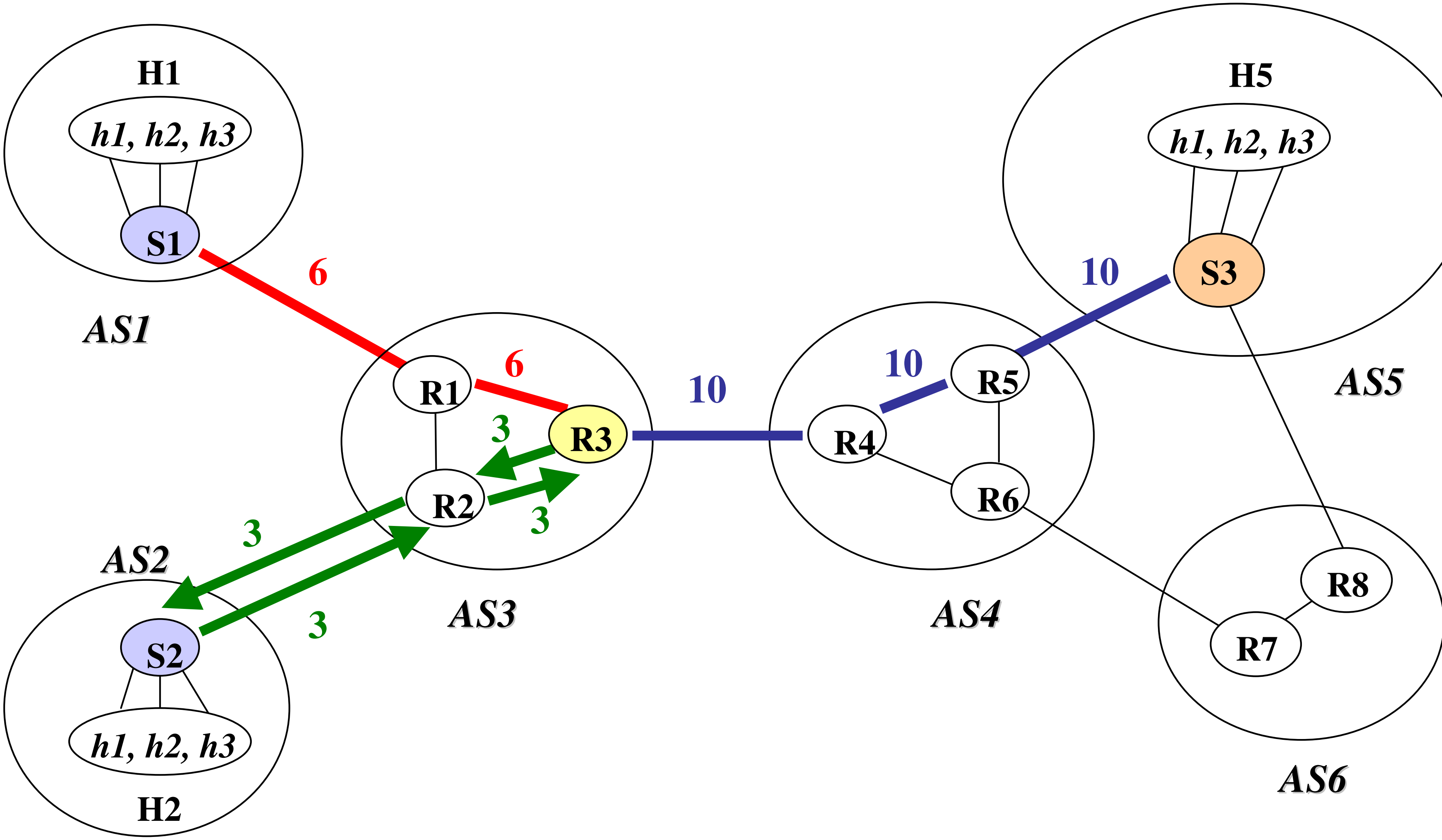


- CIDR Labeling
- Quiet Grafting

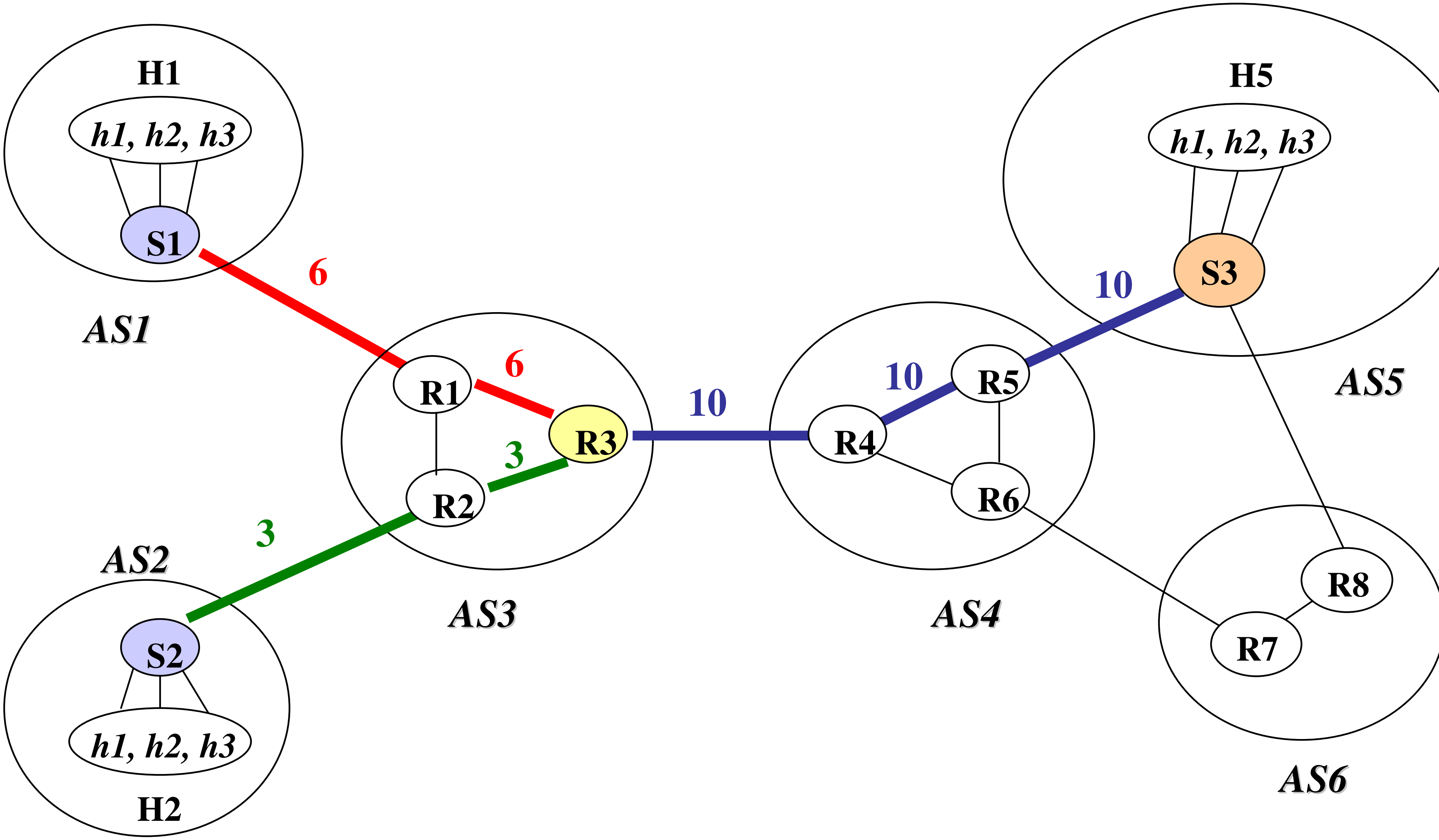
Quiet Grafting: 1st Branch



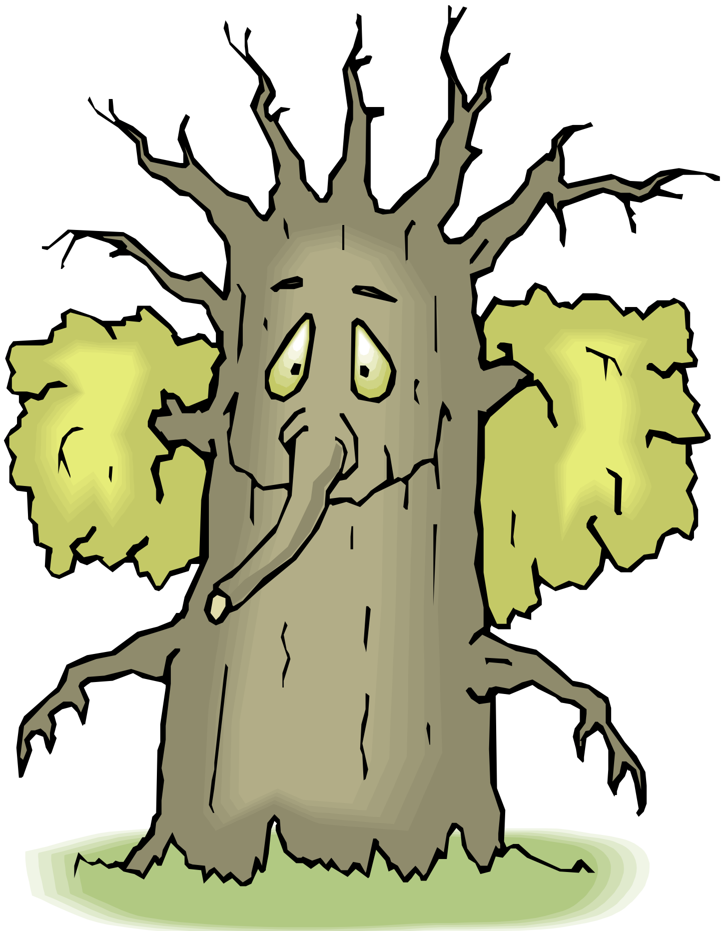
Quiet Grafting: 2nd Branch



Quiet Grafting: Complete



Problem: Broken Branches



- Self-Healing
- Filtering Route Changes

Performance Evaluation

Show BGRP benefits
as function of:

