Report on DIMACS Workshop on Resource Management and Scheduling in Next Generation Networks (March 26-27, 2001)

Report Author:
Matthew Andrews, Bell Labs

Workshop Organizers:
Matthew Andrews, Bell Labs
M. Franklin, University of California, Berkeley
S. Muthukrishnan, AT&T Labs
Rajmohan Rajaraman, Northeastern University
Roy Yates, Rutgers University

1 Workshop Focus

The next generation of communication networks will have a number of features that give rise to new optimization problems. In particular, it is becoming clear that future networks will be characterized by increased mobililty and heterogeneity of users and increased distribution of content. Although these trends have obvious benefits, they also raise fundamental questions about the best way to manage network resources.

For example, the traditional packet scheduling literature assumes that servers have a fixed processing rate that must be shared between clients. However, in a mobile environment the rate at which we can transmit to clients is fundamentally affected by their position. New scheduling algorithms are needed to deal with this spatial aspect. The advent of distributed storage networks gives rise to a number of new problems. For instance, we must make sure that all new content is distributed to the storage sites in a timely fashion and we must ensure the consistency of the content. In addition, whenever a client wishes to retrieve some content, the network must work out which is the best storage site for that client.

This workshop focused on all optimization problems that will arise in new network settings. Specific topics included:

- Physical Layer Issues: channel allocation; rate and power control; contention resolution and multiple access schemes.
- Network and Transport Layer Issues: location management (role of kinetic geometric algorithms and dynamic data structures), distributed directory services.

- Application Layer Issues: data dissemination by broadcast servers, management of replicated data in wide-area heterogeneous networks, (caching, replication, and consistency), addressing the limited capability of mobile stations and adapting web content.

2 Talk Abstracts

The following were prepared by the speakers and mildly edited.

Increased Wireless Capacity Via Scattering

Mike Andrews, Bell Labs

Results on increasing wireless capacity that have recently received a lot of attention will be presented. We will show that in a scattering environment it is possible to increase capacity by transmitting through channels with different polarization states.

Fine-Grained Layered Multicast

John Byers, Boston University

Traditional approaches to receiver-driven layered multicast have advocated the benefits of cumulative layering, which enables coarse-grained congestion control and complies with TCP-friendliness equations over large time scales. In this talk, we qualify and quantify the costs and benefits of using non-cumulative layering and present a scalable multicast congestion control scheme which provides a fine-grained approximation to the behavior of TCP additive increase / multiplicative decrease (AIMD).

Our results show that careful design of non-cumulative layering sequences can provide an effective, fine-grained, TCP-friendly congestion control solution. In contrast to the conventional wisdom, we demonstrate that fine-grained rate control can be achieved with only modest increases in the number of layers and aggregate bandwidth consumption while using a small constant number of join and leave operations per rate adjustment.

We will demonstrate how our technique integrates naturally with existing methods for content delivery using reliable multicast implemented with fast forward error correcting (FEC) codes.

Joint work with Michael Luby (Digital Fountain) and Michael Mitzenmacher (Harvard).

Frame Structure Adaptation for QoS-based Resource Management in Wireless Packet Data Networks

Sajal K. Das, University of Texas at Arlington

In this talk we will present an algorithmic framework for bandwidth allocation in wireless packet data networks, based on the data rate and inter-packet delay requirements of the traffic. The proposed framework is designed to operate at an optimal point which maximizes the system performance in terms of the number of users served and the promised QoS requirements. The operational point changes with varying airlink frame loss probability, which is captured through adaptation of the frame structure, thus controlling the admission of packets into the system.

The performance of our framework will be studied analytically as well as through simulation experiments. Results will be compared with the performance of IS-136 system. We will show an improvement of up to 500% in the number of users in the system at the cost of a drop in the throughput by at most 46%. In spite of the lower system throughput, the adaptive frame structure algorithm has fewer QoS delay violations with changes in the airlink frame loss rate.

