Service Discrimination Based on Delay Priority in Integrated Services Internets

Lee Breslau
Xerox PARC
November 6, 1996

XEROX

Current Internet Service

Single class of best-effort service

- All packets are generally treated the same

New applications (audio, video) require revisiting this key architectural feature

Service Discrimination

Can differentiate service along several axes:

Reserved vs. Unreserved

Scheduling Priority

Dropping Priority

Preemptable service

XEROX

Resource Reservations

Consensus that resource reservations are needed to support real-time applications in the Internet

- assurance of minimum level of service for acceptable performance

IETF moving to standardize relevant protocols and services

- Real-time services
  - Controlled Load
  - Guaranteed
- RSVP

Our Focus

Should Internet include service discrimination based on delay priority? What benefits does it provide?

Within the context of real-time services, is a single level of service sufficient or do multiple levels of service differentiated by delay provide value?

Important because set of available services determines what applications can be supported and how efficiently

Fundamental question about amount of functionality in the network and in the applications

Methodology to understand this question

Costs of Multiple Service Levels

General architectural complexity

Impact on reservation setup protocol in the context of multicast applications

XEROX

Outline

I. Motivation

II. Delay Adaptive Applications

III. Performance Issues

IV. Simulation Experiments/Results

Delay Adaptive Applications

In presence of variable network delays, applications buffer packets for later playback

- for how long?

For real-time applications in the Internet, common mode of operation is to adapt the playback point to changing network delays

- applications can work over different qualities of service
- easier for the network (no delay bounds)
- lower average delays (even with delay bounds)