- Region Size
- Topology
- Traffic Load
- Refresh Rate
- Quantum Size

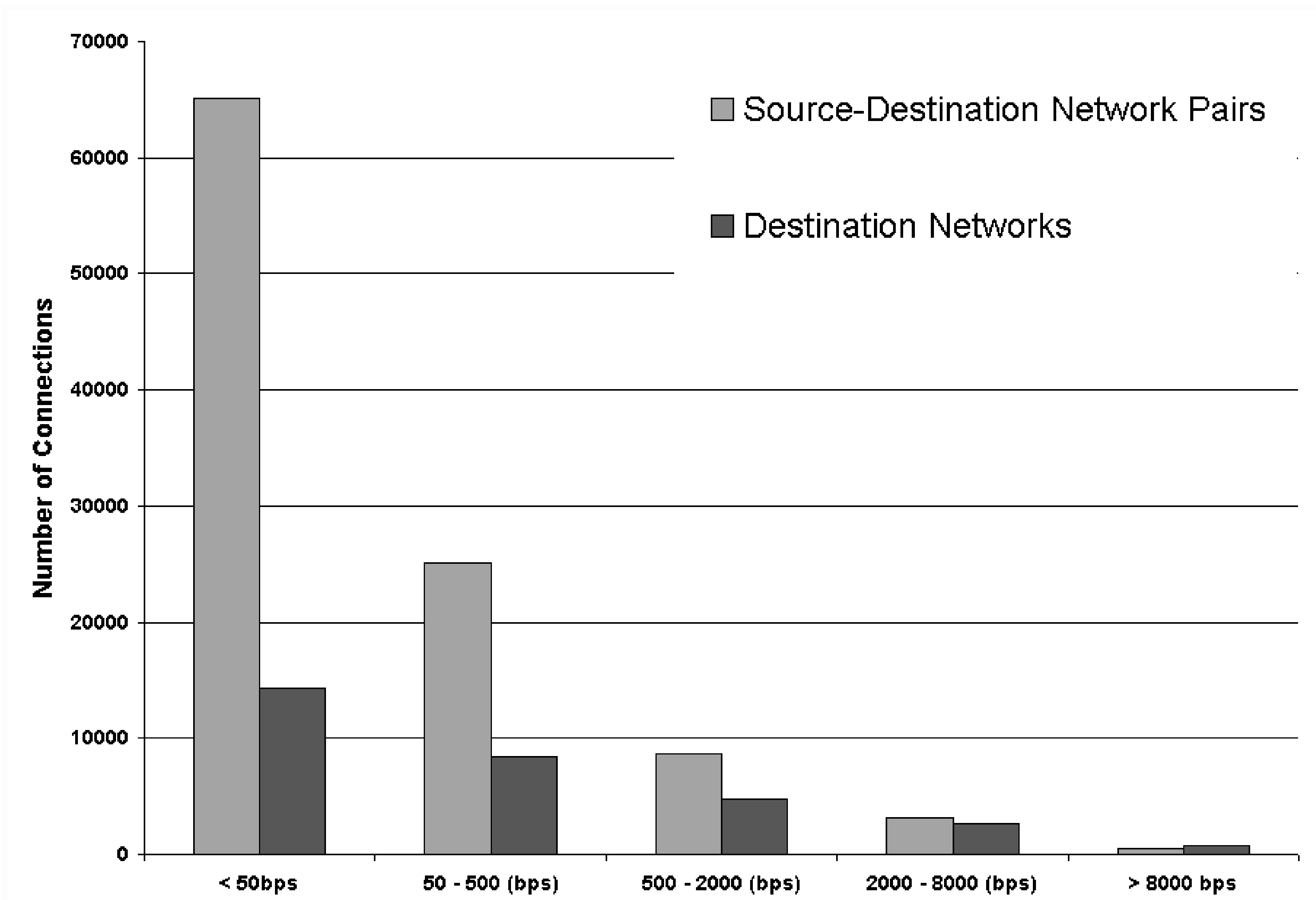
Flow Counts vs. Region Size

Time Interval	Region Granularity	# Source-Destination Pairs (For RSVP)	# Destinations (For BGRP)	Ratio
90 sec.	Application	339,245	208,559	1.6
	Host	131,009	40,538	3.2
	Network	79,786	20,887	3.8
	Domain	20,857	2,891	7.2
1 month	Domain	7,900,362	5,001	1,579.8

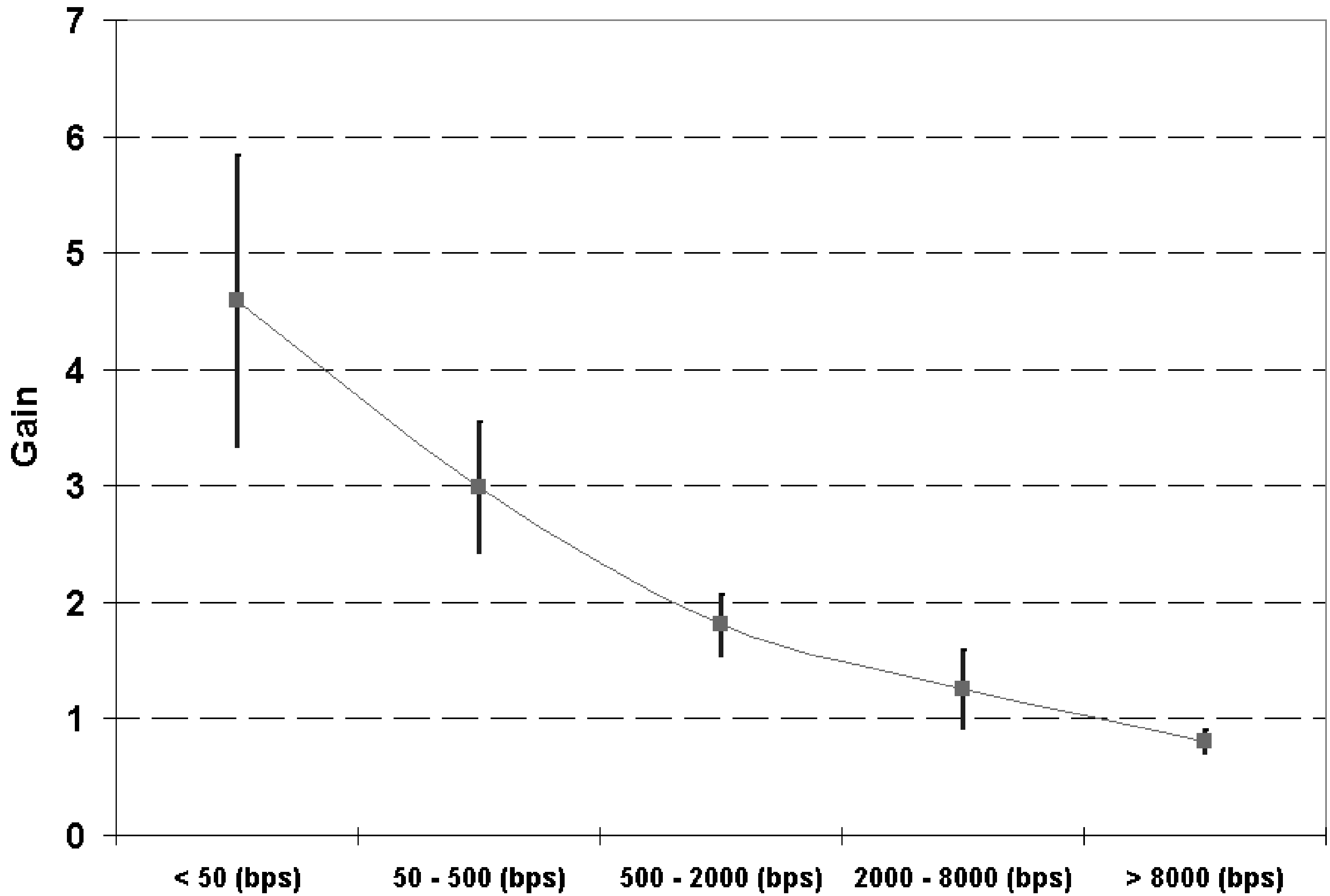
Flow Counts vs. Region Size

- Assume reserv'n is popular.
- Aggregation is needed !
- Region-based aggregation works.
- BGRP helps most when:
  - Aggregating Region is Large.
  - Reserv'n Holding Time is Long.
- Theoretical "N vs. N^2" problem is real !

