Improving Utilization for Deterministic Real-time Service

Hui Zhang
The Tenet Group
University of California at Berkeley
and International Computer Science Institute

and

Information and Computing Sciences Division
Lawrence Berkeley Laboratory

ILP Conference
March 10, 1994

Synopsis

- What is the problem?
  - performance guarantees, bursty traffic, high average utilization
- What is the solution?
  - tighter delay bounding technique: deterministic fluid model analysis

Problem Specification

- Support new applications in packet networks
  - scientific visualization, medical imaging, video conferencing
- Alternative solutions:
  - traditional best-effort service
  - guaranteed service
  - predicted service

Considerations

- Environment: wide-area public packet-switching networks

	best-effort	predicted	guaranteed
protection	no	?	yes
QOS	bad	moderate	good
utilization	high	?	?
complexity	low	moderate	moderate

- Guaranteed service: utilization is the key consideration

Guaranteed Services

- Deterministic service: for the k^{th} packet on connection j

$$D_j^k \leq \overline{D}_j \quad \forall\ k$$

- Statistical service:

$$Prob(D_j^k \leq \overline{D}_j) \geq z_j$$

Providing Guaranteed Services

Two levels of control:

- Connection admission control: RCAP
  - reserves resources
- Service discipline: RTIP
  - allocates resources according to the reservation

Original Tenet Admission Control Test

- Limit the sum of peak rates to be less than link speed

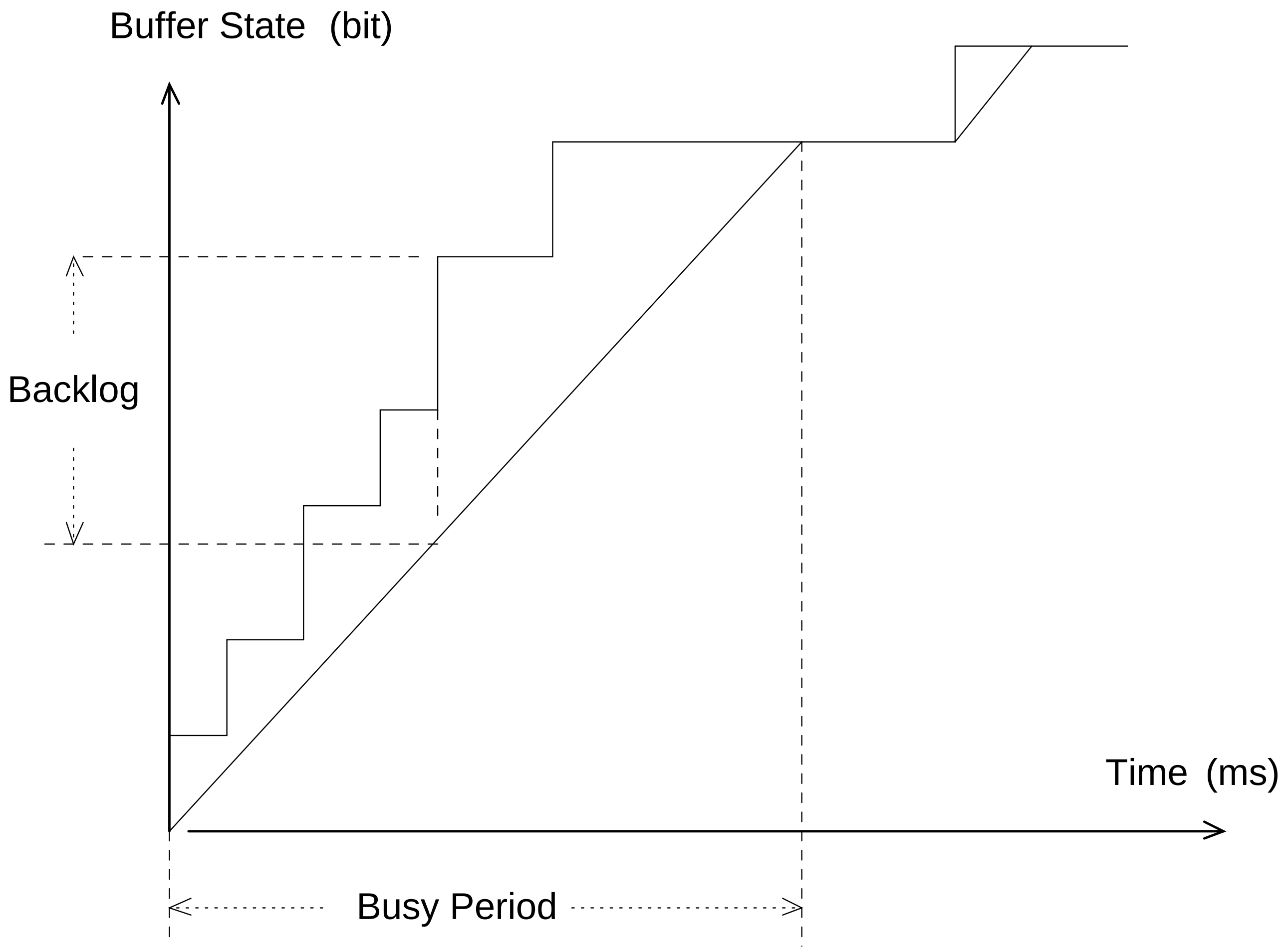
- Low average link utilization if peak-average ratio is high
  e.g. peak-average ratio 5 $\Longrightarrow$ link utilization $\leq$ 20%