XEROX

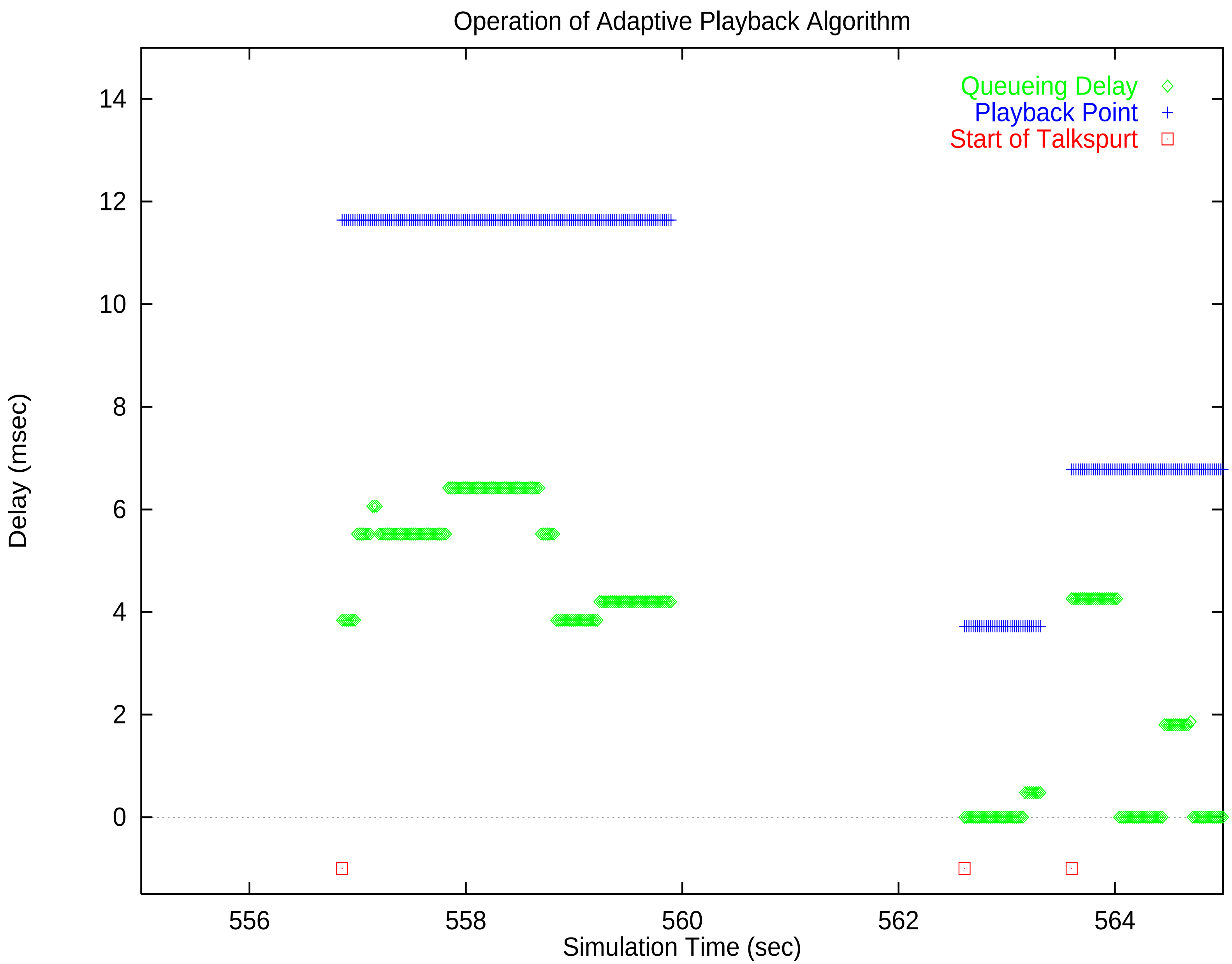
Delay Adaptive Audio Applications

For audio applications, set a playback time on receipt of first packet in a talkspurt

Buffer all packets in the talkspurt so that total delays (queueing + playout) are equal

What to do about late arriving packets?

- drop or adjust playback point



Operation of Adaptive Playback Algorithm
Queueing Delay
Playback Point
Start of Talkspurt
Delay (msec)
14
12
10
8
6
4
2
0
556
558
560
562
564
Simulation Time (sec)

XEROX

Performance of Delay Adaptive Applications

Total Delay -- network+playback

Distortion -- caused by adjustments in playback point over time

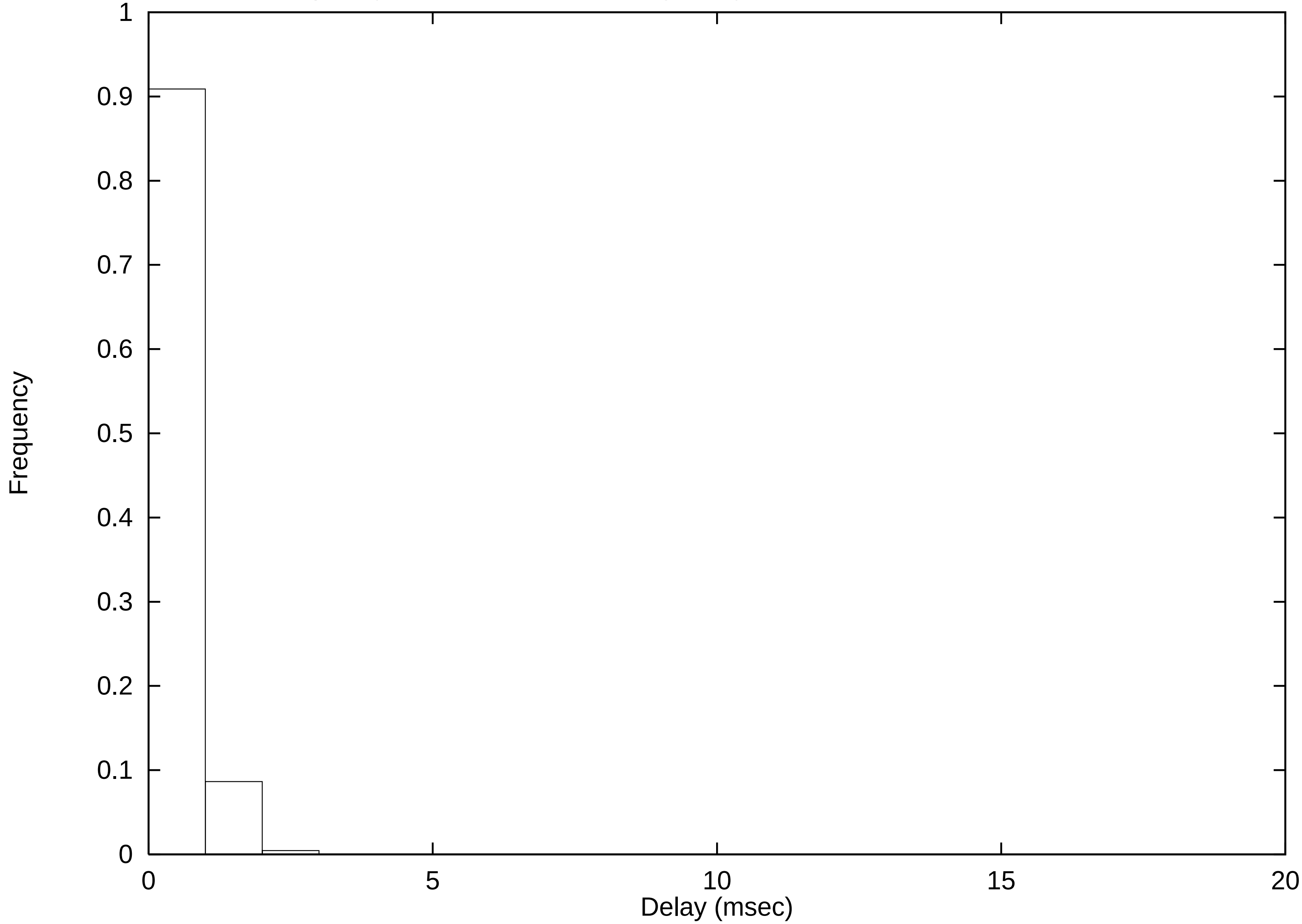
Application performance is a function of the service provided by the network and the adaptation algorithm used by the application

