To appear in the International Journal of Computer Integrated Manufacturing
Vol. 00, No. 00, Month 20XX, 1–18

REALSEACH PAPER

Setting production capacities for production agents making selfish routing decisions

H. Blunck\textsuperscript{a}*, D. Armbruster\textsuperscript{b} and J. Bendul\textsuperscript{a}

\textsuperscript{a}Jacobs University Bremen, School of Mathematics & Logistics; \textsuperscript{b}Arizona State University, School of Mathematical and Statistical Sciences

(October 2016)

Traditionally, the capacity dimensioning step within the manufacturing system design process takes a known and fixed distribution of production flow across path alternatives to derive capacity demand based on a desired target utilization level. Setting target levels for machine utilization provides limits on throughput-times and allows to provide capacity buffers against changes in the production mix. In this contribution, we transfer this rationale into the world of Industry 4.0, where Cyber-Physical Systems can make autonomous and selfish routing decisions. Under this new framework, non-cooperative agents make decisions based on the capacity allocation and the decisions of all other agents, thus creating a feedback between flow and capacity distribution that makes existing methods for capacity dimensioning inapplicable. We use methods and insight from algorithmic game theory and operations research to investigate the capacity dimensioning process in this context. We proof properties of the throughput-time optimal allocation of production capacity under fixed target utilization for important queue classes. Our findings not only provide a quantitative, easy to operationalize tool for production system designers, but explore a trade-off between cost and flexibility that arises naturally in this regime.

Keywords: Capacity Dimensioning; Cyber-Physical Systems; Industry 4.0; Algorithmic Game Theory

1. Introduction & Problem Motivation

The digitalization of production is currently investigated under different terms in research efforts around the world (c.f. European Commission 2016; Blanchet and Rinn 2016). They envision to deploy Cyber-Physical Systems (CPS), physical objects, equipped with sensing, computation and communication capacities (Lee 2008) to manufacturing settings to herald a “fourth industrial revolution”, the rise of “Industry 4.0” (Lasi et al. 2014). It is today commonly believed, that future conceptions of production have to be build on the loose federation of autonomous production entities (Monostori, Váncaza, and Kumara 2006), which constitute the production system behavior in an emergent manner through their collaboration and interaction. It has long been understood, that designing such heterarchically controlled production systems requires a new approach towards many aspects of production system design and management (Scherer 1998; Cavalieri et al. 2000; Dilts, Boyd, and Whorms 1991; Rogers and Brennan 1997; Verstraete et al. 2008; Van Dyke Parunak 1999; Trentesaux 2009).

In this contribution we tackle one particular problem of the manufacturing system design process — capacity distribution and dimensioning — and investigate it in the context of intelligent, selfish acting products.

*Corresponding author. Email: h.blunck@jacobs-university.de
1.1. Capacity Dimensioning — The Classical Approach

In classical factory planning approaches (e.g., consider the “0+5+X Planning Model” by Schenk, Wirth, and Müller (2010)), capacity dimensioning and distribution is embedded into a sequence of activities, that translate the expected production program into capacity demand for individual functions and then subsequently calculates the amount of capacity, \(a_e\), necessary to meet such demand. Crucially, two assumptions are made: (1) The distribution of production flows across the production path alternatives is fixed (given an expected production volume \(\lambda\)), leading to known flow requirements, \(f_e\), for each machine (or edge) and (2) capacities are dimensioned based on expected demand and (likewise, a priori established) notion of desired utilization rates \(u_e^*\). Fig. 1 depicts this understanding of capacity dimensioning.

![Traditional Capacity Dimensioning Approach](image)

**Traditional Capacity Dimensioning Approach (Single Commodity)**

(1) Nodes are buffers, edges are machines.
(2) Define possible production paths.
   *Here: solid and dashed path over 4 different machines combined.*
(3) Distribute flow demand among (subset of) paths.
   *Here: Total flow \(\lambda = 1\), path flows \(\lambda_1 = \frac{1}{2}\), \(\lambda_2 = \frac{1}{2}\), \(\lambda_3 = \frac{1}{2}\), \(\lambda_4 = \frac{1}{2}\).*
(4) Calculate capacity demands per machine/edge \((f_1 - f_4)\).
(5) Set resource capacities \((a_1 - a_4)\), given \(u_e^*\), i.e. \(a_i = \frac{\lambda_i}{u_e^*}\).
   *Here: \(u_e^* = 0.8\) for all edges.*

Figure 1. Traditional approach to capacity dimensioning. Notably, capacities are set, after flow distribution is known, so this traditional approach to capacity dimensioning does not work, when flow distribution is decided at run time!

