

An EPTAS for Scheduling on Unrelated Machines of Few Different Types*

Klaus Jansen and Marten Maack

Christian-Albrechts-Universität zu Kiel, 24118 Kiel, Germany,
{kj,mmaa}@informatik.uni-kiel.de

Abstract. In the classical problem of scheduling on unrelated parallel machines, a set of jobs has to be assigned to a set of machines. The jobs have a processing time depending on the machine and the goal is to minimize the makespan, that is, the maximum machine load. It is well known that this problem is NP-hard and does not allow polynomial time approximation algorithms with approximation guarantees smaller than 1.5, unless P=NP. We consider the case that there is only a constant number K of machine types. Two machines have the same type, if all jobs have the same processing time for them. We present an efficient polynomial time approximation scheme (EPTAS) for this problem, that is, for any $\varepsilon > 0$ an assignment with makespan of length at most $(1+\varepsilon)$ times the optimum can be found in polynomial time in the input length and the exponent is independent of $1/\varepsilon$. In particular we achieve a running time of $2^{\mathcal{O}(K \log(K)^{1/\varepsilon} \log^4 1/\varepsilon)} + \text{poly}(|I|)$, where $|I|$ denotes the input length. Furthermore, we study the case where the minimum machine load has to be maximized and achieve a similar result.

1 Introduction

We consider the problem of scheduling jobs on unrelated parallel machines—or unrelated scheduling for short—in which a set \mathcal{J} of n jobs has to be assigned to a set \mathcal{M} of m machines. Each job j has a processing time p_{ij} for each machine i and the goal is to find a schedule $\sigma : \mathcal{J} \rightarrow \mathcal{M}$ minimizing the *makespan* $C_{\max}(\sigma) = \max_{i \in \mathcal{M}} \sum_{j \in \sigma^{-1}(i)} p_{ij}$, i.e., the maximum machine load. The problem is one of the classical scheduling problems studied in approximation. In 1990 Lenstra, Shmoys and Tardos [20] showed that there is no approximation algorithm with an approximation guarantee smaller than 1.5, unless P=NP. Moreover, they presented a 2-approximation and closing this gap is a rather famous open problem in scheduling theory and approximation (see e.g. [23]).

In particular we study the special case where there is only a constant number K of *machine types*. Two machines i and i' have the same type, if $p_{ij} = p_{i'j}$ holds for each job j . In many application scenarios this scenario is plausible, e.g. when considering computers which typically only have a very limited number of different

* This work was partially supported by the German Research Foundation (DFG) project JA 612/16-1.

types of processing units. We denote the processing time of a job j on a machine of type $t \in [K]$ by p_{tj} and assume that the input consist of the corresponding $K \times n$ processing time matrix together with machine multiplicities m_t for each type t , yielding $m = \sum_{t \in [K]} m_t$. Note that the case $K = 1$ is equivalent to the classical scheduling on identical machines.

We will also consider the reverse objective of maximizing the minimum machine load, i.e., $C_{\min}(\sigma) = \min_{i \in \mathcal{M}} \sum_{j \in \sigma^{-1}(i)} p_{ij}$. This problem is also known as max-min fair allocation or the Santa Claus problem. The intuition behind these names is that the jobs are interpreted as goods (e.g. presents), the machines as players (e.g. children), and the processing times as the values of the goods from the perspective of the different players. Finding an assignment that maximizes the minimum machine load, means therefore finding an allocation of the goods that is in some sense fair (making the least happy kid as happy as possible). We will refer to the problem as Santa Claus problem in the following, but otherwise will stick to the scheduling terminology.

We study approximation algorithms: Given an instance I of an optimization problem, an α -approximation A produces a solution in time $\text{poly}(|I|)$, where $|I|$ denotes the input length. For the objective function value $A(I)$ it is guaranteed that $A(I) \leq \alpha \text{OPT}(I)$, in the case of an minimization problem, or $A(I) \geq (1/\alpha) \text{OPT}(I)$, in the case of an maximization problem, where $\text{OPT}(I)$ is the value of an optimal solution. We call α the *approximation guarantee* or *rate* of the algorithm. In some cases a polynomial time approximation scheme (PTAS) can be achieved, that is, for each $\varepsilon > 0$ an $(1 + \varepsilon)$ -approximation. If for such a family of algorithms the running time can be bounded by $f(1/\varepsilon) \text{poly}(|I|)$ for some computable function f , the PTAS is called *efficient* (EPTAS), and if the running time is polynomial in both $1/\varepsilon$ and $|I|$ it is called *fully polynomial* (FPTAS).