XEROX

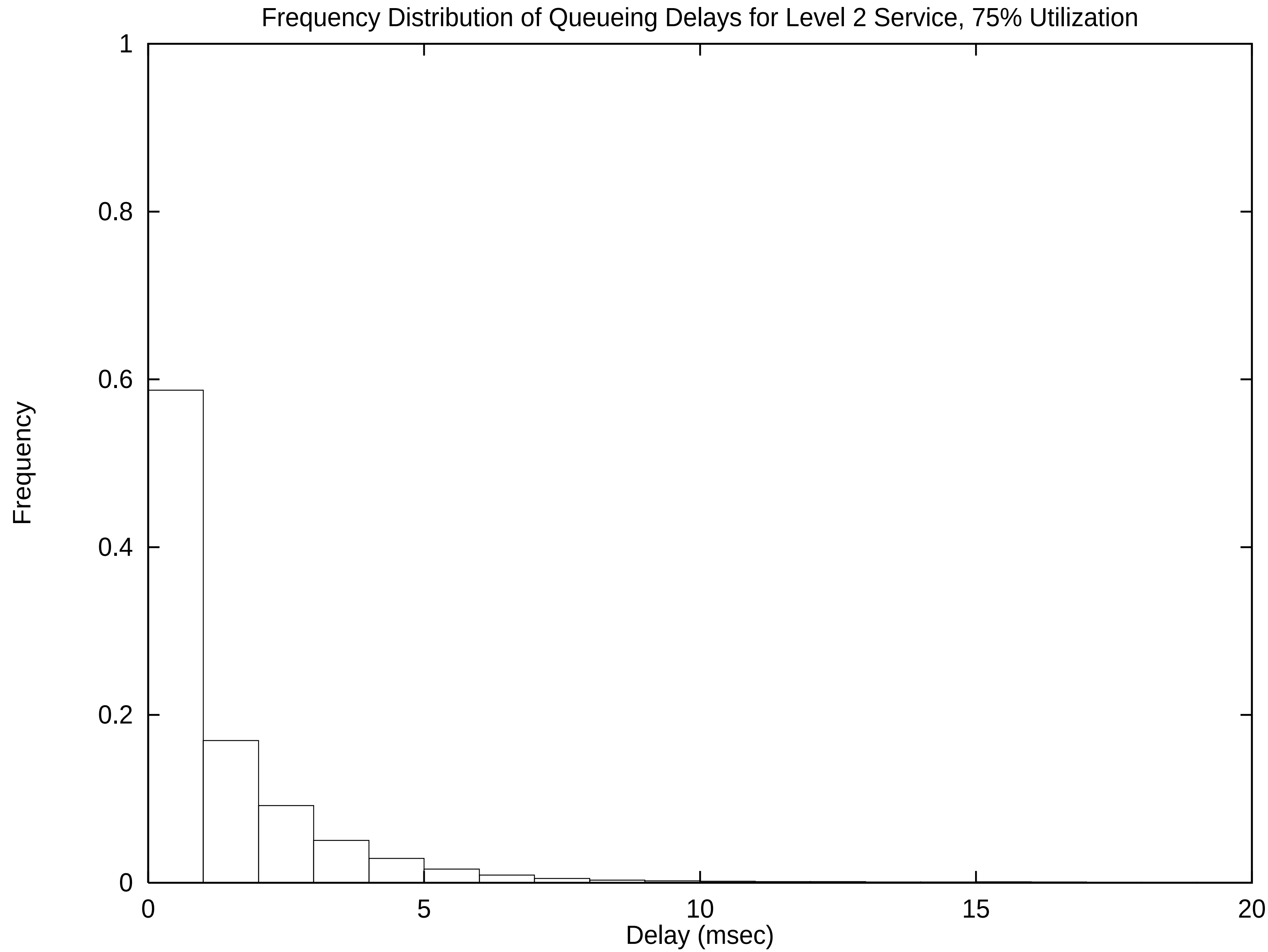
How Many Levels of Service?

Given real-time applications with different delay requirements (e.g., interactive vs. non-interactive), answer should be obvious

Frequency Distribution of Queueing Delays for Level 1 Service, 75% Utilization



XEROX


Frequency Distribution of Queueing Delays for Level 2 Service, 75% Utilization
Frequency
1
0.8
0.6
0.4
0.2
0
0
5
10
15
20
Delay (msec)

XEROX

How Many Levels of Service?