- Can we bound delay when sum of peak rates is greater than link speed?

Bounding Techniques: Deterministic Fluid Model Analysis

- Pioneer work by R. Cruz, 1987
- More recent work:
  - Fixed Rate Servers, S. Low and P. Varaiya, 1991
  - General Processor Sharing, A. Parekh and R. Gallager, 1992
  - Hierarchical Round Robin, A. Banerjea and S. Keshav, 1992

Concepts: Delay, Backlog and Busy Period



Result

- First Come First Served and Static Priority solved

- Example: FCFS & homogeneous sources:

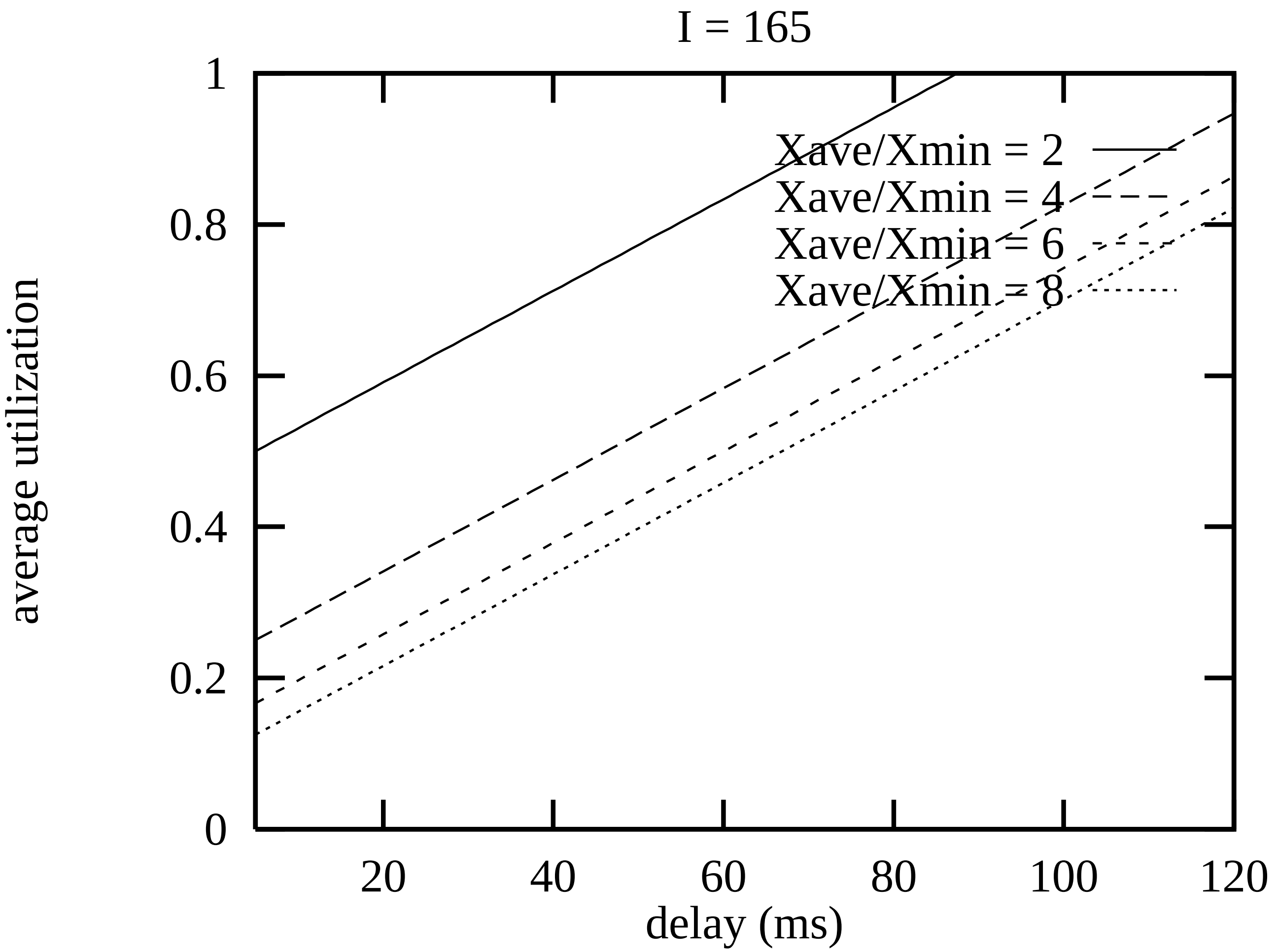
$$\overline{d} = Xmin + (\mu_{ave} - \frac{1}{burstRatio})I$$

  - $Xmin$: minimum packet inter-arrival time
  - $Xave$: average packet inter-arrival time over an interval I
  - $burstRatio$: $\frac{Xave}{Xmin}$
  - μ_{ave}: average link utilization