Related work. It is well known that the unrelated scheduling problem admits an FPTAS in the case that the number of machines is considered constant [13] and we already mentioned the seminal work by Lenstra et al. [20]. Furthermore, the problem of unrelated scheduling with a constant number of machine types is strongly NP-hard, because it is a generalization of the strongly NP-hard problem of scheduling on identical parallel machines. Therefore an FPTAS can not be hoped for in this case. However, Bonifaci and Wiese [5] showed that there is a PTAS even for the more general vector scheduling case. However, in the case considered here, their algorithm has to solve m to the power of $\mathcal{O}(K(1/\varepsilon)^{1/\varepsilon \log 1/\varepsilon})$ linear programs. Gehrke et al. [10] presented a PTAS with an improved running time of $\mathcal{O}(Kn) + m^{\mathcal{O}(K/\varepsilon^2)} (\log(m)/\varepsilon)^{\mathcal{O}(K^2)}$ for unrelated scheduling with a constant number of machine types. On the other hand, Chen et al. [7] showed that there is no PTAS for scheduling on identical machines with running time $2^{(1/\varepsilon)^{1-\delta}}$ for any $\delta > 0$, unless the exponential time hypothesis fails. Furthermore, the case $K = 2$ has been studied: Imreh [14] designed heuristic algorithms with rates $2 + (m_1 - 1)/m_2$ and $4 - 2/m_1$, and Bleuse et al. [4] presented an algorithm with rate $4/3 + 3/m_2$ and moreover a (faster) $3/2$ -approximation, for the case that for each job the processing time on the second machine type is

at most the one on the first. Moreover, Raravi and Nélis [22] designed a PTAS for the case with two machine types.

Interestingly, Goemans and Rothvoss [11] were able to show that unrelated scheduling is in P, if both the number of machine types and the number of job types is bounded by a constant. Job types are defined analogously to machine types, i.e., two jobs j, j' have the same type, if $p_{ij} = p_{ij'}$ for each machine i . In this case the matrix (p_{ij}) has only a constant number of distinct rows and columns. Note that already in the case we study, the rank of this matrix is constant. However the case of unrelated scheduling where the matrix (p_{ij}) has constant rank turns out to be much harder: Already for the case with rank 4 there is no approximation algorithm with rate smaller than $3/2$ unless $P=NP$ [8]. In a rather recent work, Knop and Koutecký [19] considered the number of machine types as a parameter from the perspective of fixed parameter tractability. They showed that unrelated scheduling is fixed parameter tractable for the parameters K and $\max p_{i,j}$, that is, there is an algorithm with running time $f(K, \max p_{i,j})\text{poly}(|I|)$ for some computable function f that solves the problem to optimality.

For the case that the number of machines is constant, the Santa Claus problem behaves similar to the unrelated scheduling problem: there is an FPTAS that is implied by a result due to Woeginger [24]. In the general case however, so far no approximation algorithm with a constant approximation guarantee has been found. The results by Lenstra et al. [20] can be adapted to show that there is no approximation algorithm with a rate smaller than 2, unless $P=NP$, and to get an algorithm that finds a solution with value at least $\text{OPT}(I) - \max p_{i,j}$, as was done by Bezkov and Dani [3]. Since $\max p_{i,j}$ could be bigger than $\text{OPT}(I)$, this does not provide an (multiplicative) approximation guarantee. Bezkov and Dani also presented a simple $(n-m+1)$ -approximation and an improved approximation guarantee of $\mathcal{O}(\sqrt{n} \log^3 n)$ was achieved by Asadpour and Saberi [1]. The best rate so far is $\mathcal{O}(n^\varepsilon)$ due to Bateni et al. [2] and Chakrabarty et al. [6], with a running time of $\mathcal{O}(n^{1/\varepsilon})$ for any $\varepsilon > 0$.

Results and Methodology. In this paper we show:

Theorem 1. *There is an EPTAS for both scheduling on unrelated parallel machines and the Santa Claus problem with a constant number of different machine types with running time $2^{\mathcal{O}(K \log(K)^{1/\varepsilon} \log^4 1/\varepsilon)} + \text{poly}(|I|)$.*

First we present a basic version of the EPTAS for unrelated scheduling with a running time doubly exponential in $1/\varepsilon$. For this EPTAS we use the dual approximation approach by Hochbaum and Shmoys [12] to get a guess T of the optimal makespan OPT . Then we further simplify the problem via geometric rounding of the processing times. Next we formulate a mixed integer linear program (MILP) with a constant number of integral variables that encodes a relaxed version of the problem. We solve it with the algorithm by Lenstra and Kannan [21, 18]. The fractional variables of the MILP have to be rounded and we achieve this with a cleverly designed flow network utilizing flow integrality and causing only a small error. With an additional error the obtained solution

can be used to construct a schedule with makespan $(1 + \mathcal{O}(\varepsilon))T$. This procedure is described in detail in Section 2. Building upon the basic EPTAS we achieve the improved running time using techniques by Jansen [15] and by Jansen, Klein and Verschae [16]. The basic idea of these techniques is to make use of existential results about simple structured solutions of integer linear programs (ILPs). In particular these results can be used to guess the non-zero variables of the MILP, because they sufficiently limit the search space. We show how these techniques can be applied in our case in Section 3. Interestingly, our techniques can be adapted for the Santa Claus Problem, which typically has a worse approximation behaviour. We discuss the ideas needed for this in the last section of the paper. More details and omitted proofs can be found in the long version of the paper [17].

2 Basic EPTAS

In this chapter we describe a basic EPTAS for unrelated scheduling with a constant number of machine types, with a running time doubly exponential in $1/\varepsilon$. Wlog. we assume $\varepsilon < 1$. Furthermore $\log(\cdot)$ denotes the logarithm with basis 2 and for $k \in \mathbb{Z}_{\geq 0}$ we write $[k]$ for $\{1, \dots, k\}$.