Scheduling Your Network Connections

Mor Harchol-Balter, Carnegie Mellon

This talk proposes a method for improving the performance of Web servers servicing static HTTP requests. The idea is to give preference to those requests which are quick, or have small remaining processing requirements, in accordance with the SRPT (Shortest-Remaining-Processing-Time) scheduling policy.

The implementation is at the kernel level and involves controlling the order in which socket buffers are drained into the network. Experiments use the Linux operating system and the Apache and Flash web servers.

Empirical results indicate that SRPT-based scheduling of connections yields significant reductions in mean response time, mean slowdown, and variance in response time at the Web server. Most importantly, and counter to intuition, the large requests are not at all (or hardly) penalized by SRPT-based scheduling.

To explain the above results, we present some of our new theorems on SRPT scheduling under heavy-tailed workloads. These theorems motivate our implementation work.

This research is joint with my students: Mukesh Agrawal, Nikhil Bansal, and Bianca Schroeder.

Practical LFU Implementation for Web Caching

George Karakostas, Telcordia Technologies, and Dimitrios Serpanos, University of Patras

Web caches can achieve high hit rates through exploitation of the properties of object access distribution, which is governed by Zipf's law, in general. Web caches need to store the most popular objects and typically employ the Least Frequently Used (LFU) replacement policy, which achieves high cache hit rates, and often the highest hit rate. Correct implementation of LFU requires that replacement decisions are made based on frequency access information (popularity), for all the objects accessed since the beginning of a cache's operation. The immensely large number of such objects renders the implementation of LFU impractical in many environments.

In this paper, we introduce an alternative implementation of LFU, the Window-LFU policy, which makes replacement decisions based on access frequency measurements in a recent past, called time-window. Window-LFU achieves cache hit rates equivalent to those of LFU, but with access information from a shorter history, leading to high performance at a low cost (significantly lower than that of LFU). We provide analytical results which enable one to estimate the appropriate window size, in order to achieve the target cache hit rate of LFU. Furthermore, we present simulation results using actual traces, which indicate that the proposed Window-LFU policy behaves as expected and in some configurations it leads to better results than theoretically expected, due to dependencies between successive Web objects requests in real environments. Our theoretical and simulation results demonstrate that Window-LFU provides an efficient solution for effective Web caches at a low cost, due to its shorter history measurements.

Broadcast Scheduling for Set Requests

Vincenzo Liberatore, Case Western Reserve University

Advances in wireless and optical communication, as well as in Internet broadcast protocols, make broadcast methods an effective solution to disseminate data. In particular, repetitive server-initiated broadcast (broadcast disks) is an effective technique in wireless systems and is a scalable solution to relieve Internet hot spots. A critical issue for the performance of broadcast disks is the schedule of the broadcast. Most previous work focused on a model where each data item is requested by clients with a certain probability that is independent of past accesses. In this paper, we consider the more complex scenario where a client accesses pages in blocks (e.g. a HTML file and all its embedded images), thereby introducing dependencies in the pattern of accesses to data. We present simple heuristics that exploit page access dependencies. We measured the resulting client-perceived delay on multiple Web server traces, and observed a speed-up over previous methods ranging from 5% to 34%. We conclude that scheduling for multi-item requests is a critical factor for the performance of Web-based broadcast disks.

Expressing Statistical Multiplexing Gain with Effective Envelopes

Jorg Liebeherr, University of Virginia

A statistical network service which allows a certain fraction of traffic to not meet its QoS guarantees can extract a large amount of additional capacity from a network by exploiting statistical properties of traffic. Recently, researchers have investigated the statistical multiplexing gain for so-called regulated adversarial traffic. Here, one assumes that each individual flow generates traffic in a worst-case fashion as permitted by a traffic conditioner, but that traffic from multiple flows is statistically independent. The talk discusses the recently developed notion of "effective envelopes", which are, with high certainty, upper bounds on the aggregate traffic of regulated flows. We show that effective envelopes can be used to devise admission control tests for a statistical service, and that admission control for a statistical service can be done in a similar fashion as with deterministic envelopes for a deterministic service. Also, we present recent results of using effective envelopes for calculating end-to-end statistical QoS guarantees.