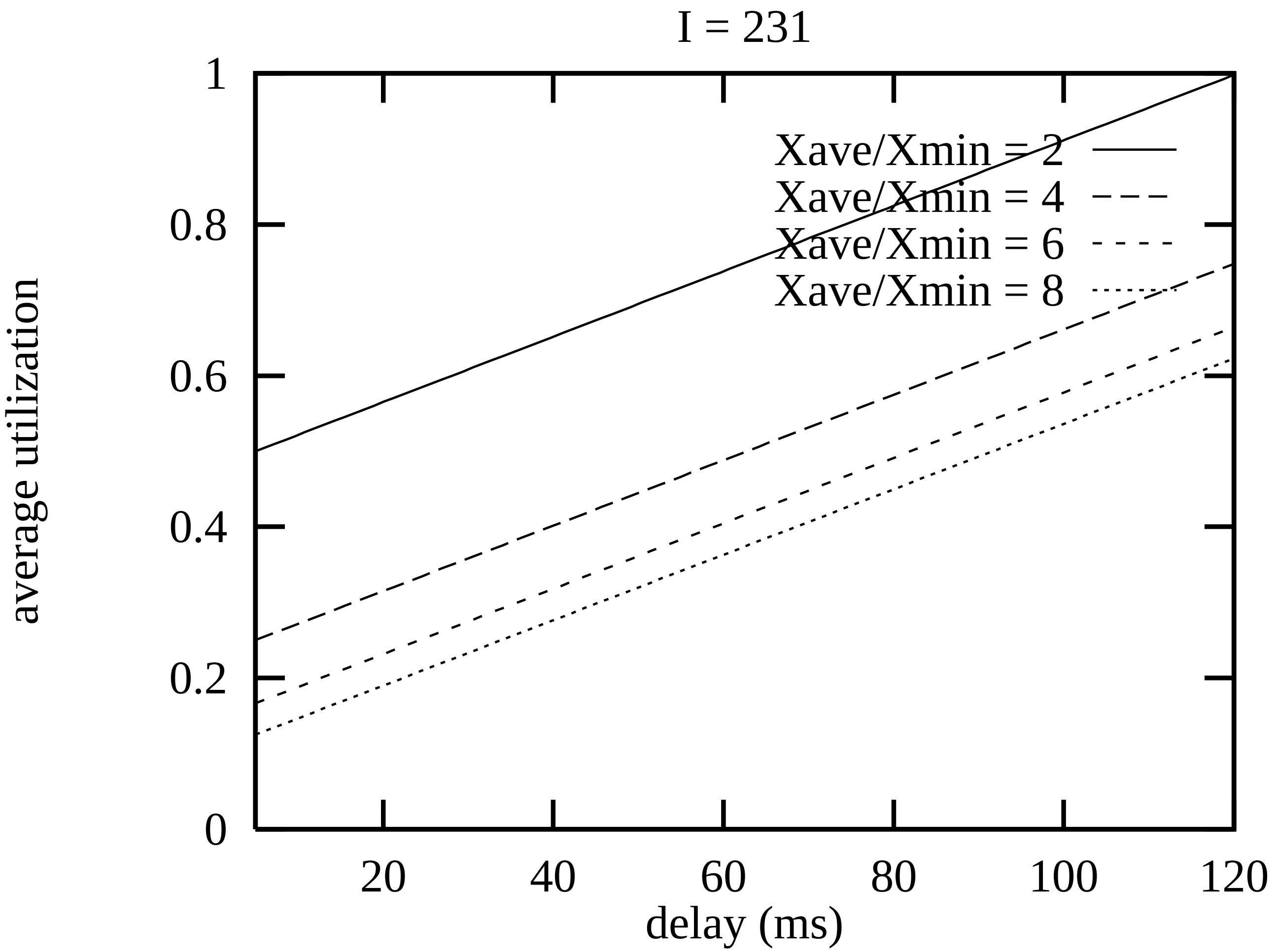
Numerical Examples

- Variable Bit Rate (VBR) video:
  - peak-to-average rate ratio $\sim 1.5 - 4$ [Kanakia 93], [Morgan 93]
  - averaging interval $\sim 3 - 7$ frame times