First, we simplify the problem via the classical dual approximation concept by Hochbaum and Shmoys [12]. In the simplified version of the problem a target makespan T is given and the goal is to either output a schedule with makespan at most $(1 + \alpha\varepsilon)T$ for some constant $\alpha \in \mathbb{Z}_{>0}$, or correctly report that there is no schedule with makespan T . We can use a polynomial time algorithm for this problem in the design of a PTAS in the following way. First we obtain an upper bound B for the optimal makespan OPT of the instance with $B \leq 2\text{OPT}$. This can be done using the 2-approximation by Lenstra et al. [20]. With binary search on the interval $[B/2, B]$ we can find in $\mathcal{O}(\log 1/\varepsilon)$ iterations a value T^* for which the mentioned algorithm is successful, while $T^* - \varepsilon B/2$ is rejected. We have $T^* - \varepsilon B/2 \leq \text{OPT}$ and therefore $T^* \leq (1 + \varepsilon)\text{OPT}$. Hence the schedule we obtained for the target makespan T^* has makespan at most $(1 + \alpha\varepsilon)T^* \leq (1 + \alpha\varepsilon)(1 + \varepsilon)\text{OPT} = (1 + \mathcal{O}(\varepsilon))\text{OPT}$. In the following we will always assume that a target makespan T is given. Next we present a brief overview of the algorithm for the simplified problem followed by a more detailed description and analysis.

Algorithm 2.

1. Simplify the input via geometric rounding with an error of εT .
2. Build the mixed integer linear program $\text{MILP}(\bar{T})$ and solve it with the algorithm by Lenstra and Kannan ($\bar{T} = (1 + \varepsilon)T$).
3. If there is no solution, report that there is no solution with makespan T .
4. Generate an integral solution for $\text{MILP}(\bar{T} + \varepsilon T + \varepsilon^2 T)$ via a flow network utilizing flow integrality.
5. The integral solution is turned into a schedule with an additional error of $\varepsilon^2 T$ due to the small jobs.

Simplification of the Input. We construct a simplified instance \bar{I} with modified processing times \bar{p}_{tj} . If a job j has a processing time bigger than T for a machine type $t \in [K]$ we set $\bar{p}_{tj} = \infty$. We call a job *big* (for machine type t), if $p_{tj} > \varepsilon^2 T$, and *small* otherwise. We perform a geometric rounding step for each job j with $p_{tj} < \infty$, that is, we set $\bar{p}_{tj} = (1 + \varepsilon)^x \varepsilon^2 T$ with $x = \lceil \log_{1+\varepsilon}(p_{tj}/(\varepsilon^2 T)) \rceil$.

Lemma 1. *If there is a schedule with makespan at most T for I , the same schedule has makespan at most $(1 + \varepsilon)T$ for instance \bar{I} and any schedule for instance \bar{I} can be turned into a schedule for I without increase in the makespan.*

We will search for a schedule with makespan $\bar{T} = (1 + \varepsilon)T$ for \bar{I} .

We establish some notation for the rounded instance. For any rounded processing time p we denote the set of jobs j with $\bar{p}_{tj} = p$ by $J_t(p)$. Moreover, for each machine type t let S_t and B_t be the set of small and big rounded processing times. Obviously we have $|S_t| + |B_t| \leq n$. Furthermore $|B_t|$ is bounded by a constant: Let N be such that $(1 + \varepsilon)^N \varepsilon^2 T$ is the biggest rounded processing time for all machine type. Then we have $(1 + \varepsilon)^{N-1} \varepsilon^2 T \leq T$ and therefore $|B_t| \leq N \leq \log(1/\varepsilon^2)/\log(1 + \varepsilon) + 1 \leq 1/\varepsilon \log(1/\varepsilon^2) + 1$ (using $\varepsilon \leq 1$).

MILP. For any set of processing times P we call the P -indexed vectors of non-negative integers $\mathbb{Z}_{\geq 0}^P$ *configurations* (for P). The *size* $\text{size}(C)$ of configuration C is given by $\sum_{p \in P} C_p p$. For each $t \in [K]$ we consider the set $\mathcal{C}_t(\bar{T})$ of configurations C for the big processing times B_t and with $\text{size}(C) \leq \bar{T}$. Given a schedule σ , we say that a machine i of type t obeys a configuration C , if the number of big jobs with processing time p that σ assigns to i is exactly C_p for each $p \in B_t$. Since the processing times in B_t are bigger than $\varepsilon^2 T$ we have $\sum_{p \in B_t} C_p \leq 1/\varepsilon^2$ for each $C \in \mathcal{C}_t(\bar{T})$. Therefore the number of distinct configurations in $\mathcal{C}_t(\bar{T})$ can be bounded by $(1/\varepsilon^2 + 1)^N < 2^{\log(1/\varepsilon^2 + 1)/\varepsilon \log(1/\varepsilon^2) + 1} \in 2^{\mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)}$.