What impact does the tail of the delay distributions have on adaptation algorithms and resulting application performance?

Simulation Methodology

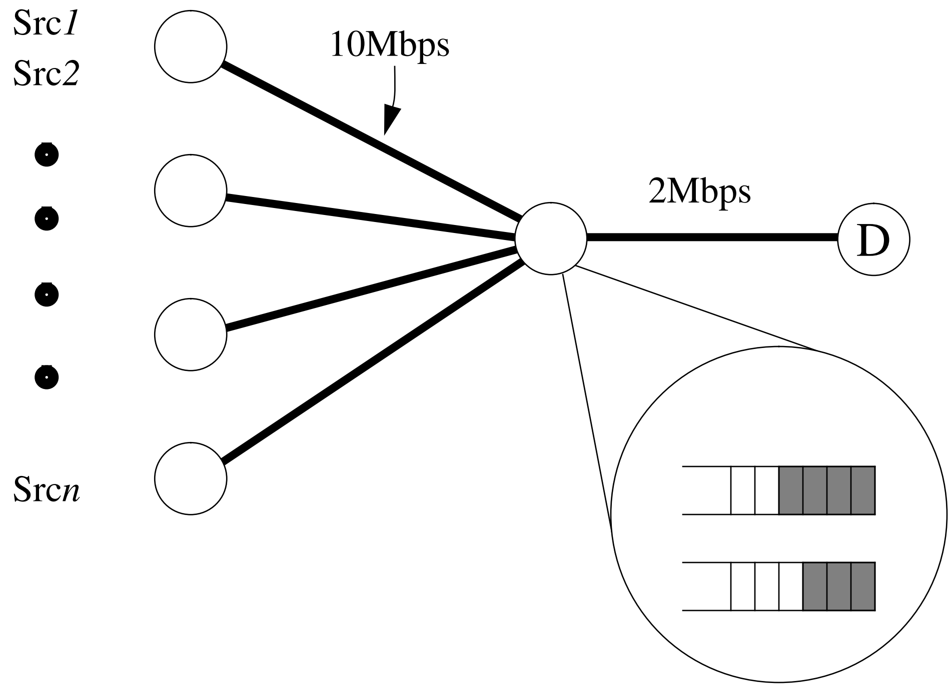
Compare a single level of service with 2 levels of service

Assume different kinds of applications using application appropriate adaptation algorithms

Measure delay and distortion

Compute network utility based on assumptions about application-specific utility

Simulation Toplogy



2 Adaptation Algorithms

1. Aggressive algorithm like that employed in the *vat* application
   - at start of talkspurt, playout delay is a function of computed delay variance

2. Conservative algorithm
   - start with a large playback delay and adjust delay upward (as needed) but not downward

Algorithms not claimed to be optimal; chosen to represent two ends of the spectrum