Resource Allocation Under SLA Constraints

Zhen Liu - IBM

Service Level Agreements (SLAs) have been receiving increasing interest in a number of computer application domains. In this paper we study the problem of maximizing profits based on SLA contracts. Our approach consists of formulating the SLAP as a network flow model with a separable set of concave cost functions at the servers (or sink nodes). The SLA classes are taken into account in both the constraints and the cost function. We then iteratively solve a pair of optimization problems based on queueing-theoretic formulas. This formulation simultaneously yields both the load balancing and server scheduling parameters.

Joint with Mark S. Squillante and Joel L. Wolf

Duality Model of TCP and Queue Management Algorithms

Steve Low, California Institute of Technology

We describe a duality model of TCP and active queue management algorithms. Congestion control is the interaction of source rates with certain congestion measures in the network. The basic idea is to regard source rates as primal variables and congestion measures as dual variables, and congestion control a Lagrangian method that iterates on source rates and congestion measures to maximize aggregate source utility subject to a capacity constraint. In TCP, the primal iteration is carried out by source algorithms such as Reno or Vegas, and the dual iteration is carried out by queue management such as DropTail, RED or REM. We present these algorithms and derive their utility functions.

Scheduling Broadcasts in Wireless Networks

Kirk R. Pruhs, University of Pittsburgh

We consider problems involving how to schedule broadcasts in a pulled-based data-dissemination service, such as the DirecPC system, where data requested by the clients is delivered via broadcast. In particular, we consider the case where all the data items are of equal size and preemption is not allowed. We give an offline O(1)-speed O(1)-approximation algorithm for the problem of minimizing the average response time. We provide worst-case analysis, under various objective functions, of the online algorithms that have appeared in the literature, namely, Most Requests First, First Come First Served, and Longest Wait First.

Resource Management Issues in Content Delivery Networks (CDNs)

Michael Rabinovich, AT&T Labs - Research

This talk outlines some research issues concerning resource management in CDNs. First, I will discuss potential pitfalls and merits of DNS-based request scheduling, the prevalent request distribution mechanism of today's CDNs. Conflicting claims have been made on the severity of these pitfalls and the authoritative study quantifying these claims is long overdue. Some methodological remarks on conducting such a study will be provided. The second direction concerns an emerging use of CDNs for providing scalable access to applications. As an application may require large amounts of underlying data to execute, existing replica placement methods that simply cache responses at the edge servers may not be appropriate. I will discuss some alternative approaches for replica placement and their interaction with server scheduling. Finally, I will describe an interesting facility location problem that arises in CDNs for streaming data.

On-Line Algorithms for the Maximum Disjoint Paths Problem

Christian Scheideler, Johns Hopkins University

The maximum disjoint paths problem is one of the classical problems in the routing theory area. Given a set of source-destination pairs, the problem is to find the maximum subset of pairs for which there exist edge disjoint paths. In my talk I will give an overview of what is known about this problem in general and for specific graphs. Afterwards, I will present different ways of characterizing networks with the help of suitable parameters and I will demonstrate that there are classes of on-line algorithms that can get close to an optimal compeititive ratio when only these parameters are known. Furthermore, I will show how the on-line techniques can be used to obtain good approximation algorithms for generalizations of the maximum disjoint paths problem.

Content Delivery Using Error Correcting Codes

Amin Shokrollahi, Digital Fountain

We will describe new coding schemes that can be used for the distribution of popular content to large numbers of recipients.

Path Computation and Network Planning Problems in Optical Networks

Rakesh K. Sinha, Ciena Core Switching Division

Today's Internet, for the most part, deploys best effort routing. That is, its primary focus is on providing connectivity without any assurance of service quality. However there has been an increasing demand for quality of service. The current generation of routers and switches invariably support classification of IP packets, multi-constrained path computation for MPLS or ATM, and dynamic path restoration.