We define a mixed integer linear program MILP(\bar{T}) in which configurations are assigned integrally and jobs are assigned fractionally to machine types. To this amount we introduce variables $z_{C,t} \in \mathbb{Z}_{\geq 0}$ for each machine type $t \in [K]$ and configuration $C \in \mathcal{C}_t(\bar{T})$, and $x_{j,t} \geq 0$ for each machine type $t \in [K]$ and job $j \in \mathcal{J}$. For $\bar{p}_{tj} = \infty$ we set $x_{j,t} = 0$. Besides this, the MILP has the following constraints:

$$\sum_{C \in \mathcal{C}_t(\bar{T})} z_{C,t} = m_t \quad \forall t \in [K] \quad (1)$$

$$\sum_{t \in [K]} x_{j,t} = 1 \quad \forall j \in \mathcal{J} \quad (2)$$

$$\sum_{j \in J_t(p)} x_{j,t} \leq \sum_{C \in \mathcal{C}_t(\bar{T})} C_p z_{C,t} \quad \forall t \in [K], p \in B_t \quad (3)$$

$$\sum_{C \in \mathcal{C}_t(\bar{T})} \text{size}(C) z_{C,t} + \sum_{p \in S_t} p \sum_{j \in J_t(p)} x_{j,t} \leq m_t \bar{T} \quad \forall t \in [K] \quad (4)$$

With constraint (1) the number of chosen configurations for each machine type equals the number of machines of this type. Due to constraint (2) the variables

$x_{j,t}$ encode the fractional assignment of jobs to machine types. Moreover, for each machine type it is ensured with constraint (3) that the summed up number of big jobs of each size is at most the number of big jobs that are used in the chosen configurations for the respective machine type. Lastly, (4) guarantees that the overall processing time of the configurations and small jobs assigned to a machine type does not exceed the area $m_t\bar{T}$. It is easy to see that the MILP models a relaxed version of the problem:

Lemma 2. *If there is schedule with makespan \bar{T} there is a feasible (integral) solution of $\text{MILP}(\bar{T})$, and if there is a feasible integral solution for $\text{MILP}(\bar{T})$ there is a schedule with makespan at most $\bar{T} + \varepsilon^2 T$. \square*

We have $K2^{\mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)}$ integral variables, i.e., a constant number. Therefore $\text{MILP}(T)$ can be solved in polynomial time, with the following classical result due to Lenstra [21] and Kannan [18]:

Theorem 3. *A mixed integer linear program with d integral variables and encoding size s can be solved in time $d^{\mathcal{O}(d)} \text{poly}(s)$.*

Rounding. In this paragraph we describe how a feasible solution $(z_{C,t}, x_{j,t})$ for $\text{MILP}(\bar{T})$ can be transformed into an integral feasible solution $(\bar{z}_{C,t}, \bar{x}_{j,t})$ of $\text{MILP}(\bar{T} + \varepsilon T + \varepsilon^2 T)$. This is achieved via a flow network utilizing flow integrality.

For any (small or big) processing time p let $\eta_{t,p} = \lceil \sum_{j \in J_t(p)} x_{j,t} \rceil$ be the rounded up (fractional) number of jobs with processing time p that are assigned to machine type t . Note that for big job sizes $p \in B_t$ we have $\eta_{t,p} \leq \sum_{C \in \mathcal{C}_t(\bar{T})} C_p z_{C,t}$ because of (3) and because the right hand side is an integer.

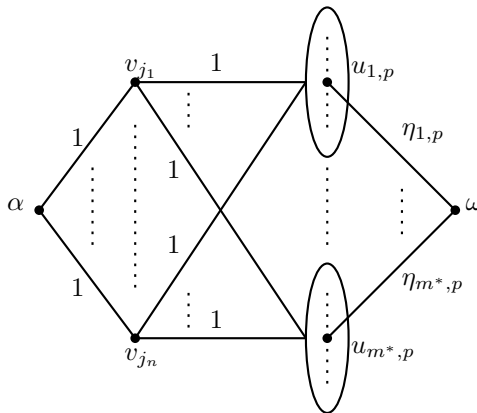


Fig. 1. A sketch of the flow network.

Now we describe the flow network $G = (V, E)$ with source α and sink ω . For each job $j \in \mathcal{J}$ there is a job node v_j and an edge (α, v_j) with capacity

1 connecting the source and the job node. Moreover, for each machine type t we have processing time nodes $u_{t,p}$ for each processing time $p \in B_t \cup S_t$. The processing time nodes are connected to the sink via edges $(u_{t,p}, \omega)$ with capacity $\eta_{t,p}$. Lastly, for each job j and machine type t with $\bar{p}_{t,j} < \infty$, we have an edge $(v_j, u_{t,\bar{p}_{t,j}})$ with capacity 1 connecting the job node with the corresponding processing time nodes. We outline the construction in Figure 1. Obviously we have $|V| \leq (K + 1)n + 2$ and $|E| \leq (2K + 1)n$.

Lemma 3. *G has a maximum flow with value n .*

Proof. Obviously n is an upper bound for the maximum flow, and the solution $(z_{C,t}, x_{j,t})$ for MILP(\bar{T}) can be used to design a flow f with value n , by setting $f((\alpha, v_j)) = 1$, $f((v_j, u_{t,\bar{p}_{t,j}})) = x_{j,t}$ and $f((u_{t,y}, \omega)) = \sum_{j \in J_t(y)} x_{j,t}$. \square

