****

MONDAY AM

Speakers: Nikhil Bansal (Eindhoven) and Adam Wiermen (Cal Tech)

Title: Worst Case and Stochastic Analysis in Scheduling: Similarities, Differences, and Bridges

Abstract: In order to provide background for the workshop participants, this talk will give a quick introduction to the two areas of worst case analysis and stochastic analysis of scheduling polices. We will introduce the basic notions in each of these areas, and then provide some examples of topics that have been studied by both of the communities. The goal of the talk will be to highlight the differences and similarities in the approaches for addressing these topics. We will also describe some examples of topics where the techniques from one community have proven useful in the other.

****

MONDAY PM

Speaker: Amit Kumar (IIT New Delhi)

Title: Online scheduling Algorithms Analyzed by Dual Fitting

Abstract: I shall talk about a general dual-fitting technique for analyzing online scheduling algorithms in the unrelated machines setting where the objective involves weighted flow-time and we allow the machines of the online algorithm to have slightly extra resources than the offline optimum (the resource augmentation model). In this framework, one can often analyze simple greedy algorithms by considering the dual (or Lagrangian dual) of the linear (or convex) program for the corresponding scheduling problem, and finding a feasible dual solution as the online algorithm proceeds. I shall also mention some recent applications of this technique for deadline scheduling problems.

Speaker: Mark Squillante (IBM Thomas J. Watson Research Center)

Title: Stochastic Optimal Control for a Class of Dynamic Resource Allocation Problems

Abstract: We consider a class of general dynamic resource allocation problems within a stochastic optimal control theoretic framework. This class of problems arises in a wide variety of applications, each of which intrinsically involves resources of different types and demand with uncertainty and/or variability. The goal is to determine the allocation capacity for every resource type in order to serve the uncertain/variable demand and maximize the expected profit (utility) over a time horizon of interest based on the rewards and costs associated with the different resources. We derive the optimal (online) control policy within a singular stochastic optimal control setting, which includes simple expressions for governing the dynamic adjustments to resource allocation capacities over time. Numerical experiments investigate various issues of both theoretical and practical interest, quantifying the benefits of our approach over alternative optimization approaches. This talk is based on joint work with Xuefeng Gao, Yingdong Lu, Mayank Sharma and Joost Bosman.

****

TUESDAY AM

Speaker: Anupam Gupta (CMU)

Title: Stochastic Knapsacks and Matchings

Abstract: I will survey some work on two stochastic packing problems:
(a) stochastic knapsack where jobs of uncertain sizes and/or rewards are packed into a knapsack, and
(b) stochastic matchings where uncertain edges are packed given a set of constraints.
I'll focus on the techniques used for these problems (and how they give solutions for extensions and generalizations), and the many open questions.

Speaker: Devavrat Shah (MIT)

Title: Optimal queue-size scaling in switched networks

Abstract: We consider a switched (queueing) network in which there are constraints on which queues may be served simultaneously; such networks have been used to effectively model input- queued switches, wireless networks and more recently data-centers. The scheduling policy for such a network specifies which queues to serve at any point in time, based on the current state or past history of the system. Designing a scheduling policy with optimal average queue-size for switched network has been a question of interest for a while now. As the main result, we shall discuss a new class of online scheduling policies that achieve optimal scaling for average queue-size for a class of switched networks including input-queued switches. Talk is based on work with Neil Walton (U of Amsterdam)+ Yuan Zhong (UC Berkeley).

****

TUESDAY PM

Speaker: Cliff Stein (Columbia)

Title: Online Stochastic Matching

Abstract: Online matching is an important and well-studied problem and has received particular attention recently because of its connections to allocation problems in internet advertising.

In the traditional worst-case model, on-line algorithms are well understood, with tight competitive ratios of 1/2 for deterministic algorithms and (e-1)/e for randomized algorithms. Several recent works have considered relaxing the worst-case model and considering various probabilistic models for on-line matching. We will survey some of the models and results, and also describe an online algorithm in a model in which an adversary chooses the graph, but does not control the order in which the online nodes are revealed (This work is joint with Feldman, Korula, Henzinger, and Mirrokni). We will also show how these ideas generalize to online packing problems

Speaker: Gideon Weiss (Univ. Haifa)

Title: FCFS infinite matching, queues with skill based routing, and organ transplants

Abstract: This talk is a survey of recent work with Ivo Adan and several other collaborators, including Cor Hurkens, Marko Boon, Ana Busic, and Jean Mairesse. It is based on earlier work of Rishi Talreja and Ward Whitt, and of Rene Caldentey and Ed Kaplan.