Taxonomy of Real-Time Applications

		Delay	
		sensitive	insensitive
Distortion	sensitive		
	insensitive		

Performance Measures

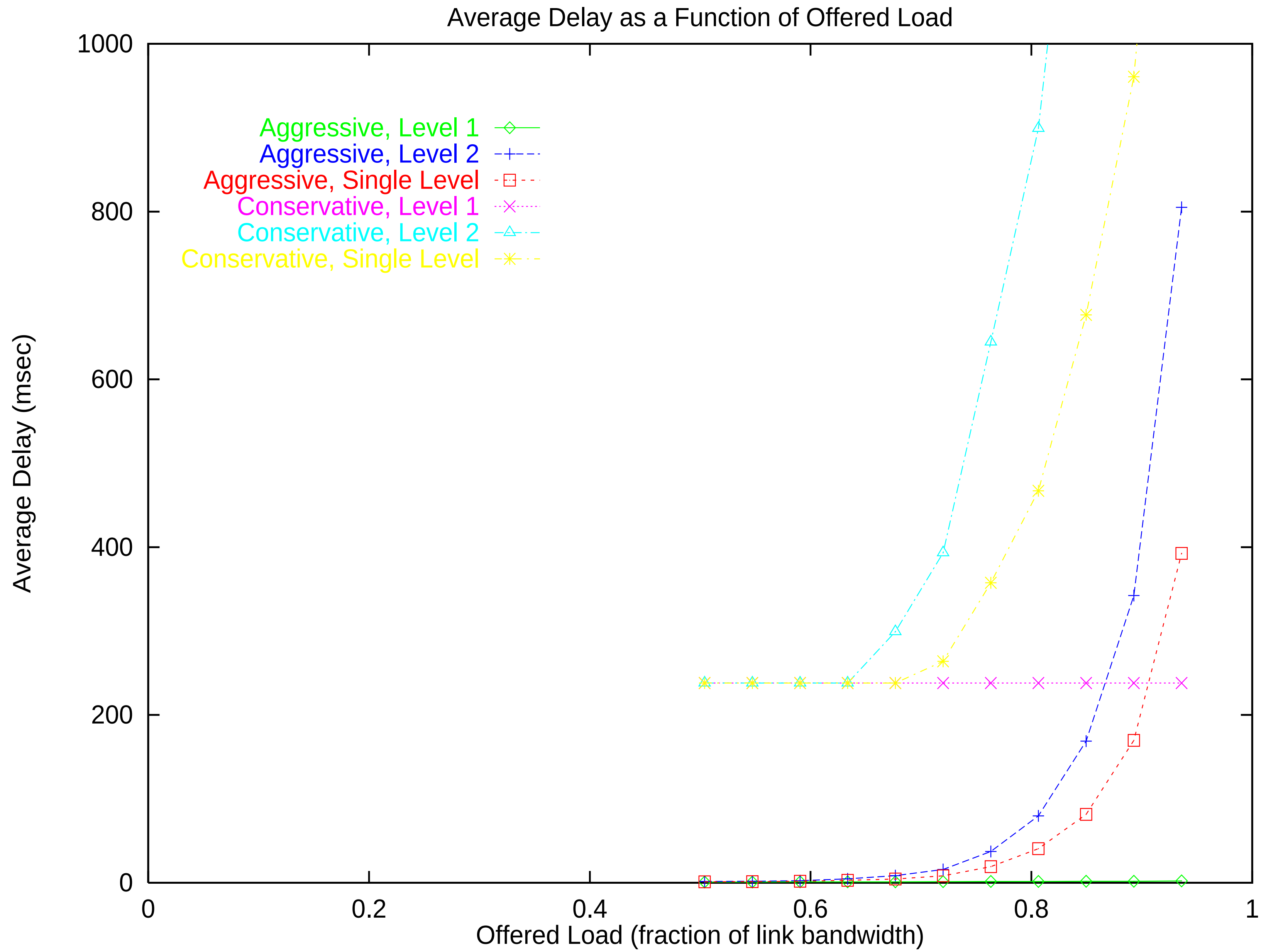
Delay:

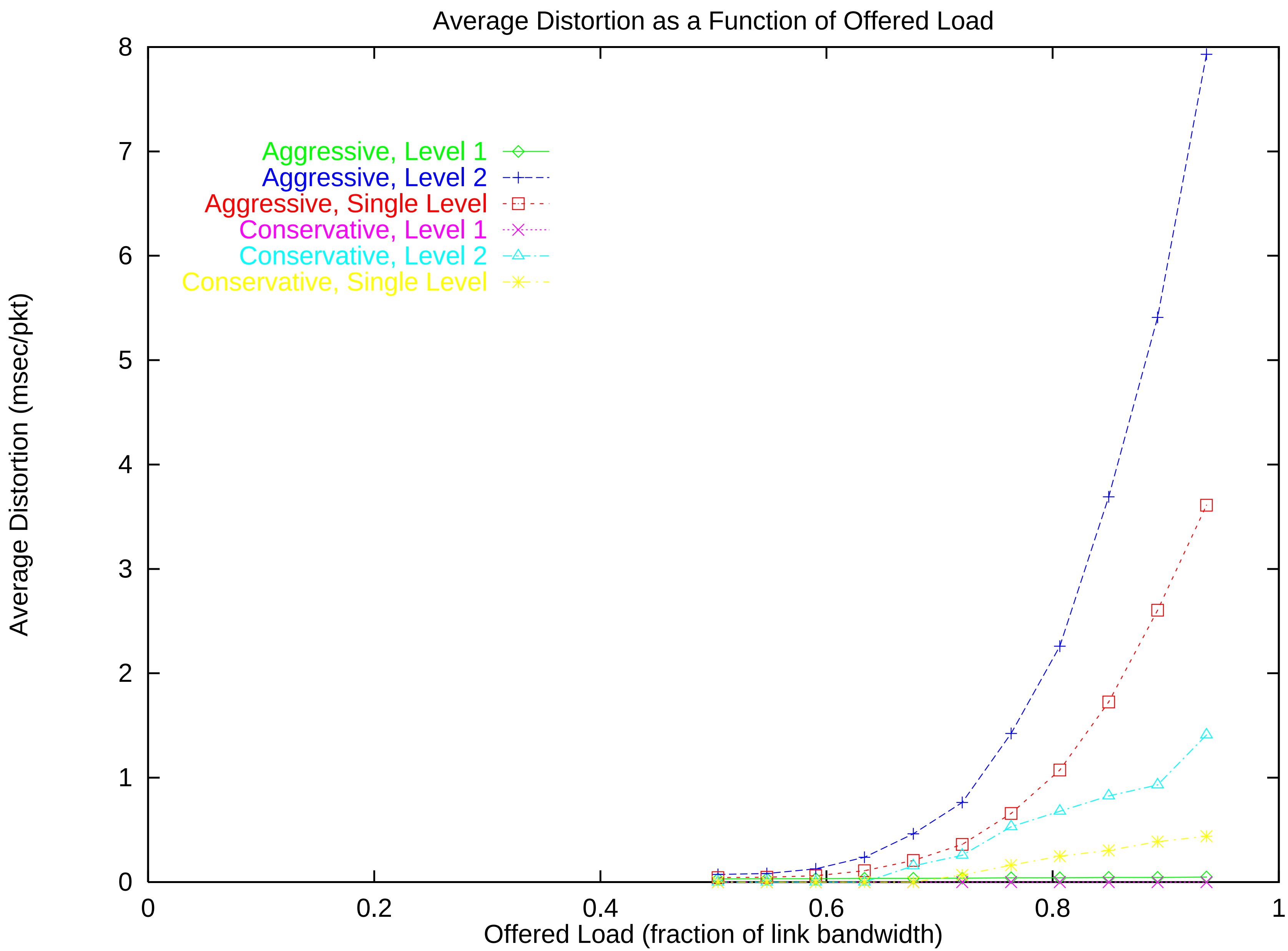
$$Del = \frac{\sum d_i}{n}$$

Distortion:

$$e_i = \left| \frac{d_i - d_{i-1}}{t_i - t_{i-1}} \right|$$

$$Dist = \sqrt{\frac{\sum_i e_i^2}{n}}$$


Average Delay as a Function of Offered Load
Aggressive, Level 1
Aggressive, Level 2
Aggressive, Single Level
Conservative, Level 1
Conservative, Level 2
Conservative, Single Level
Average Delay (msec)
0
200
400
600
800
1000
0
0.2
0.4
0.6
0.8
1
Offered Load (fraction of link bandwidth)


Average Distortion as a Function of Offered Load
Average Distortion (msec/pkt)
Offered Load (fraction of link bandwidth)
Aggressive, Level 1
Aggressive, Level 2
Aggressive, Single Level
Conservative, Level 1
Conservative, Level 2
Conservative, Single Level
0
1
2
3
4
5
6
7
8
0
0.2
0.4
0.6
0.8
1

Application Utility

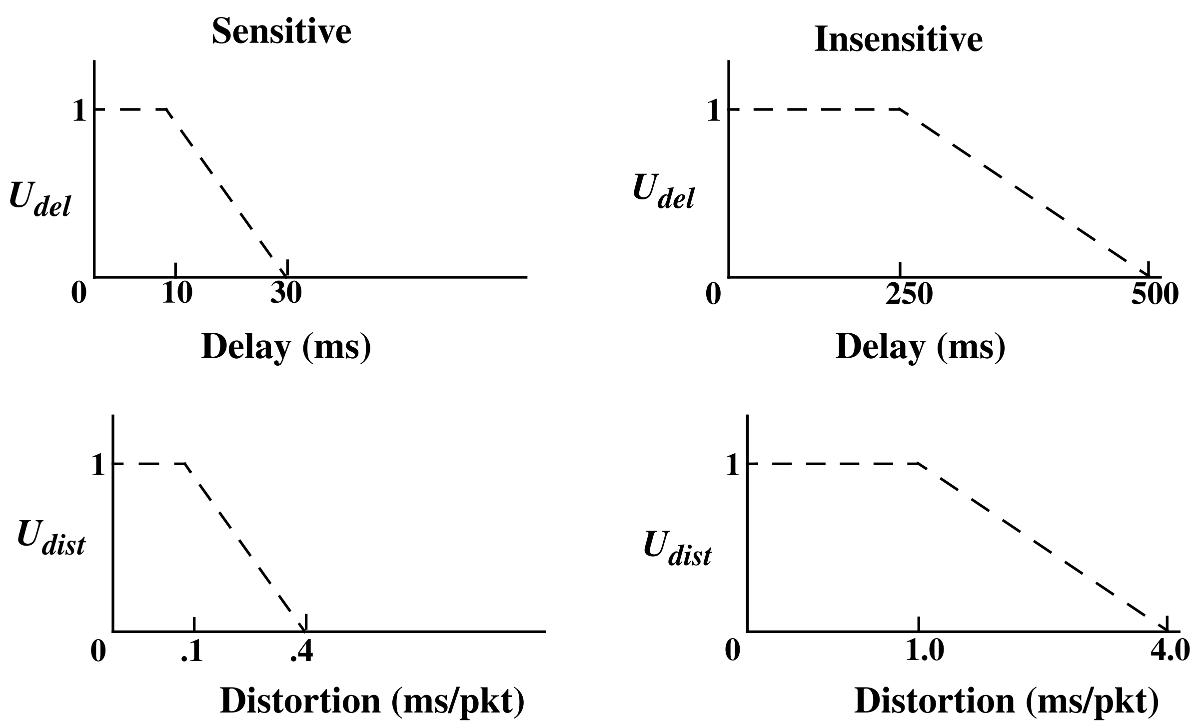
How do differences in service levels affect application utility?