1.1.1. Fixed Flow Distribution Assumption

The classic approach to determine the amount of required capacity by dividing the capacity demand (the product of deterministic or expected processing time demands) by the total working time per period can be traced back to the 1950s (c.f. Miller and Davis 1977, for an early review). Subsequently, optimization models were devised to calculate the optimal number of machines considering e.g. capacity, floor space, cost, etc. (Kusiak 1987, provides a review).

Processing path alternatives and the desirability of flexibility within the manufacturing system were brought into the discussion with the rise of Flexible Manufacturing Systems (FMS). Research in this domain focussed on pre-determining the (cost-)optimal flow distribution across process and machine alternatives, leaving the decision between alternatives explicitly not open until runtime: Optimization models that seek to find the optimal capacities and flow distributions in conjunction are provided e.g. by Lee, Srinivasan, and Yano (1991), Bard and Feo (1991), Tetzlaff (1995), and Askin (2013). Other analytical support for the design of FMS is usually provided through simulation and queueing models (reviewed by Solot and van Vliet 1994) that also allow the prediction of waiting times and inventory levels and hence a more concise picture of the dynamic behavior.

1.1.2. Target Utilization Assumption

It has long been acknowledged and subject of scholarly investigation that there exist a trade-off between capacity distribution and other performance indicators (in particular: WIP and throughput time) (c.f. e.g. Bitran and Tirupati 1989; Bitran and Morabito 1999; Negri da Silva and Morabito 2009). Machine utilization, the ratio of actual workload and system capacity (Schönslében 2012,
Ch. 1.2.4), is the key gauge by which this trade-off can be navigated.

While the optimization approaches discussed above usually only constrain the resulting utilization to be less than 1 (c.f. Lee, Srinivasan, and Yano 1991), in a continuous approximation, where any capacity value is feasible, we have the flexibility to set production capacities (and hence expected utilization levels) more freely.

It is widely assumed that the theoretically available capacity should exceed the capacity demand. Schönsleben (2012, Ch. 13.1.1) calls this deliberate deviation from the ideal of 100% utilization “tactical underutilization”, meant to ensure high performance across all performance indicators (including those competing with high utilization). The reasons given in literature to maintain such “buffer” between the technically required capacity and the installed capacity include:

- Capacity buffer may be required to compensate setup times, machine unavailability, sick times, etc. (Schenk, Wirth, and Müller 2010, Ch. 3.2.3).
- Limiting \( u \) gives a upper bound on the throughput time. Better yet, fixing \( u \) gives a predictable throughput time.
- Since also the first derivative of the clearing function generally increases in the utilization rate \( u \), limiting \( u \) means limiting the impact of deviations in utilization on throughput time.
- Capacity buffer provide protection against expected fluctuations or long-term changes in capacity demand. Where overload leads to resource failure, adding extra capacity can prevent cascading effects upon the failure of single nodes (discussed in the context of complex networks e.g. by Motter and Lai 2002; Zhao, Park, and Lai 2004).

One may conceptualize this intentional installation of excess capacity beyond the calculated demand by assuming a “target utilization” value \( u^* < 1 \) for every machine and installing excess capacity such that the ratio of expected capacity demand and technical capacity equals \( u^* \). According to the Total Productive Maintenance (TPM) literature, an effectiveness (OEE) value of 85% would establish a world class operation (Dal, Tugwell, and Greatbanks 2000). This does not prevent the (planned) existence of bottleneck machines. Where certain machines are meant to be bottlenecks, given their large relative investment cost or for production control purposes (c.f. Goldratt 1999), we may simply assume a different, higher target utilization level for this machine.

1.2. The (Un)predictability of Cyber-Physical System Behavior

Existing approaches for the distribution of capacity fundamentally rely on the knowledge of (expected) production flows along process alternatives and eventually machines. When intelligent products make routing decisions based on real-time information however, such a priori information is not available, since (1) routing and other decisions are postponed to runtime and (2) routing often depends on the system state: Tay and Ho (2008) observe that in distributed production control systems, routing decisions are often based on the queue length of the available machine alternatives, expressing the individual product’s intention to optimize its throughput time. With the queue length being directly influenced by the available capacity, we induce a feedback between capacity distribution and flow distribution, that have so far been considered separate and allocated sequentially. Fig. 2 visualizes this new situation.