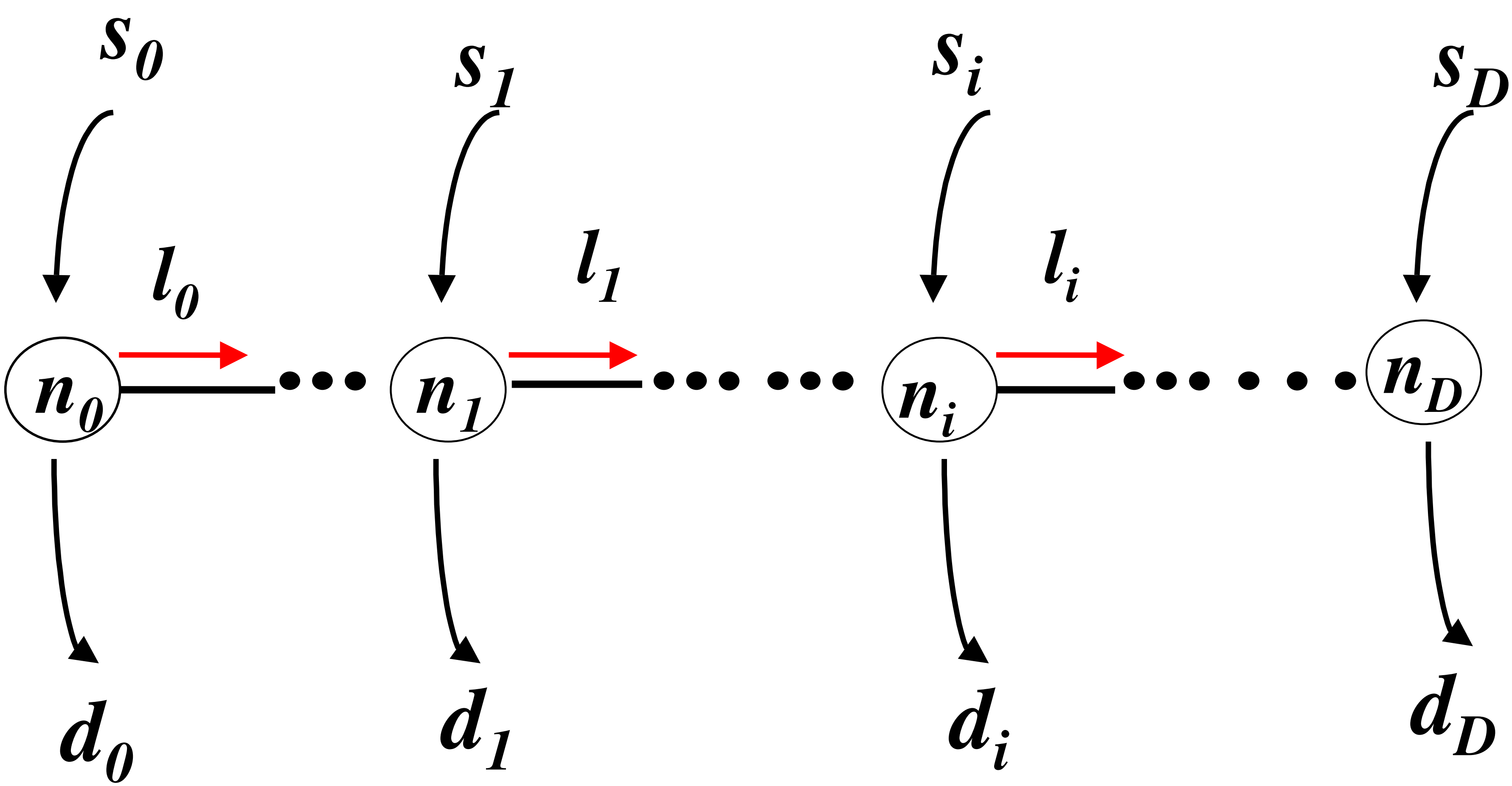
Number of Flows (broken down by BW)