Using the Ford-Fulkerson algorithm, an integral maximum flow f^* can be found in time $\mathcal{O}(|E|f^*) = \mathcal{O}(Kn^2)$. Due to flow conservation, for each job j there is exactly one machine type t^* such that $f((v_j, u_{t^*,y^*})) = 1$, and we set $\bar{x}_{j,t^*} = 1$ and $\bar{x}_{j,t} = 0$ for $t \neq t^*$. Moreover we set $\bar{z}_{C,t} = z_{C,t}$. Obviously $(\bar{z}_{C,t}, \bar{x}_{j,t})$ fulfils (1) and (2). Furthermore (3) is fulfilled, because of the capacities and because $\eta_{t,p} \leq \sum_{C \in \mathcal{C}_t(\bar{T})} C_p z_{C,t}$ for big job sizes p . Utilizing the geometric rounding and the convergence of the geometric series, as well as $\sum_{j \in J_t(p)} \bar{x}_{j,t} \leq \eta_{t,p} < \sum_{j \in J_t(p)} x_{j,t} + 1$, we get:

$$\sum_{p \in S_t} p \sum_{j \in J_t(p)} \bar{x}_{j,t} < \sum_{p \in S_t} p \sum_{j \in J_t(p)} x_{j,t} + \sum_{p \in S_t} p < \sum_{p \in S_t} p \sum_{j \in J_t(p)} x_{j,t} + \varepsilon^2 T \frac{1 + \varepsilon}{\varepsilon}$$

Hence, we have $\sum_{C \in \mathcal{C}_t(\bar{T})} \text{size}(C) \bar{z}_{C,t} + \sum_{s \in S_t} \sum_{j \in J_{t,s}} p_{j,t} \bar{x}_{j,t} < m_t(\bar{T} + \varepsilon T + \varepsilon^2 T)$ and therefore (4) is fulfilled as well.

Analysis. The solution found for MILP(\bar{T}) can be turned into an integral solution for MILP($\bar{T} + \varepsilon T + \varepsilon^2 T$). This can easily be turned into a schedule with makespan $\bar{T} + \varepsilon T + \varepsilon^2 T + \varepsilon^2 T \leq (1 + 4\varepsilon)T$. It is easy to see that the running time of the algorithm by Lenstra and Kannan dominates the overall running time. Since MILP(\bar{T}) has $\mathcal{O}(K/\varepsilon \log 1/\varepsilon + n)$ many constraints, Kn fractional and $K2^{\mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)}$ integral variables, the running time of the algorithm can be bounded by $2^{K2^{\mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)}} \text{poly}(|I|)$.

3 Better running time

We improve the running time of the algorithm using techniques that utilize results concerning the existence of solutions for integer linear programs (ILPs) with a certain simple structure. In a first step we can reduce the running time to be only singly exponential in $1/\varepsilon$ with a technique by Jansen [15]. Then we further improve the running time to the one claimed in Theorem 1 with a very recent result by Jansen, Klein and Verschae [16]. Both techniques rely upon the following result about integer cones by Eisenbrandt and Shmonin [9].

Theorem 4. *Let $X \subset \mathbb{Z}^d$ be a finite set of integer vectors and $b \in \text{int-cone}(X) = \{\sum_{x \in X} \lambda_x x \mid \lambda_x \in \mathbb{Z}_{\geq 0}\}$. Then there is a subset $\tilde{X} \subseteq X$, such that $b \in \text{int-cone}(\tilde{X})$ and $|\tilde{X}| \leq 2d \log(4dM)$, with $M = \max_{x \in X} \|x\|_\infty$.*

For the first improvement of the running time this Theorem is used to show:

Corollary 1. *MILP(\bar{T}) has a feasible solution where for each machine type $\mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)$ of the corresponding integer variables are non-zero.*

We get the better running time by guessing the non-zero variables and removing all the others from the MILP. The number of possibilities of choosing $\mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)$ elements out of a set of $2^{\mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)}$ elements can be bounded by $2^{\mathcal{O}(1/\varepsilon^2 \log^4 1/\varepsilon)}$. Considering all the machine types we can bound the number of guesses by $2^{\mathcal{O}(K/\varepsilon^2 \log^4 1/\varepsilon)}$. The running time of the algorithm by Lenstra and Kannan with $\mathcal{O}(K/\varepsilon \log^2 1/\varepsilon)$ integer variables can be bounded by $2^{\mathcal{O}(K/\varepsilon \log K/\varepsilon \log^2 1/\varepsilon)} \text{poly}(|I|)$. This yields a running time of $2^{\mathcal{O}(K \log(K) 1/\varepsilon^2 \log^4 1/\varepsilon)} \text{poly}(|I|)$.

In the following we first proof Corollary 1 and then introduce the technique from [16] to further reduce the running time.

Proof of Corollary 1. We consider the so called *configuration ILP* for scheduling on identical machines. Let m' be a given number of machines, P be a set of processing times with multiplicities $k_p \in \mathbb{Z}_{>0}$ for each $p \in P$ and let $\mathcal{C} \subseteq \mathbb{Z}_{>0}^P$ be some finite set of configurations for P . The configuration ILP for m' , P , k , and \mathcal{C} is given by:

$$\sum_{C \in \mathcal{C}} C_p y_C = k_p \quad \forall p \in P \tag{5}$$

$$\sum_{C \in \mathcal{C}} y_C = m' \tag{6}$$

$$y_C \in \mathbb{Z}_{\geq 0} \quad \forall C \in \mathcal{C} \tag{7}$$

The default case that we will consider most of the time is that \mathcal{C} is given by a target makespan T that upper bounds the size of the configurations.