While in the literature on autonomous, intelligent products, the interaction between intelligent products and machines has been addressed and (short-term) capacity adjustments in response to unplanned events have been discussed (c.f. e.g. Jeken et al. 2012), the authors are not aware of literature that discusses the interplay of capacity dimensioning and routing decisions by intelligent products.

Understanding and shaping emergent system properties (such as flow distributions) in next generation manufacturing systems is a worthwhile endeavor: Unexpected and unpredictable performance metrics of production systems under autonomous control have been identified as a key hindrance for the adoption of the latter in industry (Trentesaux 2009; Mařík and McFarlane 2005;
Mařík and Lažanský 2007). The lack of understanding for the relationship between system design decisions and emergent properties today requires excessive “fine-tuning” and difficult “debugging” during a “bottom-up” development of manufacturing systems (Van Brussel et al. 1998; Monostori et al. 2015).

This paper investigates, how the assumptions of classical capacity dimensioning and distribution can be transferred to situations with autonomous, intelligent products, providing the designer of such novel production systems with the predictability and bounds on performance metrics that can boost the industrial adoption of these ideas.

1.3. Paper Outline

The remainder of this paper is structured as follows: Sec. 2 introduces congestion games as a suitable minimal model of flows of intelligent agents in networks of limited capacity and reviews the relevant literature from that domain. Sec. 3 introduces the specific model notation used in this paper. In Sec. 4 we briefly review our previous work on the problem (Blunck, Armbruster, and Bendul 2016) and its limitations. In Sections 5 and 6 we extend this previous work by introducing an analytical approach to both the simultaneous and successive distribution of capacities and flows. It is shown in Sec. 5 that a simultaneous flow and capacity allocation leads to stable distributions, yet highly inflexible manufacturing systems. Sec. 6 discusses the possibility to trade-off cost (for work in progress) against process flexibility to address this issue.

2. Modeling Production Flows as Congestion Games on Graphs

2.1. Review of Concepts

Game theory can be used to analytically investigate coordination problems among selfish agents. From Ossowski (1999, p. 23) we learn that “a solution to a coordination problem constitutes an equilibrium, a compromise that assures somehow ‘maximal’ attainment of the different interests of all involved individuals”. In particular a Nash equilibrium describes an equilibrium in non-cooperative games by assigning a strategy mix to each player such that no player has an incentive to deviate from its strategy, since it is maximizing its expected profit (or minimizing its cost).

Congestion Games (Rosenthal 1973) are a particular type of game, where every player has to choose a combination of facilities. The payoff for each player is a function of the number of players who choose this strategy, meaning that players experience congestion whenever they concentrate on certain facilities. Rosenthal (1973) already suggested to use congestion games to model the behavior of road users and the interplay of supply and demand of production capacity (Holzman and Law-Yone 1997; Meyers and Schulz 2012). In particular, we may equate strategies with path choices in a network and attain a network congestion game (c.f. Meyers and Schulz 2012). By
studying these games, we can bring the advantages of game theory to the study of interactions between players and queuing networks.

Under the fluid assumption (modeling flows of products instead of individual products) we can, instead of focussing on individual users, consider demands for certain “origin-destination pairs” (road-users with the same origin and destination, products of the same type, . . . ) as players (Haurie and Marcotte 1985), giving each a strategy space that allows to split the total flow arbitrarily among path alternatives. Under this mental model, the Nash equilibrium is equivalent to the so-called Wardrop equilibrium (Haurie and Marcotte 1985; Altman et al. 2006), which describes a situation among nonatomic users that is described in “Wardrop’s first principle”:

“The journey times in all routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route.” — Wardrop (1952, p. 345)

It is important to note that Nash Equilibria constitute a compromise between selfish agents. They do not represent a socially desirable solution as they can be inefficient (Dubey 1986). A commonly applied measure for the efficiency of a flow distribution is its induced “social cost”, the combined latency accrued across all flows and obtained by multiplying the path latencies with the path flows. Applied to the domain of production systems, the social cost are a measure of the total throughput time and the work in process (WIP) that accumulates in the system. The ratio between the social cost of the Nash equilibrium and the socially optimal solution is commonly called the “price of anarchy” (Papadimitriou 2001; Koutsoupias and Papadimitriou 2009).