BGRP / RSVP Gain for each BW Class

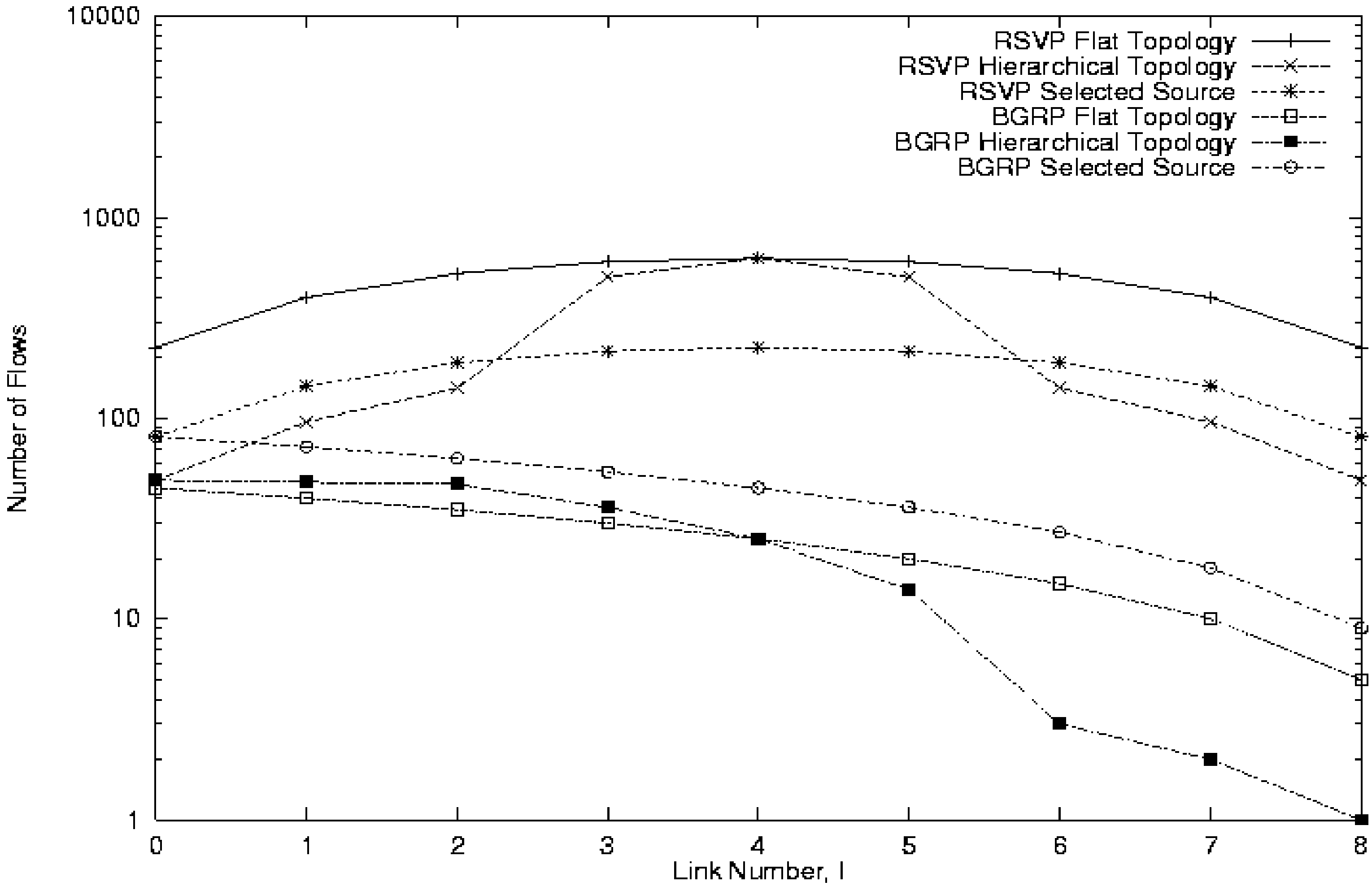


Modeling the Topological Distribution of Demand

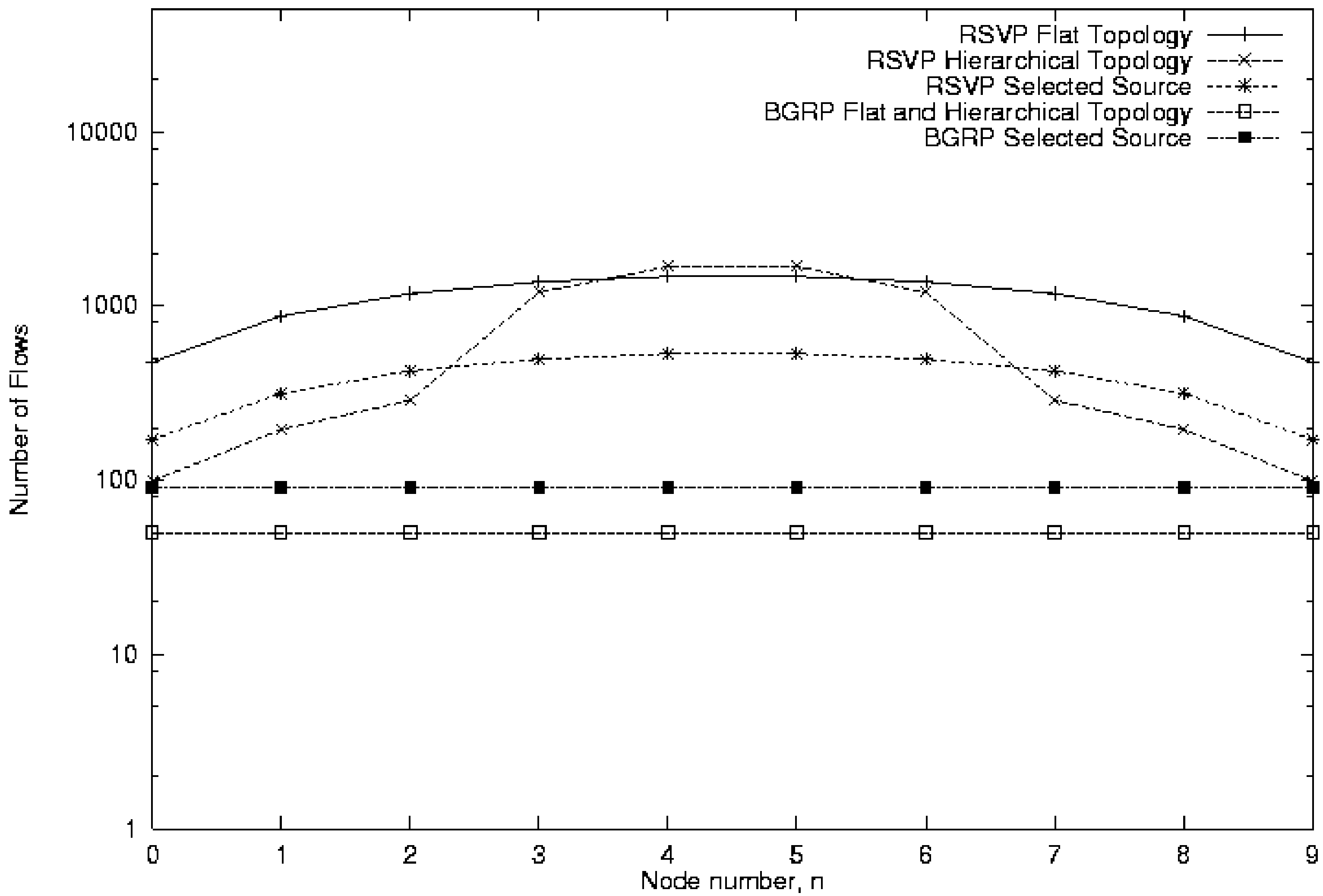


3 distributions: **Flat**, **Hierarchical**, **Selected Source**

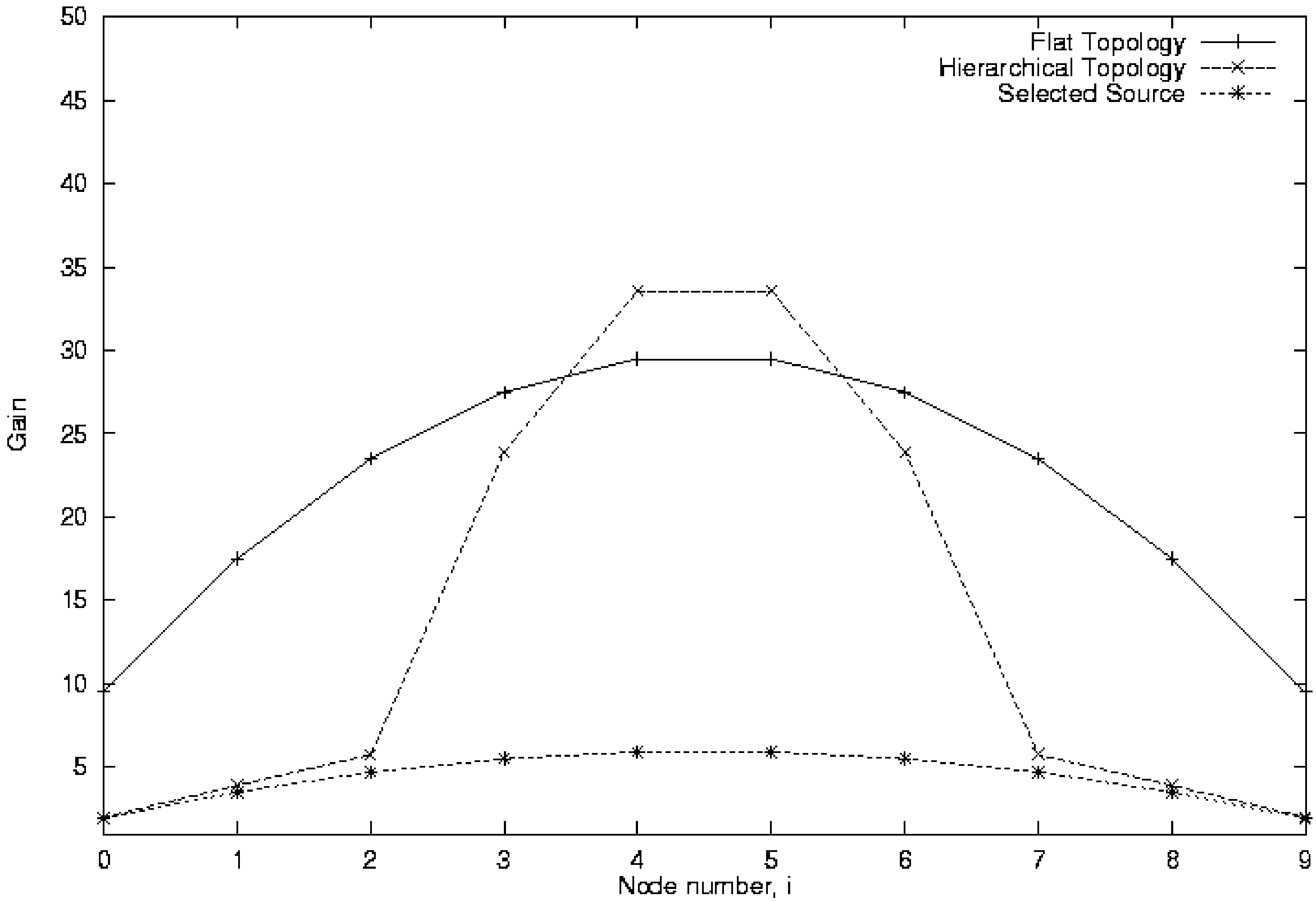
Reservation Count vs. Link Number



Reservation Count vs. Node Number



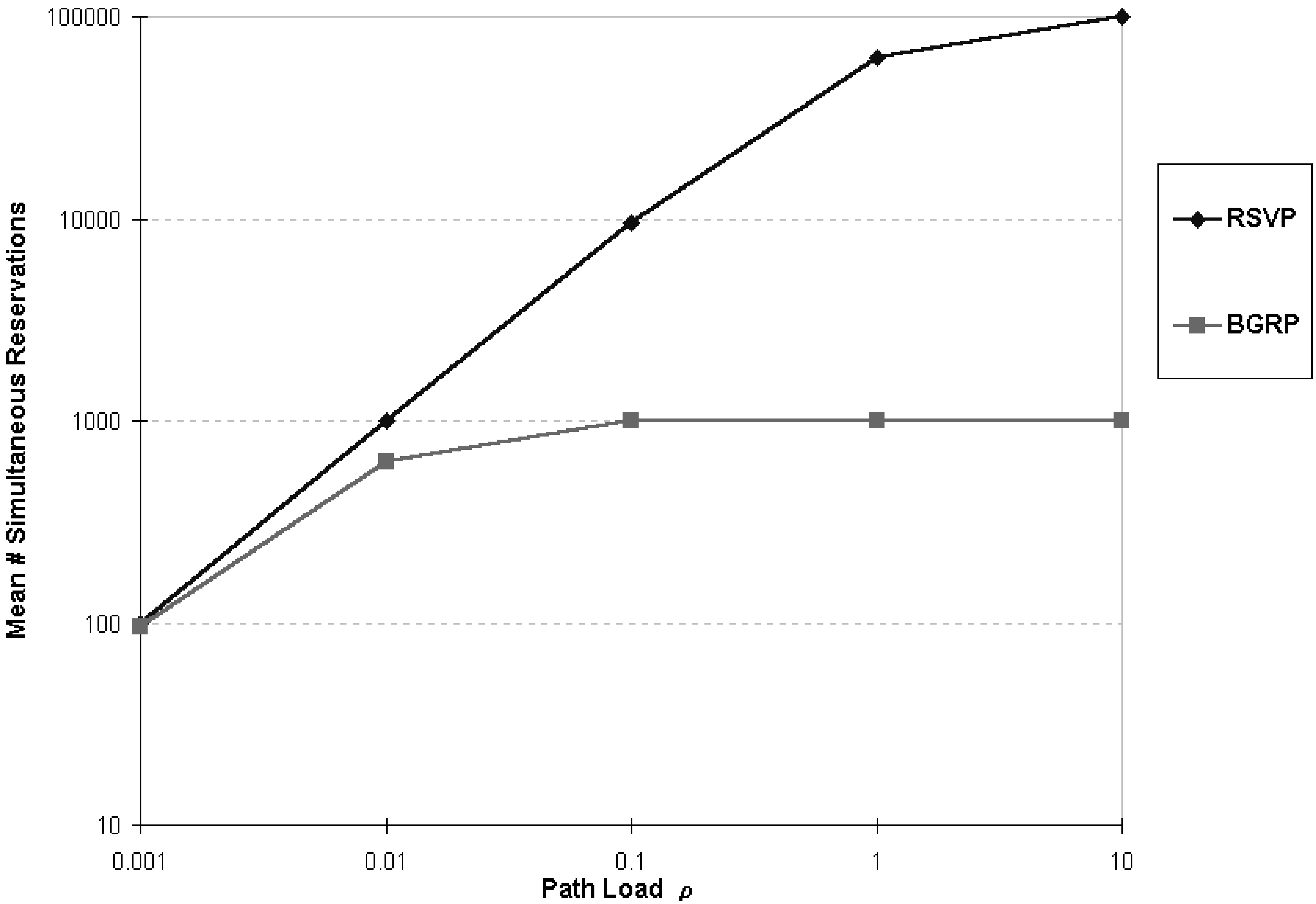
Gain: N^{rsvp} / N^{bgrp}



Reservation Count vs. Traffic Load

- Model for given hop H:
  - P paths thru H
  - T sink trees thru H
  - ρ micro-flows @ path (Poisson λ, μ, ρ)
- # RSVP reserv'ns = $(1-e^{-\rho})\cdot P$
- # BGRP reserv'ns = $(1-e^{-\rho\cdot F/T})\cdot T$
- BGRP helps most for large ρ
  - Gain ~ $P\,/\,T$
- Graph: P =100000, T =1000