Lets assume we had a feasible solution $(\tilde{z}_{C,t}, \tilde{x}_{j,t})$ for MILP(\bar{T}). For $t \in [K]$ and $p \in B_t$ we set $\tilde{k}_{t,p} = \sum_{C \in \mathcal{C}_t(\bar{T})} C_p \tilde{z}_{C,t}$. We fix a machine type t . By setting $y_C = \tilde{z}_{C,t}$ we get a feasible solution for the configuration ILP given by m_t , B_t , \tilde{k}_t and $\mathcal{C}_t(\bar{T})$. Theorem 4 can be used to show the existence of a solution for the ILP with only a few non-zero variables: Let X be the set of column vectors corresponding to the left hand side of the ILP and b be the vector corresponding to the right hand side. Then $b \in \text{int-cone}(X)$ holds and Theorem 4 yields that there is a subset \tilde{X} of X with cardinality at most $2(|B_t|+1) \log(4(|B_t|+1)1/\varepsilon^2) \in \mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)$ and $b \in \text{int-cone}(\tilde{X})$. Therefore there is a solution (\check{y}_C) for the ILP with $\mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)$ many non-zero variables. If we set $\check{z}_{C,t} = \check{y}_C$ and $\check{x}_{j,t} = \tilde{x}_{j,t}$ and perform corresponding steps for each machine type, we get a solution $(\check{z}_{C,t}, \check{x}_{j,t})$ that obviously satisfies constraints (1),(2) and (3) of MILP(\bar{T}). The last constraint is also satisfied, because the number of covered big jobs of each size does not change and therefore the overall size of the configurations does not change either for each machine type. This completes the proof of Corollary 1.

Further Improvement of the Running Time. The main ingredient of the technique by Jansen et al. [16] is a result about the configuration ILP, for the case that there is a target makespan T' upper bounding the configuration sizes. Let $\mathcal{C}(T')$ be the set of configurations with size at most T' . We need some further notation. The *support* of any vector of numbers v is the set of indices with non-zero entries, i.e., $\text{supp}(v) = \{i \mid v_i \neq 0\}$. A configuration is called *simple*, if the size of its support is at most $\log(T' + 1)$, and *complex* otherwise. The set of complex configurations from $\mathcal{C}(T')$ is denoted by $\mathcal{C}^c(T')$.

Theorem 5. *Let the configuration ILP for m', P, k , and $\mathcal{C}(T')$ have a feasible solution and let both the makespan T' and the processing times from P be integral. Then there is a solution (y_C) for the ILP that satisfies the following conditions:*

1. $|\text{supp}(y|_{\mathcal{C}^c(T')})| \leq 2(|P| + 1) \log(4(|P| + 1)T')$ and $y_C \leq 1$ for $C \in \mathcal{C}^c(T')$.
2. $|\text{supp}(y)| \leq 4(|P| + 1) \log(4(|P| + 1)T')$.

We will call such a solution *thin*. Furthermore they argue:

Remark 1. The total number of simple configurations is in $2^{\mathcal{O}(\log^2(T') + \log^2(|P|))}$.

The better running time can be achieved by determining configurations that are equivalent to the complex configurations (via guessing and dynamic programming), guessing the support of the simple configurations, and solving the MILP with few integral variables. The approach is a direct adaptation of the one in [16] for our case. We now explain the additional steps of the modified algorithm in more detail and analyze its running time.

We have to ensure that the makespan and the processing times are integral and that the makespan is small. After the geometric rounding step we scale the makespan and the processing times, such that $T = 1/\varepsilon^3$ and $\bar{T} = (1 + \varepsilon)/\varepsilon^3$ holds and the processing times have the form $(1 + \varepsilon)^x \varepsilon^2 T = (1 + \varepsilon)^x / \varepsilon$. Next we apply a second rounding step for the big processing times, setting $\check{p}_{t,j} = \lceil \bar{p}_{t,j} \rceil$ for $\bar{p}_{t,j} \in B_t$ and denote the set of these processing times by \check{B}_t . Obviously we have $|\check{B}_t| \leq |B_t| \leq 1/\varepsilon \log(1/\varepsilon^2) + 1$. We denote the corresponding instance by \check{I} . Since for a schedule with makespan T for instance I there are at most $1/\varepsilon^2$ big jobs on any machine, we get:

Lemma 4. *If there is a schedule with makespan at most T for I , the same schedule has makespan at most $(1 + 2\varepsilon)T$ for instance \check{I} and any schedule for instance \check{I} can be turned into a schedule for I without increase in the makespan.*

We set $\check{T} = (1 + 2\varepsilon)T$ and for each machine type t we consider the set of configurations $\mathcal{C}_t(\lfloor \check{T} \rfloor)$ for \check{B}_t with size at most $\lfloor \check{T} \rfloor$. Rounding down \check{T} ensures integrality and causes no problems, because all big processing times are integral. Furthermore let $\mathcal{C}_t^c(\lfloor \check{T} \rfloor)$ and $\mathcal{C}_t^s(\lfloor \check{T} \rfloor)$ be the subsets of complex and simple configurations. Due to Remark 1 we have:

$$|\mathcal{C}_t^s(\lfloor \check{T} \rfloor)| \in 2^{\mathcal{O}(\log^2 \lfloor \check{T} \rfloor + \log^2 |\check{B}_t|)} = 2^{\mathcal{O}(\log^2 1/\varepsilon)} \tag{8}$$