In recent years it has become very important to investigate service systems which serve cus- tomers of several types, and which employ servers of various skills. Such service systems are referred to in current literature as queues with skill based routing. These have several types of customers, and several types of servers, and a bipartite compatibility graph to indicate which types of servers can serve which types of customers. Applications include such varied fields as call centers, outsourcing, manufacturing process, cloud computing and health systems.

A somewhat different model is the matching of applicants and positions, of organ donors and patients, of adoptive parents and children, the so called marriage model. These two types of service models motivate our research, we note that these applications are also studied by schedulers and combinatorial optimizers.

There is a significant difference between these two models: While in a queueing model the sequence of interarrival times and the sequence of service times get hopelessly entangled through the busy/idle cycles of the servers, this is absent from the matching applications: Here customers and servers play a symmetric role and each customer server encounter has no after effects on the rest of the system. The taxi rank is a simple example of that.

A further simplification is to consider just the sequence of customers and of servers, ordered by arrivals and classified by types: We think of those as i.i.d infinite sequences of customer types (all the customers that will appear in the future) and of server types (all the services that will be given).

For these sequences it is natural to define a FCFS matching: The first server is matched to the first compatible customer, and the nth server is matched to the first compatible customer that none of the previous n − 1 servers picked up.

This infinite FCFS model has a much more combinatorial flavor than the standard queueing models. It turns out that it is quite tractable. The key property which we discovered is that this infinite matching model is in some sense time reversible. Reversibility plays a key role in the theory of Markov chains and of queueing models. It is the property which underlies product form results. It is a sad fact that most queueing modelsare complicated and often intractable. This is true even for the single server queue, with theshining exception of the M/M/1 queue, which in stationary form has exponential sojourn times and geometric (1 − ρ) ρ n queue length distribution, and for which many other quantities can be calculated by explicit formulae. Research on queueing networks would have gone nowhere if it weren't for Jackson's discovery that Jackson networks have steady state distribution

given byProduct (1 − ρ _i) ρ _^n_i . From that time onwards, product form results have been keenly sought after, as ithey seem the best way of getting explicit solutions and useful insight for more general models.
Product form is always related to some form of reversibility. The reversibility of the infinite matching model underlies all our further results.
Coming back to queues with skilled based routing, we focus on FCFS-ALIS — first come first served, assign longest idle server — policy.
This means that whenever a server becomes available he will go to the longest waiting customer in the system which he can serve, and whenever a customer arrives to find several idle servers he will be assigned to the longest idle compatible server. This policy has several attractive features: First and foremost it is fair to both customers and servers — in many systems e.g. organ donations, public housing assignment, FCFS is dictated by law. ALIS is the best way to equalize the efforts of the servers, and thus it encourages diligent service.
The policy is also very natural and easy to implement, and it requires minimal information about the parameters of the system and its current state.
As a result it is useful in systems in which load and staffing keep changing over the operating horizon.
Our exact results are that under a FCFS-ALIS policy when arrivals are Poisson and services exponential we have product form queue length distributions, and closed form expressions for waiting times. Furthermore, these results can be used to obtain excellent approximations for much more general queueing models, under many server scaling. These include general link dependent service times, abandonments, and efficiency driven systems.
We will present some examples of how to use these results in the design of call centers, and in the planning of an organ transplants policy.

****

WEDNESDAY AM

Speaker: Kamesh Munagala (Duke)

Title: Weakly Coupled Stochastic Systems and Linear Programming

Abstract: Several problems in stochastic optimization and decision theory have the property that they are composed of many independent decision sub-problems coupled together by a few constraints. We present several examples of such problems from diverse application areas such as wireless communication, design of experiments, mechanism design, and budgeted allocations. We present unifying solution techniques based on linear programming and duality, and show connections to well-known heuristics used in practice. The talk is self-contained and will primarily focus on recent results in auction design.

Speaker: John Hasenbein (Univ. Texas)

Title: When Does Stochasticity Matter? Fluid Models and Scheduling Queueing Networks

Abstract: We discuss the connection between scheduling multiclass stochastic and fluid (deterministic) networks. First we overview the relationship via a few of classic results and examples. Next, we present more recent research on a stochastic combinatorial scheduling problem in which the "macro" stochasticity must be taken into account, but the "micro" stochasticity is less important.

****

THURSDAY AM

Speaker: Yossi Azar (Tel-Aviv University)

Title: Online Scheduling in the Cloud

Abstract: Online task scheduling of jobs on cloud computing infrastructures poses new challenges theoretically and practically. In particular the number of machines or virtual machines(VM) is not fixed any more and the goal is to minimize the cost of the computation as well as minimize the delay or the load.