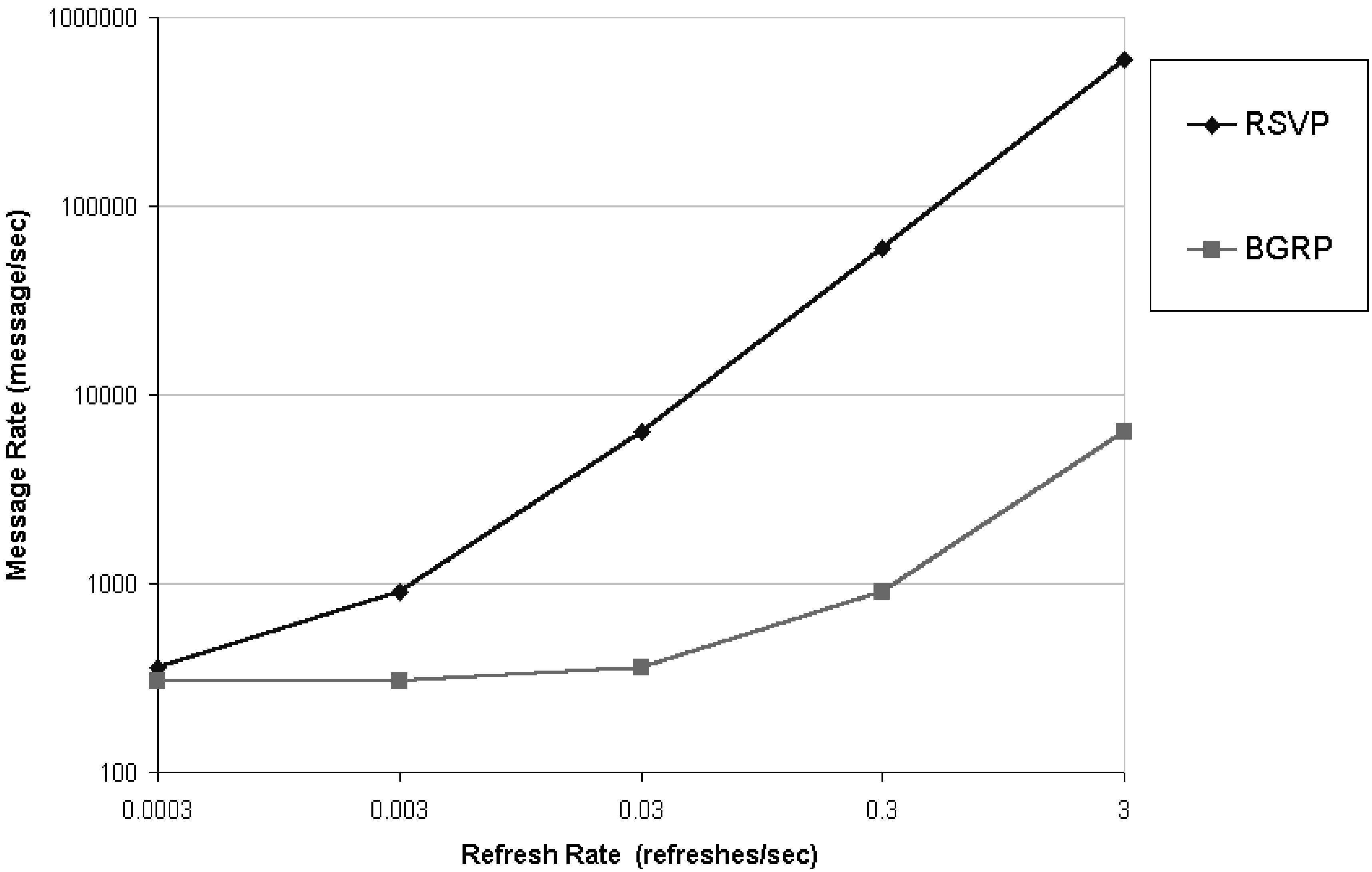
Reservation Count vs. Traffic Load



Message Rate vs. Refresh Rate

- Model for given hop H:
  - P paths thru H
  - T sink trees thru H
  - ρ micro-flows @ path (Poisson λ, μ, ρ)
  - η refresh rate
- RSVP msg rate = $3\lambda \cdot P + 2\eta \cdot P \cdot (1 - e^{-\rho})$
- BGRP msg rate = $3\lambda \cdot P + 2\eta \cdot T \cdot (1 - e^{-\rho \cdot P/T})$
- BGRP helps most for $\eta >> \lambda$, $\rho >> 1$
  - Gain ~ P / T
- Graph: P =100K, T =1000, ρ =10, λ =.001