Due to Theorem 5 (using the same considerations concerning configuration ILPs like in the last paragraph) we get that there is a solution $(\check{z}_C, \check{x}_{j,t})$ for MILP(\check{T}) (adjusted to this case) that uses for each machine type t at most $4(|\check{B}_t|+1) \log(4(|\check{B}_t|+1)|\check{T}|) \in \mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)$ many configurations from $\mathcal{C}_t(|\check{T}|)$. Moreover at most $2(|\check{B}_t|+1) \log(4(|\check{B}_t|+1)|\check{T}|) \in \mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)$ complex configurations are used and each of them is used only once. Since each configuration corresponds to at most $1/\varepsilon^2$ jobs, there are at most $\mathcal{O}(1/\varepsilon^3 \log^2 1/\varepsilon)$ many jobs for each type corresponding to complex configurations. Hence, we can determine the number of complex configurations m_t^c for machine type t along with the number of jobs $k_{t,p}^c$ with processing time p that are covered by a complex configuration for each $p \in \check{B}_t$ in $2^{\mathcal{O}(K/\varepsilon \log^2 1/\varepsilon)}$ steps via guessing. Now we can use a dynamic program to determine configurations (with multiplicities) that are equivalent to the complex configurations in the sense that their size is bounded by $|\check{T}|$, their summed up number is m_t^c and they cover exactly $k_{t,p}^c$ jobs with processing time p . The dynamic program iterates through $[m_t^c]$ determining \check{B}_t -indexed vectors y of non-negative integers with $y_p \leq k_{t,p}^c$. A vector y computed at step i encodes that y_p jobs of size p can be covered by i configurations from $\mathcal{C}_t(|\check{T}|)$. We denote the set of configurations the program computes with \check{C}_t and the multiplicities with \check{z}_C for $C \in \check{C}_t$. It is easy to see that the running time of such a program can be bounded by $\mathcal{O}(m_t^c (\prod_{p \in \check{B}_t} (k_{t,p}^c + 1))^2)$. Using $m_t^c \in \mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)$ and $k_{t,p}^c \in \mathcal{O}(1/\varepsilon^3 \log^2 1/\varepsilon)$ this yields a running time of $K^2 \mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)$, when considering all the machine types.

Having determined configurations that are equivalent to the complex configurations, we may just guess the simple configurations. For each machine type, there are at most $2^{\mathcal{O}(\log^2 1/\varepsilon)}$ simple configurations and the number of configurations we need is bounded by $\mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)$. Therefore the number of needed guesses is bounded by $2^{\mathcal{O}(K/\varepsilon \log^4 1/\varepsilon)}$. Now we can solve a modified version of MILP(\check{T}) in which z_C is fixed to \check{z}_C for $C \in \check{C}_t$ and only variables $z_{C'}$ corresponding to the guessed simple configurations are used. The running time for the algorithm by Lenstra and Kannan can again be bounded by $2^{\mathcal{O}(K \log K^{1/\varepsilon} \log^3 1/\varepsilon)} \text{poly}(|I|)$. Thus we get an overall running time of $2^{\mathcal{O}(K \log K^{1/\varepsilon} \log^4 1/\varepsilon)} \text{poly}(|I|)$. Considering the two cases $2^{\mathcal{O}(K \log K^{1/\varepsilon} \log^4 1/\varepsilon)} < \text{poly}(|I|)$ and $2^{\mathcal{O}(K \log K^{1/\varepsilon} \log^4 1/\varepsilon)} \geq \text{poly}(|I|)$ yields the claimed running time of $2^{\mathcal{O}(K \log(K)^{1/\varepsilon} \log^4 1/\varepsilon)} + \text{poly}(|I|)$ completing the proof of the part of Theorem 1 concerning unrelated scheduling.

4 The Santa Claus Problem

Adapting the result for unrelated scheduling we achieve an EPTAS for the Santa Claus problem. It is based on the basic EPTAS together with the second running time improvement. In the following we briefly discuss the needed adjustments.

The dual approximation method can be applied in this case as well. However, since we have no approximation algorithm with a constant rate, the binary search is more expensive. For the simplification of the input it has to be taken into

account that their may be huge jobs that are bigger than the optimal makespan, but otherwise it can be done similarly.

Moreover, in the Santa Claus problem it makes sense to use configurations of size bigger than \check{T} . Let $P = \lfloor \check{T} \rfloor + \max\{\check{p}_{t,j} \mid t \in [K], j \in \check{B}_t\}$. It suffices to consider configurations with size at most P and for each machine type t we denote the corresponding set of configurations by $\mathcal{C}_t(P)$. Again we can bound $\mathcal{C}_t(P)$ by $2^{\mathcal{O}(1/\varepsilon \log^2 1/\varepsilon)}$. The MILP has integral variables $z_{C,t}$ for each such configuration and fractional ones like before. The constraints (1) and (2) are adapted changing only the set of configurations and for constraint (3) additionally in this case the left-hand side has to be at least as big as the right hand side. The last constraint (4) has to be changed more. For this we partition $\mathcal{C}_t(P)$ into the set $\hat{\mathcal{C}}_t(P)$ of big configurations with size bigger than $\lfloor \check{T} \rfloor$ and the set $\check{\mathcal{C}}_t(P)$ of small configurations with size at most $\lfloor \check{T} \rfloor$. The changed constraint has the following form:

$$\sum_{C \in \hat{\mathcal{C}}_t(P)} \text{size}(C)z_{C,t} + \sum_{p \in S_t} p \sum_{j \in J_t(p)} x_{j,t} \geq (m_t - \sum_{C \in \check{\mathcal{C}}_t(P)} z_{C,t})\check{T} \quad \forall t \in [K] \quad (9)$$

To solve the MILP with the claimed running time, some additional non-trivial considerations are needed that are omitted in this version of the paper.