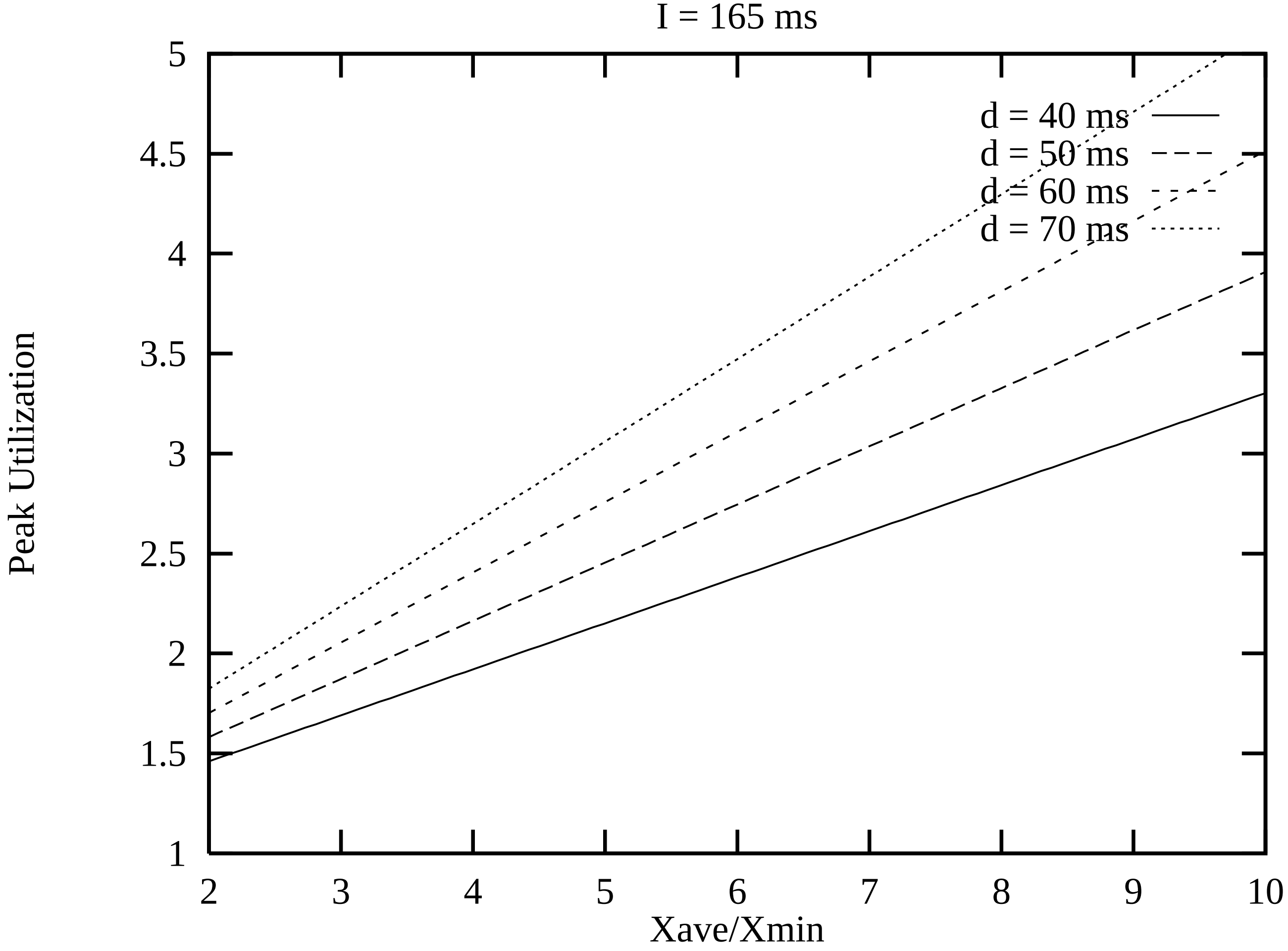
Average utilization vs. $\overline{d}$: I = 165 ms