Beckmann, McGuire, and Winsten (1955) showed more than 60 years ago that for non-atomic flows (i.e. the demand of any player can be infinitesimally subdivided) both, the Nash equilibrium and the socially optimal flow distribution given a set of edge cost functions, could, under the modest assumption of cost function convexity, be computed through simple convex optimization problems (c.f. Sec. 5.1). With the the rise of algorithmic game theory (Nisan and Ronen 1999) came increased interest in the ability to not only predict the outcome of games but to understand the result as a function of the game parameters and ultimately to be able to engineer and influence it. In particular for congestion games, Roughgarden (2005, Ch. 3) showed upper bounds for the price of anarchy in congestion games for many popular queue classes\(^1\). His work inspired subsequent investigations on upper bounds on the price of anarchy for more general cost functions (Czumaj, Krysta, and Vöcking 2010), finite games (Christodoulou and Koutsoupias 2005), and situations where the maximum flow is limited (Correa, Schulz, and Stier-Moses 2004). In parallel, network design for selfish agents has seen game-theoretic treatment: Korilis, Lazar, and Orda (1995, 1997, 1999) study strategies to distribute (additional) capacity across network edges with the goal of reducing the social cost of the Nash equilibrium. Roughgarden (2001) showed that determining a subset of graph edges that optimize the Nash equilibrium social cost is hard.

### 2.2. Modeling Assumptions

To develop an appropriate model to understand the process of capacity dimensioning in the environment of industry 4.0 and therefor for autonomous agents, we make the following assumptions:

**No product dependent processing time.** This is a common assumption for queuing models (Papadopoulos and Heavey 1996).

**Throughput times** (or “Latency”) functions \(c_e(f_e)\) are generally considered to be non-decreasing in flow \(f_e\), a requirement typically held by all relevant clearing function classes (c.f. Armbruster and Uzsoy 2012). In particular, they are considered to be semiconvex, i.e. \(c_e(f_e) \cdot f_e\) is convex.

---

1Linear, general polynomial, and M/M/1.
Since we express the behavior of intelligent products as Nash equilibria in games, the usual assumptions on Nash equilibria hold. In particular, these include:

**Perfect information** is assumed for agents when considering Nash equilibria. This is at odds with the common perception of distributed production control that usually considers agents with limited information horizon (c.f. e.g. Trentesaux (2009); Leitão (2009), and Schneeweiss (2003, Ch. 1.1)). However, approaches such as stigmergy can lead to the build-up of global information through locally available pheromones (c.f. Peeters et al. 2001; Van Dyke Parunak 1997; Armbruster et al. 2006). Thus the assumption of visibility of the current state (queue length) across path alternatives for an intelligent product is reasonable even within the context of distributed control. Note that we do not assume any sort of scheduling or coordination between agents.

**Non-atomic flows** are assumed in Wardrop’s principle. While this certainly does not represent actual production flows, it is a valid simplification for strategic decision making, inherently present in the reviewed literature on capacity distribution. In addition, fluid models have shown to be useful also in the context of distribution control research (c.f. e.g. Scholz-Reiter, Jagalski, and Bendul 2008; Jeken et al. 2012; Armbruster et al. 2006).

**Unused paths are possible** in the Nash equilibrium, since Wardrop’s condition only applies to paths with flow. This complicates the solution process, since it includes the identification of path subsets to be used, which may cause prohibitive computational cost (Powell and Sheffi 1982).

### 3. Model

#### 3.1. Notation

We use the existing notation for transport systems (based here on Roughgarden 2005; Raith et al. 2014) also used in Fig. 1.