In this talk, I will describe a number of these algorithmic and network planning problems. Several of these are variations of clasical problems, with additional constraints on reliability. This is an excellent opportunity for the algorithm community to influence the design and architecture of next generation switches and routers.

Algorithms and Architectures for Optical Packet Fabrics

Dimitri Stiliadis, Bell Labs

Due to developments in optical and DWDM technologies, link speeds are approaching or even exceeding memory bandwidths making the design of high-capacity packet switches and routers extremely challenging. Input-Output buffered crossbars are popular building blocks for scalable high-speed switching because they require minimum speed-up of memory bandwidth. However, scaling the design of systems based on electronic crossbar switches to large capacities is limited by technology issues such as the reconfiguration of high-speed fabrics, very high power consumption, and complexities in distributing the building blocks of the system over several bays of equipment. In addition, when crossbar architectures are used in the context of IP networks, where packets are of variable size, they typically schedule and transfer fixed size *envelopes*, suffering from the well known fragmentation effect.

In this talk, we will discuss architectures and algorithms that enable the use of all optical switching fabrics as the building blocks for high-capacity input/output crossbars targeted explicitly to IP networks. We will also present methods that allow us to solve the problem of packet fragmentation with minimal penalty in terms of latency.

Content Delivery Networks - Practice and Theory

Ravi Sundaram, R&D, Akamai Technologies, Inc.

We present an overview of the Internet from the standpoint of content delivery. We identify issues and bottlenecks in the speedy delivery of content, setting the stage for a description of the Akamai solution.

The talk then briefly describes the history and evolution of Akamai, its products and services. We walk through an interaction between a browser and an Akamaized website.

We close with a description of some (theoretical) problems arising out of the practical considerations at Akamai. The problems range from load balancing to detection of network failures.

TCP and Active Queue Management

D. Towsley, University of Massachusetts

In this talk we use fluid models to model the interactions between sessions using the Internet TCP transport protocol and routers implementing active queue management (AQM). Using these models, we examine the behavior of RED (random early detect), the AQM policy currently in use in the Internet. Careful analysis of the model is able to explain many of the problems manifested by RED that have been previously observed by others. We then show how these problems can be ameliorated by a new AQM policy based on the classical PI controller.

Joint work with W. Gong, C. Hollot, and V. Misra.

3 Workshop Talks Summary

The allocation of scarce resources is one of the most prevalent problems in the design and operation of current and future networks. If multiple users wish to access a shared medium then scheduling decisions must be made as to who should access it when. Our aim was to invite a mix of speakers from academia and industry who are working at the forefront of resarch on these problems.

The first session of the workshop was devoted to wireless networks. Mike Andrews of Bell Labs began by presenting results on increasing wireless capacity that have recently received a lot of attention. He demonstrated that in a scattering environment it is possible to increase capacity by transmitting through channels with different polarization states. Sajal Das from the University of Texas then showed how TDMA networks can be optimized by altering the TDMA frame structure. There were two talks on TCP and active queue management. Don Towsley of the University of Massachusetts began with a primer on TCP and the Random Early Detection (RED) scheme for queue management. He presented a control-theoretic framework for tuning parameters and demonstrated that RED is reactive to congestion whereas it should be proactive. He then presented a new scheme called PI that addresses some of the shortcomings of RED. Steven Low of Caltech demonstrated that TCP and RED can be viewed as two components of a primal-dual algorithm that converges to optimal transmission rates in an asynchronous environment. Ravi Sundaram from Akamai Technologies described how Akamai is tackling many of the congestion problems in the Internet by pushing content out to the network edge.