Simple models of application utility as a function of delay and distortion

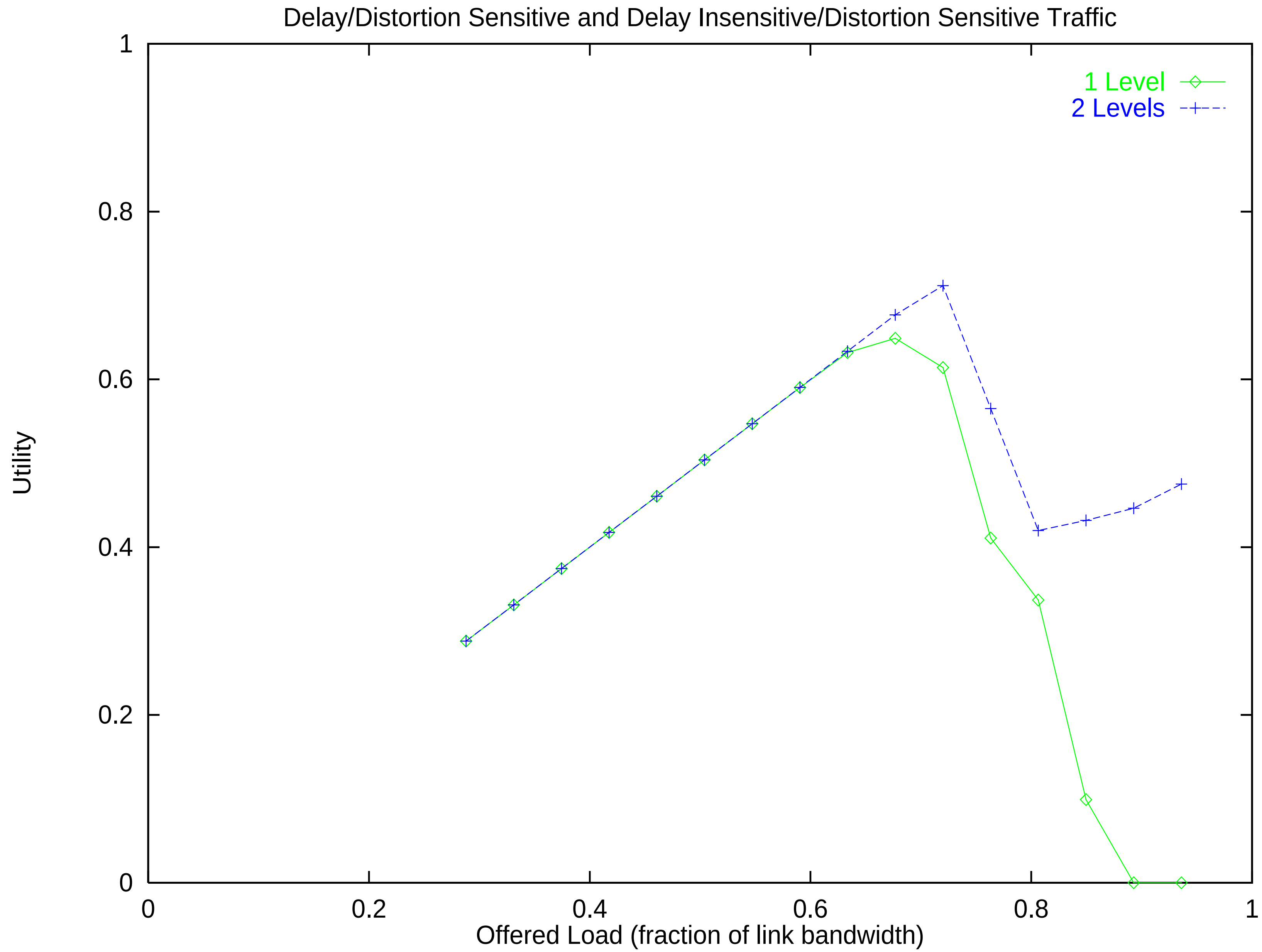
- $U_{del} = f(Del)$
- $U_{dist} = g(Dist)$
- $U_{tot} = U_{del} \times U_{dist}$