- Let \((G, r, c)\) be a transportation problem, where \(G\) is a directed graph \(G = (V, E)\), \(c\) are cost (latency) functions given for each edge \(e \in E\) and \(r\) is a set of source-sink pairs of vertices \(\in V\). We will also call a source-sink pair a “commodity”.
- Let \(P\) be the set of all paths between a source and a sink in \(G\). In particular, let \(P_i\) be the set of paths that connect the \(i\)th source-sink pair (decision alternatives for the \(i\)th commodity).
- Let \(\rho = |P|\) be the number of paths.
- Let \(m = |E|\) be the number of all machines (edges).
- Then \(\Phi \in \mathbb{N}^{\rho \times P}\) (where \(\mathbb{N}\) is assumed to contain also 0) is the flow-path incidence matrix, where elements \(\phi_{i,j}\) indicates the number of times path \(j\) visits machine \(i\).
- Let \(\lambda \in \mathbb{R}^\rho\) be the flow distribution among paths. Then \(\lambda_P\) is a scalar, indicating the flow on path alternative \(P \in P\); \(\lambda_P \geq 0\).
- The required flow (demand) for commodity \(i\) is given by \(r_i\), i.e. \(\sum_{P \in P_i} \lambda_P = r_i \forall i\).
- Let \(f \in \mathbb{R}^m\) be the the edge flow distribution that results from path flow distribution \(\lambda\). We can calculate \(f\) from path flows as \(f = \Phi \cdot \lambda\). Then \(f_e\) is a scalar, indicating flow on edge \(e\). Given the constraints on \(\lambda\) and \(\Phi\), \(f_e \geq 0 \forall e\).

For every edge \(e \in E\) we assume a semi-convex latency (or: cost) function \(c_e(f_e)\), parametrized by some non-negative value \(a_e\), called the capacity of the machine.

We refer to the ratio \(f_e/a_e = u_e\) as the utilization of the machine. When intending to find a flow distribution that attains a certain capacity utilization per machine, we will indicate the desired utilization level by \(u^*_e\) (or \(u^*\) when no distinction between edges is made).
Given a vector $c_e$ of edge latency functions, we can calculate the latency of a path alternative as the sum of the constituting edges (Raith et al. 2014), resulting in a path-latency vector $c_P$,

$$c_P = \Phi^T \cdot c_e. \quad (1)$$

We are particularly interested in flow distributions that we call “utilization-attaining”:

**Definition 1** (Utilization-attaining flow distribution). A flow distribution $f$ is called utilization-attaining if the utilization of all machines with positive flow is equal to some constant $u^*_e (u^*_e \in (0,1) \forall e | f_e > 0)$.

### 3.2. Cost Function Classes

Two latency function classes are of particular interest:

- **Affine Latency Functions**, $c_e(f_e) = f_e/a_e + b_e$, are a simplified cost function class that has received most attention (c.f. Roughgarden 2005, Sec. 3.7 for a review).
- **M/M/1 Latency Function**, $c_e(f_e) = 1/(a_e f_e)$, represent a well understood cost function class for a queueing system with exponential arrival and exponential departure times that is closely related to more general cost function classes such as $M/G/1$ and $G/G/1$ (Kingman 1961).

We limit ourselves to these cost function classes, disregarding others, for which selfish routing has been considered (c.f. Roughgarden 2003), because of (1) their distinct role in modeling manufacturing systems and (2) their direct mathematical representation of machine capacity (and hence utilization).

### 4. Previous Work: Iterative Capacity Adjustment

In Blunck, Armbruster, and Bendul (2016), we suggest an iterative approach that alternates between the re-assignment of capacities (given flows) and flows (given capacities). The Nash equilibrium for the flow distribution given a capacity distribution is approximated using the “Method of Successive Averages” (MSA) (Sheffi and Powell 1981). The approach is tested on abstracted, real production networks, assuming exponentially distributed inter-arrival and processing times. We demonstrate that for all investigated networks, the approach converges to a capacity and flow distribution with utilization levels within $\pm 2\%$ of $u^*$. Fig. 4 visualizes the evolution of utilization levels for one such network.

There are however problems with such an iterative approach: The solution depends on the starting point of the algorithm, which is why it is usually recommended to compare multiple runs, using different initial flow distributions (Queiroz and Humes 2003). While convergence for all our examples is reached within an acceptable number of iterations (usually $\leq 10$) and for all networks, we can neither guarantee convergence or bound the convergence speed, nor can we prove any other particular properties of the solution.

In this contribution, we chart a different approach to the same problem to obtain more concrete insights.