Rakesh Sinha of Ciena gave an open-problem talk on problems arising in the design of optical networks. A problem that generated a lot interest was the fault-tolerant routing of lightpaths in the case of coordinated link failures. Here each link belongs to a number of failure groups. Each lightpath must have a primary path and a backup path such that if any failure group of links goes down, all of the affected light paths can be quickly restored. Dimitri Stiliadis from Bell Labs continued the optical theme by presenting a queueing scheme for an optical switch fabric. He showed that the scheme can give good performance even if the fabric cannot be reconfigured on a packet timescale. Zhen Liu of IBM finished the session with a theoretical analysis of Service Level Agreements in communication networks.

Rajiv Vijayakumar of the University of Michigan gave a talk on power allocation in wireless networks. He presented an iterative algorithm that outperforms autonomous power allocation algorithms but which only requires communication between neighboring basestations. Christian Scheideler of Johns Hopkins University gave a new algorithm for the online maximum disjoint paths problem that uses a new parameter called the routing number.

The second day began with two talks on coding and multicast. Amin Shokrollahi of Digital Fountain described new coding schemes that can be used for the distribution of popular content to large numbers of recipients. John Byers of Boston University then presented a new scheme for multicast which achieves true additive increase, multiplicative decrease congestion control. This scheme transmits in layers, each of which corresponds to a different bandwidth.

There were five presentations on scheduling. Mor Harchol-Balter from Carnegie Mellon University demonstated that if a web server uses the Shortest Remaining Processing Time (SRPT) scheduling discipline then jobs have a better response time than if Processor Sharing (PS) were used. She also described an implementation in which the socket queues on a Linux web server were drained according to SRPT. Jorg Liebeherr from the University of Virginia showed how statistical Quality-of-Service delay bounds may be derived using the concept of effective envelopes. Sasha Stolyar from Bell Labs finished the session by describing how we can improve performance in wireless networks if we always try to transmit to users that are in "relatively good" positions. Then there were two talks on scheduling for broadcast disks. A broadcast disk is a single server (e.g., a satellite) that can broadcast to a large number of clients simultaneously. Vincenzo Liberatore from Case Western Reserve University described how to schedule the disk when there is no uplink communication from the clients to the server. Kirk Pruhs of the University of Pittsburgh analyzed the problem when there is a small uplink channel from the clients to the server.

The workshop concluded with two presentations related to web caching and content delivery. George Karakostas of Telcordia Technologies presented a practical approximation to the Least Frequently Used (LFU) web caching scheme that only takes into account recent requests. He showed that on real workloads the approximation often outperforms true LFU. Michael Rabinovich of AT&T analyzed the problem of overlaying trees for content delivery on top of a physical topology. He then presented a number of problems related to DNS request distribution in Content Delivery Networks.

4 Research Directions

The workshop demonstrated that networking is an extremely vibrant area of research. Some of the problems raised are tough problems that have been around for a number of years now but still allow for a great deal of fruitful study. Examples include wireless scheduling and the optimization of TCP-like flow control.

The workshop also considered some newer topics that have inpired a large body of research in a short time. One example is Content Delivery Networks (CDNs). The success of Akamai has shown that pushing content out to the network edge can bypass many of the bottlenecks present in today's networks. However, the problem of how to manage a CDN leads to a host of rich and interesting problems. These problems were addressed directly in the presentations of Ravi Sundaram and Michael Rabinovich. Moreover, the talks on multicast, broadcast disks and web caching were also extremely pertinent to the study of CDNs.

Another area that has received a lot of attention recently and is likely to be an extremely active research area in the future is optical networking. This is an interesting development since many of the problems are related to the routing of lightpaths and these have a “circuit-switched” flavor to them. In recent years the trend in research has been to move away from circuit-switching and focus on more “packet-oriented” problems such as packet scheduling. The types of problems raised in Rakesh Sinha’s talk are likely to bring people back to the circuit switching world. Dimitri Stiliadis’ talk demonstrated that one of the major challenges for the future is to create flexible optical switches. Optical systems have become the technology of choice for transmission but their lack of flexibility in comparison to electronic systems has meant that most switches are still electronic. If we are to scale today’s networks to even faster rates then it is important that we learn how to build efficient optical switches.