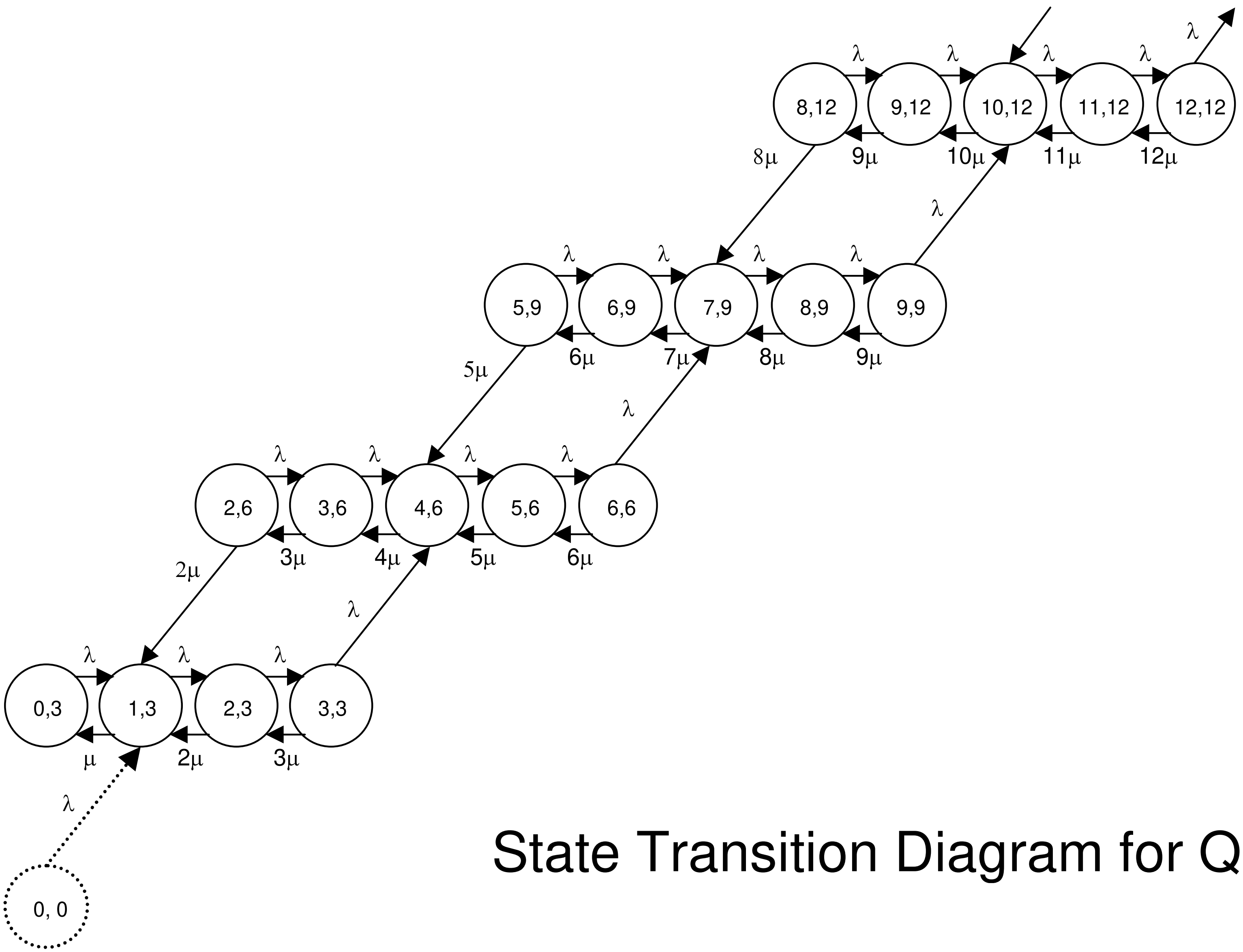
Message Rate vs. Refresh Rate



Message Reduction vs. Quantum Size

- Single hop H (tree leaf)
- ρ micro-flows on H (birth/death, Poisson)
- Each micro-flow needs 1 unit of BW
- H manages aggregate BW reserv’n
- Quantization: Reserv’n must be $k \cdot Q$
- Hysteresis: Descent lags by Q

Quantization with Hysteresis



State Transition Diagram for Q=3

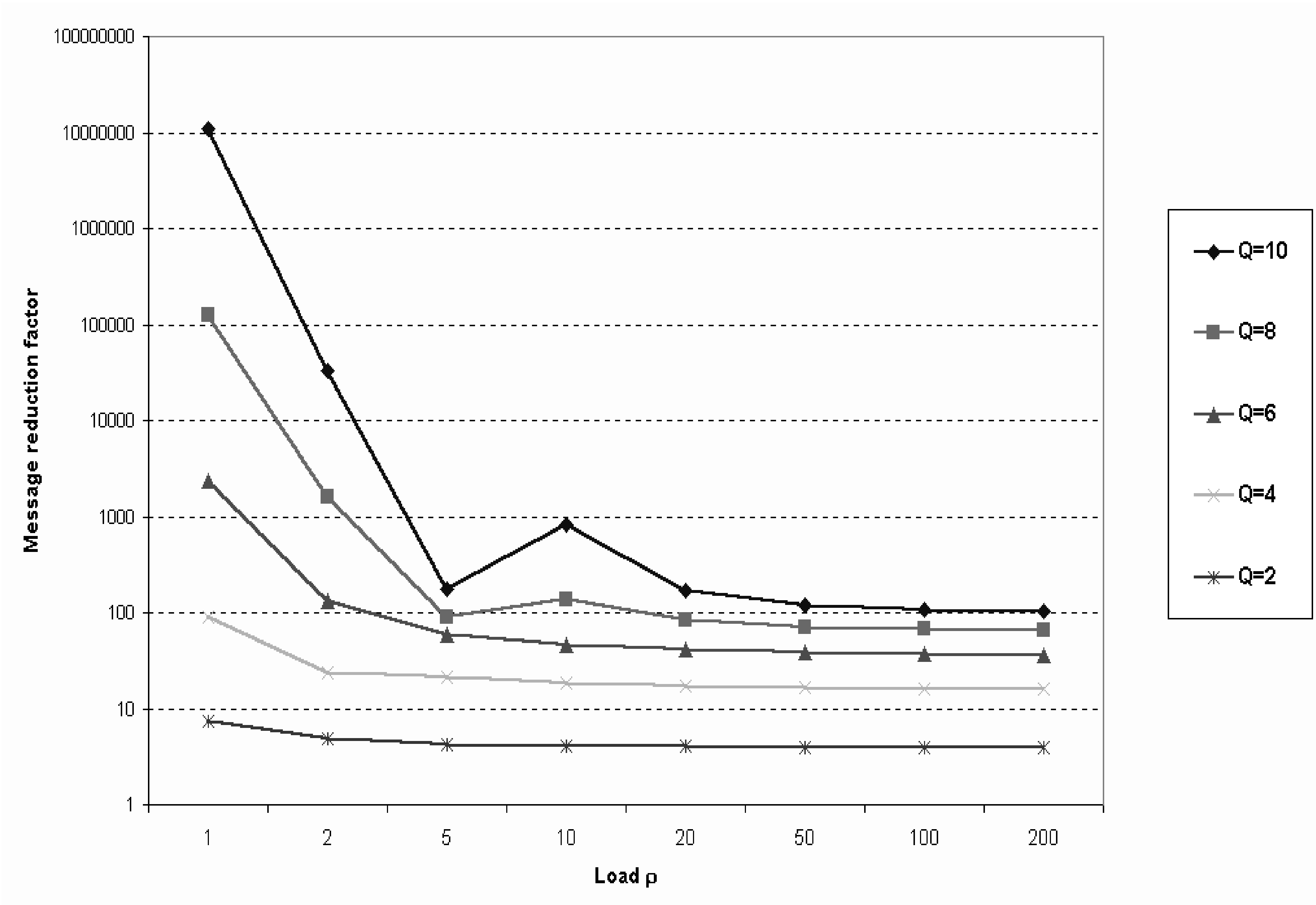
Message Reduction vs. Quantum Size

- Closed-form expression for state probabilities
- Quantization & Hysteresis cut message rate by:

$$e^{-\rho} \cdot \sum_{k=1}^{\infty} \left(\rho^{(k \cdot Q)} / \sum_{i=0}^{Q-1} [\rho^{i} \cdot (k \cdot Q - i)!] \right)$$

- E.g., ρ=*100* & *Q=10,* message rate cut by 100
- Multi-hop model with Quiet Grafting:
  - Further improvement
  - Approximate analysis
  - Simulation

Message Reduction vs. Load & Quantum Size



Conclusions

BGRP meets Challenges

- Scalable Protocol State
- Scalable Protocol Processing
- Scalable Protocol Bandwidth
- Scalable Data Forwarding
- Inter-Domain Administration

Future Work

- Detailed Protocol Specification
- Simulation
- Reference Implementation
- MPLS
- Lucent products
- Internet 2 (Q-bone)
- IETF: Draft; BOF; Working Group

Future Work: Bandwidth Broker Model