### 5. Parallel Assignment of Flows and Capacities

We begin by assigning flows and capacities simultaneously. We will show for both cost function classes discussed in section 3.2 that utilization-attaining Nash equilibria exist that exhibit low social cost but also low flexibility, in that only a small subset of available path alternatives is used.
The parallel assignment of flows and capacities assumes that demand, changing its routing decisions "takes the capacity" with it, such that on both the old and new path, the target utilization levels are attained. This is clearly beyond any reasonable expectation about the power of intelligent products in real manufacturing systems, but is a valid modeling trick in the design stage, to ensure utilization attainment. The usefulness of this assumption rests on the following two Lemmas:

**Lemma 1** (Equilibrium Persistency). A utilization attaining Nash equilibrium obtained through the simultaneous assignment of flows and capacities will maintain its path and social cost, utilization levels, and Nash equilibrium property, under a relaxation of the utilization constraint, allowing variation of flows and keeping the capacities at the equilibrium levels.

*Proof.* This is the immediate consequence of the experimental setup: We calculate flows and corresponding capacities such that, given the capacity values, the resulting Nash flow will be a utilization-attaining Nash equilibrium.

**Lemma 2** (Price of Predictability). A flow and capacity distribution that generates minimum social cost under the utilization constraint may not remain the social cost minimum under a variation of flows keeping the capacities at the equilibrium levels.

### 5.1. Calculating Nash- and Socially Optimal Flow Distributions through Convex Optimization

For given fixed capacities, i.e. fixed cost functions both, the socially optimal flow distribution and the Nash equilibrium flow distribution, can be calculated through simple, convex optimization problems. The social cost for a given flow distribution can be computed as the sum of the social cost for each machine (edge). The socially optimal flow distribution, the flow distribution that incurs the lowest social cost, is then (given known and fixed capacities) the optimal solution to the following optimization problem (c.f. Beckmann, McGuire, and Winsten (1955) and Roughgarden (2005, Sec. 2.4)):

![Figure 3](image-url). Attained utilization levels during iterative capacity allocation process. With the sixth reiteration, all utilization levels are within the ±2% of the target utilization level (80%).
Cost Type | Expression | Affine | M/M/1
--- | --- | --- | ---
Latency | \( c_e(f_e) = u_e^* + b_e \) | \( = \frac{u_e^*}{1-u_e^*} \cdot f_e^{-1} \cdot \text{sgn}(f_e) \) | \( = \text{constant} \)
Social Cost | \( c_e(f_e) \cdot f_e \) | \( = (u_e^* + b_e) \cdot f_e \) | \( = \frac{u_e^*}{1-u_e^*} \cdot \text{sgn}(f_e) \)

Table 1. Overview of the functional form of the latency and social cost functions with implemented capacity constraint. \( \text{sgn}(x) \) indicates the sign function, returning 1 for strictly positive values of \( x \) and 0 for \( x = 0 \) (by definition).

5.2. Inserting the Target-Utilization Constraint

Given a known target utilization rate for every edge \( (u_e^*) \), we can express the capacity of an edge \( (a_e) \) with positive flow \( (f_e > 0) \) purely as a linear function of the flow,

\[
a_e = \frac{1}{u_e^*} \cdot f_e, \quad \forall \left\{ e \in E \mid f_e > 0 \right\}.
\]

For the remainder of this paper, we will assume that in the case where \( f_e = 0 \), i.e. when no capacity demand is present for a machine, this machine will not be purchased and the latency and social cost associated with this edge will be 0.

Inserting the target utilization constraint (Eq. 3) into the latency and social cost functions discussed in Sec. 3.2 yields new expressions for them as a function of the flow and the target utilization rate. The results are given in Table 1.
5.3. Conclusions for Affine Cost Functions

Table 1 shows a constant latency under the assumption that capacity and flow are assigned simultaneously for affine cost functions leading to the following theorem.

**Theorem 1** (Shortest paths yield a utilization-attaining Nash equilibrium). Given a set of source-sink pairs with associated flow requirements, known target utilizations $u^*_e$ and affine cost function parameters $u^*_e + b_e$ for all edges $e \in E$, any flow distribution among the shortest weighted paths for each commodity is a Nash equilibrium.

**Proof.** To be a Nash equilibrium, the flow distribution has to exhibit two properties (Wardrop 1952):

1. the latency of all paths with flow has to be equal,
2. any path without flow has to have a higher latency than that of used paths.

Both properties obviously hold for the set of shortest weighted paths which have to be of equal length (otherwise, one would be shorter) and shorter than any alternative path for that commodity (otherwise, it would be a shortest path). Since the latency term is constant, the length of the paths does not change as we apply more flow.