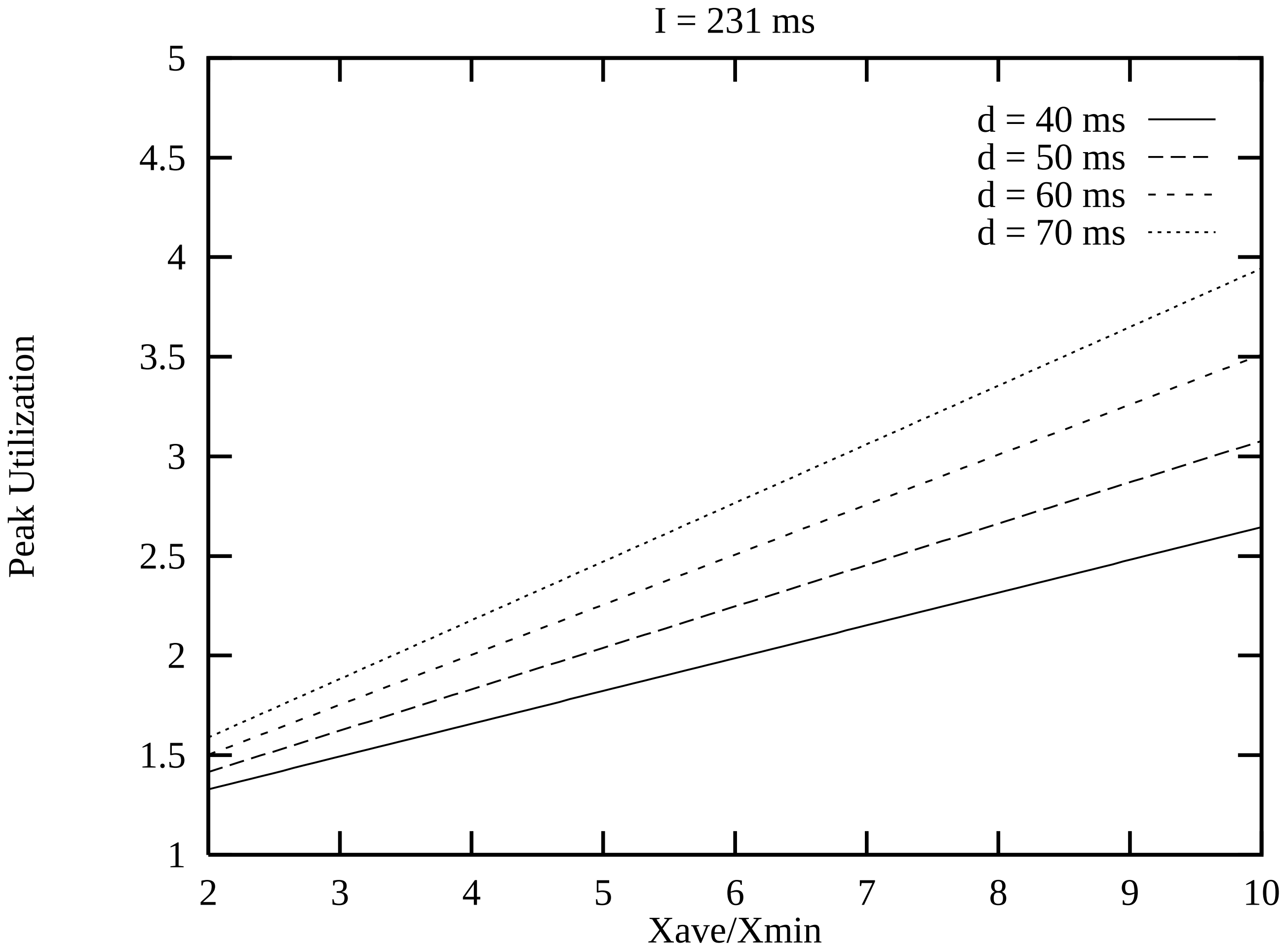
Average utilization vs. $\overline{d}$: I = 231 ms



Improvements: I = 165 ms



Improvements: I = 231 ms



Improvements

- Example: VBR video with peak rate 3 Mbps over T3 (45Mbps) link
- Old Admission Control: maximum 15 connections
- New Admission Control:

I	d	Burst Ratio	# Connection	Factor
165	40	3	25	1.69
165	50	4	32	2.16
165	60	5	41	2.76
231	50	4	27	1.83
231	60	5	33	2.26

Summary

- Biggest concern for guaranteed service: network under-utilized
- We can achieve high average utilization for bursty traffic even for deterministic service
- Higher utilization can be achieved by
  - statistical service: Ed Knightly's talk
- Why not guaranteed services, or Tenet protocols?