XEROX

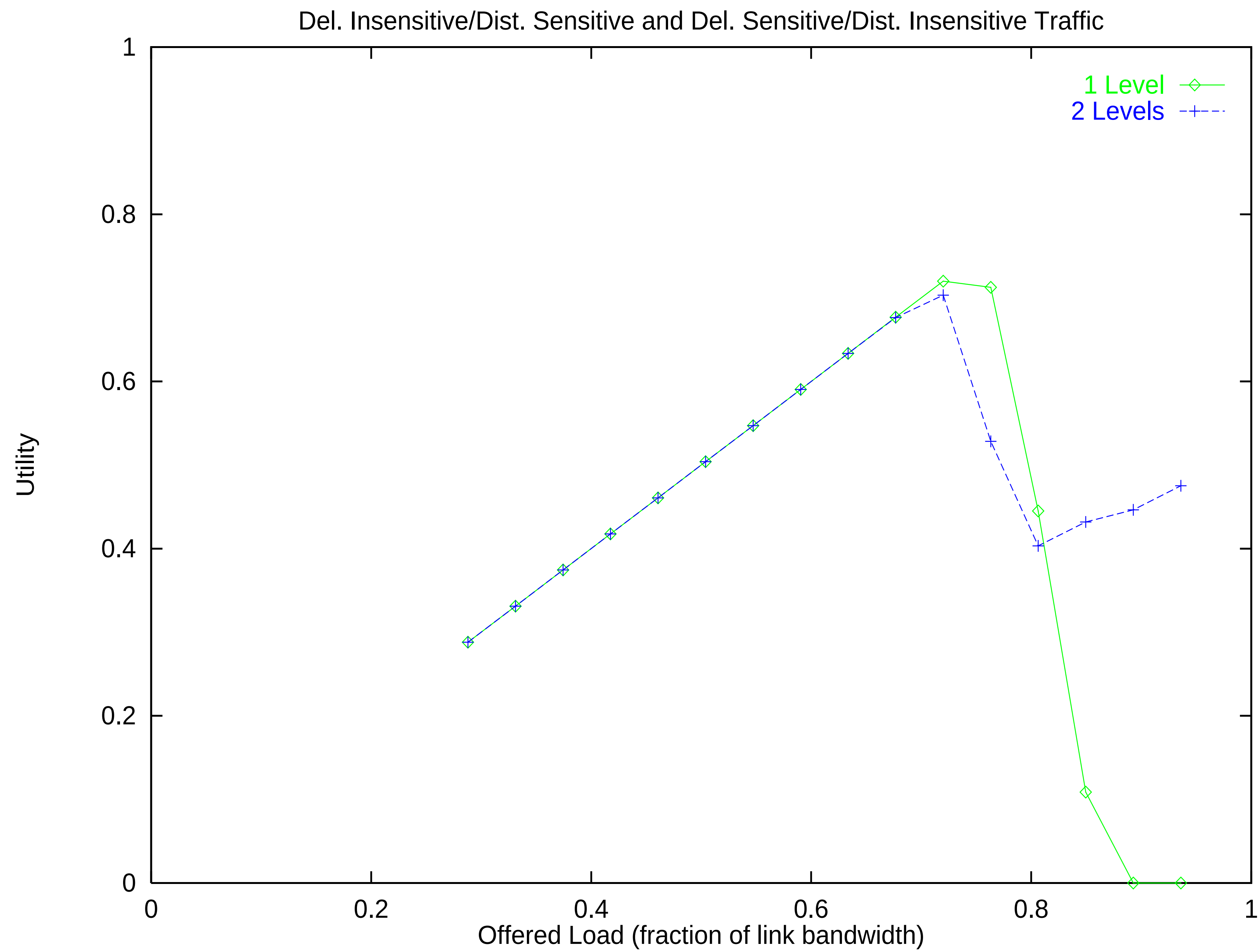
Utility Functions



$$U_{tot} = U_{del} \times U_{dist}$$



Delay/Distortion Sensitive and Delay Insensitive/Distortion Sensitive Traffic
1 Level
2 Levels
Utility
Offered Load (fraction of link bandwidth)
0
0.2
0.4
0.6
0.8
1



Del. Insensitive/Dist. Sensitive and Del. Sensitive/Dist. Insensitive Traffic
1 Level
2 Levels
Utility
Offered Load (fraction of link bandwidth)
0
0.2
0.4
0.6
0.8
1

Discussion

Value of multiple levels of service depends on:

- the nature of real-time applications and their utility
- amount of non-real-time traffic in the network

Limits to what adaptive applications can do

- hard to provide both low delay and low distortion
- if network is highly loaded, need mechanism inside the network to maximize utility

Results consistent with arguments for admission control

Ongoing/Future Work

Different source models

- burstier traffic

Service discrimination based on dropping priority

Assess value of delay-bounded services