There always exists at least one flow/capacity-distribution that is a utilization-attaining Nash equilibrium. If multiple path alternatives have the same length, all possible flow/capacity distributions among them have the same latency and are Nash equilibria. Additionally, we can also state that:

**Theorem 2.** The social cost for affine cost functions for any flow distribution over shortest weighted paths are the same and are the minimum value.

**Proof.** The resulting social cost term is linear in the edge flow $f_e$. This means that the additional social cost induced by adding flow on a certain path (a certain set of edges) is independent of the flow already associated with these edges and a linear function of solely the additional flow $\Delta f_e$. Hence, regardless of previous allocation decisions, additional flow (e.g. flow of an additional source-sink pair) is best allocated to the path alternatives with the smallest sum of slopes $\sum_{P \subseteq P; e \in P} (u^*_e + b_e)$ which is just the a-priori known “weight” measure of path length and hence equal across all shortest weighted paths and optimal in the sense that there is no path alternative for that commodity that would yield lower social cost when given the same amount of additional flow.

Together, Theorems 1 and 2 show that, using the set of shortest weighted paths, we can always find a set of paths for each commodity such that any flow distribution among these paths is a Nash equilibrium (if capacity is set accordingly) and that any such flow and capacity distribution has the same and lowest possible social cost across all utilization-attaining flow distributions.

It is worthwhile to compare the utilization constraint case with similar findings for the non-utilization constrained case for affine cost functions: The price of anarchy is always 1 for every affine set of cost functions in the former case. In the latter the price of anarchy is 1 if and only if, the cost functions are linear, i.e. $b_e = 0 \forall e$ (Roughgarden 2005, Lemma 3.2.2). Assuming a constraint on the desired level of utilization, allows us to show a similar result for a broader cost function class.

5.4. Conclusions for $M/M/1$ Cost Functions

For $M/M/1$ cost functions Table 3 shows constant social cost when flows and capacities are assigned simultaneously.
Theorem 3 (Minimal set of machines minimizes social cost). We call a machine (edge) active if the flow through the machine is nonzero. Then the lowest social cost for a utilization-attaining flow distribution in a network of $M/M/1$ cost functions is attained by flows that use the minimum weighted number of active machines (weighted by $\frac{u^*_e}{1-u^*_e}$).

Proof. Follows immediately from the considerations made above.

Above mentioned minimal set can be obtained with the following Mixed-Integer Linear Program (MILP) with a total of $|P| + |E|$ binary variables. Let

- $B$ be the binary path incidence matrix, where matrix element $\phi_{e,P}$ states whether edge $e$ is in path $P$ (at all) ($b_{i,j} = 1 \iff \phi_{i,j} > 0$, otherwise: 0). We will refer to column vectors as $B(P)$.
- $x_e$ be the binary decision variable, indicating whether or not to choose machine $e$, composing together the decision vector $x$.
- $y_j$ be an auxiliary binary variable, indicating if all machines belonging to path $j$ are part of the selected subset.
- $M$ be a large scalar (“Big M Method”).

The optimization problem to be solved is the following

$$\min \sum_{e \in E} \frac{u^*_e}{1 - u^*_e} \cdot x_e$$

subject to

$$\sum_{P \in P} y_P \geq 1 \quad \forall i,$$  

$$-(1-y_P) \cdot M \leq x \cdot B(P) - \sum_{e \in E} b_{e,P} \quad \forall P \in P,$$

$$x_e \in \{0,1\} \quad \forall e \in E,$$

$$y_P \in \{0,1\} \quad \forall P \in P.$$  

The optimization problems seeks to minimize the social cost across all “selected machines” ($\sum_{e \in E} \frac{u^*_e}{1 - u^*_e} \cdot x_e$), such that at least one path alternative is viable for each commodity (Eq. (4b)). Eq. (4c) ties the availability of a path ($y_P = 1$) to the availability of all machines on that path ($x_e = 1 \forall e \in P$). This is done, by comparing on the right hand side of Eq. (4c) the number of all machines necessary for a given path ($\sum_{e \in E} b_{e,P}$) with the the number of elements in the union of machines required for $P$ and available in vector $x$ (calculated as the scalar product of the two binary vectors $x$ and $B(P)$).

The related path subset that can be maintained with the minimal weighted machine set is then given by the indicator variable $y_P$.