We will discuss various models and questions in this area concerned with identical vs heterogeneous machines, fixed setup time vs arbitrary setup cost and single vs multi dimension job requirements.

The talk is based on three papers (but the time is too short so I will not be able to fit them all):

1. one will appear in SODA - unrelated machine scheduling with startup cost (paper is called Online mixed packing and covering)
2. second submitted to STOC - Online Vector Bin Packing
3. third (not submitted) is Cloud Scheduling with Setup Cost

Speaker: Alexander Stolyar (Alcatel-Lucent, Murray Hill)

Title: An infinite server system with customer-to-server packing constraints

Abstract: The model is motivated by the problem of efficient "packing" of virtual machines into physical host machines in a network cloud data center.

There is an infinite number of servers and multiple flows of arriving customers of different types. Each server can simultaneously serve several customers, subject to some "packing" constraints. Service times of different customers are independent -- even if customers share a server. Customers leave after their service is complete. The underlying objective is to minimize the number of occupied servers.

We show that some versions of a greedy packing strategy are asymptotically optimal as the system scale (the average total number of customers in service) goes to infinity.

****

THURSDAY PM

Speaker: Nicole Megow (TU Berlin)

Title: Approximation in Stochastic Scheduling

Abstract: Stochastic scheduling is concerned with scheduling problems in which job processing times are modeled as random variables with known probability distributions. The actual processing times are revealed only upon completion of the jobs. Such problems have been addressed since the 70s, but only more recently approximation results were derived. We give an overview of results and methods for obtaining provably good scheduling policies. This involves linear programming, lower bounding techniques borrowed from online scheduling, and index-based dynamic allocation rules known from multi-armed bandit problems. We discuss open problems, further research directions, and possible connections to other areas.

Speaker: Sem Borst (TU/e & Alcatel-Lucent Bell Labs)

Title: Wireless Random-Access Algorithms: Fluid Limits and Delay Issues

Abstract: Queue-based wireless random-access algorithms are relatively simple and inherently distributed, yet provide a striking capability to match the optimal throughput performance of centralized scheduling mechanisms in a wide range of scenarios. Unfortunately, the specific type of activation rules for which throughput optimality has been established, may result in extremelylong queues and delays.
The use of more aggressive/persistent access schemes can improve the delay performance, but does not provide any universal maximum-stability guarantees.
In order to gain qualitative insights and examine stability properties, we investigate fluid limits where the system dynamics are scaled in space and time.
Several distinct types of fluid limits can arise, ranging from ones with smooth deterministic features, to others which exhibit random oscillatory characteristics, depending on the topology of the network, in conjunction with the form of the activation rules. As we will show, these qualitatively different regimes are strongly related to short-term fairness measures and mixing times for random-access mechanisms with fixed activation rates, and carry significant implications for stability properties.
Note: based on joint work with Niek Bouman (TU/e), Javad Ghaderi (UIUC),
Johan van Leeuwaarden (TU/e), Alexandre Proutiere (KTH),
Peter van de Ven (IBM), Phil Whiting (Alcatel-Lucent Bell Labs),
Alessandro Zocca (TU/e)

****

FRIDAY AM

Speaker: Benjamin Moseley (Toyota Technological Institute)

Title: MapReduce and Distributed Scheduling

Abstract: Recently, the MapReduce parallel computing framework has become the de facto standard for processing large data. The MapReduce distributed framework consist of an elegant combination of sequential computation and network communication that naturally lends itself to efficient distributed data processing.
In a MapReduce implementation there is a centralized job tracker that coordinates job scheduling. Designing new scheduling policies has been one of the active research topics in MapReduce because of the need to balance often contradictory needs, e.g., system utilization, fairness, and response times. In this talk, we will first focus on introducing the fundamentals of MapReduce. Then we will discuss several

scheduling issues that arise in MapReduce as well as recent developments in the theoretical scheduling community that have addressed these issues.

Speaker: Ger Koole (Vrije Universiteit Amsterdam)

Title: Employee scheduling and rescheduling in call centers

Abstract: In call centers, many parameters are still uncertain the moment employees are scheduled. This leads to the necessity of real-time adjustments to the schedule. This requires different forms of flexibility in the initial schedule.
Ideally, when making agent schedules the right amount of flexibility should be introduced. In this talk we discuss the different forms of parameter uncertainty, different ways to do rescheduling, and how this can be incorporated in the initial schedule.