Lastly, for the rounding of the MILP the flow network has to be changed as well, using lower and upper bounds for the flow.

Acknowledgements. We thank Florian Mai and Jannis Mell for helpful discussions on the problem.

References

1. Asadpour, A., Saberi, A.: An approximation algorithm for max-min fair allocation of indivisible goods. *SIAM Journal on Computing* 39(7), 2970–2989 (2010)
2. Bateni, M., Charikar, M., Guruswami, V.: Maxmin allocation via degree lower-bounded arborescences. In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. pp. 543–552. ACM (2009)
3. Bezáková, I., Dani, V.: Allocating indivisible goods. *ACM SIGecom Exchanges* 5(3), 11–18 (2005)
4. Bleuse, R., Kedad-Sidhoum, S., Monna, F., Mounié, G., Trystram, D.: Scheduling independent tasks on multi-cores with gpu accelerators. *Concurrency and Computation: Practice and Experience* 27(6), 1625–1638 (2015)
5. Bonifaci, V., Wiese, A.: Scheduling unrelated machines of few different types. arXiv preprint arXiv:1205.0974 (2012)
6. Chakrabarty, D., Chuzhoy, J., Khanna, S.: On allocating goods to maximize fairness. In: *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on*. pp. 107–116. IEEE (2009)
7. Chen, L., Jansen, K., Zhang, G.: On the optimality of approximation schemes for the classical scheduling problem. In: *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. pp. 657–668. SIAM (2014)
8. Chen, L., Ye, D., Zhang, G.: An improved lower bound for rank four scheduling. *Operations Research Letters* 42(5), 348–350 (2014)

9. Eisenbrand, F., Shmonin, G.: Carathéodory bounds for integer cones. *Operations Research Letters* 34(5), 564–568 (2006)
10. Gehrke, J.C., Jansen, K., Kraft, S.E., Schikowski, J.: A ptas for scheduling unrelated machines of few different types. In: *International Conference on Current Trends in Theory and Practice of Informatics*. pp. 290–301. Springer (2016)
11. Goemans, M.X., Rothvoß, T.: Polynomiality for bin packing with a constant number of item types. In: *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. pp. 830–839. Society for Industrial and Applied Mathematics (2014)
12. Hochbaum, D.S., Shmoys, D.B.: Using dual approximation algorithms for scheduling problems theoretical and practical results. *Journal of the ACM (JACM)* 34(1), 144–162 (1987)
13. Horowitz, E., Sahni, S.: Exact and approximate algorithms for scheduling nonidentical processors. *Journal of the ACM (JACM)* 23(2), 317–327 (1976)
14. Imreh, C.: Scheduling problems on two sets of identical machines. *Computing* 70(4), 277–294 (2003)
15. Jansen, K.: An EPTAS for scheduling jobs on uniform processors: using an milp relaxation with a constant number of integral variables. *SIAM Journal on Discrete Mathematics* 24(2), 457–485 (2010)
16. Jansen, K., Klein, K., Verschae, J.: Closing the gap for makespan scheduling via sparsification techniques. In: *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11–15, 2016, Rome, Italy*. pp. 72:1–72:13 (2016)
17. Jansen, K., Maack, M.: An EPTAS for scheduling on unrelated machines of few different types. arXiv preprint arXiv:1701.03263v1 (2017), <https://arxiv.org/abs/1701.03263>
18. Kannan, R.: Minkowski’s convex body theorem and integer programming. *Mathematics of operations research* 12(3), 415–440 (1987)
19. Knop, D., Koutecký, M.: Scheduling meets n-fold integer programming. arXiv preprint arXiv:1603.02611 (2016)
20. Lenstra, J.K., Shmoys, D.B., Tardos, É.: Approximation algorithms for scheduling unrelated parallel machines. *Mathematical programming* 46(1-3), 259–271 (1990)
21. Lenstra Jr, H.W.: Integer programming with a fixed number of variables. *Mathematics of operations research* 8(4), 538–548 (1983)
22. Raravi, G., Nélis, V.: A ptas for assigning sporadic tasks on two-type heterogeneous multiprocessors. In: *Real-Time Systems Symposium (RTSS), 2012 IEEE 33rd*. pp. 117–126. IEEE (2012)
23. Williamson, D.P., Shmoys, D.B.: *The design of approximation algorithms*. Cambridge university press (2011)
24. Woeginger, G.J.: When does a dynamic programming formulation guarantee the existence of a fully polynomial time approximation scheme (FPTAS)? *INFORMS Journal on Computing* 12(1), 57–74 (2000)