Note that, since the $y_P$ in Eq. (4) are tied to the values of $x$, the number of truly independent, binary decision variables increases linearly with the number of edges (machines) in the system, which is generally much slower then the increase in path alternatives resulting from additional machines (which may be multiplicative in the number of machines).

The potential function, required to calculate the Nash equilibrium is given as

$$h_e(f_e) = \int_0^{f_e} \frac{u^*_e}{1 - u^*_e} \cdot t^{-1} dt$$  

(5)
which has a singularity of $1/t$ at $t = 0$. To approximate the problem, we introduce a small constant $\epsilon$ and essentially re-normalize the integral:

$$h_e(f_e) = \int_0^{f_e} \frac{u_e^*}{1 - u_e^*} \cdot (t + \epsilon)^{-1} \, dt$$  \hspace{1cm} (6a)$$

$$= \frac{u_e^*}{1 - u_e^*} \cdot \ln \left( \frac{f_e + \epsilon}{\epsilon} \right)$$  \hspace{1cm} (6b)$$

$$= \frac{u_e^*}{1 - u_e^*} \cdot \ln (f_e + \epsilon) - \ln (\epsilon) \cdot \frac{u_e^*}{1 - u_e^*}$$  \hspace{1cm} (6c)$$

where we may ignore the constant term for optimization purposes. This optimization problem will yield an optimal flow distribution that uses the minimal possible set of machines, since the decrease in the target function for a given edge $\frac{u_e^*}{1 - u_e^*} \cdot \ln (f_e + \epsilon)$ when shifting some flow $\theta$ from one edge to another is inverse proportional to $f_e$. In particular

$$h_e(f_e + \theta) - h_e(f_e) \propto \ln (f_e + \epsilon + \theta) - \ln (f + \epsilon)$$  \hspace{1cm} (7a)$$

$$= \ln \left( 1 + \frac{\theta}{f_e + \epsilon} \right)$$  \hspace{1cm} (7b)$$

which decreases as $f_e$ increases. Thus the concave minimization problem to find the Nash equilibrium will choose the same weighted minimal set of machines as does the MILP in Eq. (4).

**Theorem 4** (Nash equilibrium over minimal machine set). *The utilization-attaining Nash flow distribution is given by the socially optimal, utilization-attaining flow distribution, i.e. by the path alternatives that minimize the set of machines with nonzero flow.*

**Proof.** Follows, as discussed above, from the nature of the target function in Eq. (6) and the social-cost properties of the minimal weighted set approach in Theorem 3. \hfill $\square$

### 5.5. Summary

We showed that for affine and $M/M/1$ cost functions the price of anarchy for the utilization attaining flow distributions is always one. In addition, we showed that for affine cost functions the flow distributions corresponding to Nash equilibria and social optima may be a continuum, i.e. a parametrized set of distributions. In contrast, at least for single commodity problems, for the $M/M/1$ flows the flow over active machines is always maximal, since a deviation from maximal flow is penalized by $f_e^{-1}$ for the case of Nash equilibria and by a constant additive term for the social costs.

In the next section, we investigate how to trade-off between flexibility and social cost, i.e. if and how we may define utilization-attaining flow distributions that utilize a larger set of path alternatives, possibly at the ex pense of higher social cost. We refer to this increase in social cost as the *price of flexibility*.

### 6. The Price of Flexibility: Can we increase flexibility using utilization-attaining flow distributions?

#### 6.1. Affine Cost Functions

Theorem 1 has the following Corollary:
Corollary 1 (No utilization-attaining Nash equilibrium on non-shortest paths or paths of different length). There exists no utilization-attaining Nash equilibrium on networks with affine cost functions that gives strictly positive flow to path alternatives of different weighted length.

Proof. Follows immediately from the proof of Theorem 1. The first condition of Wardrop’s principle (equal latency of all paths with positive flow) cannot be met when paths of different weighted length are used.

The only opportunity to increase the number of path alternatives in a utilization-attaining Nash flow distribution is to adjust the parameters \( u^*_e \) and \( b_e \) (previously assumed fixed) such that multiple path alternatives per commodity have the same weighted length.

The social cost (i.e. the WIP) of such flow distribution is a function only of the achieved “length” of the shortest path. If we adjust edge lengths by increasing target utilization and fixed throughput time components leading to an increase in WIP, we increase the length of the shortest path alternatives (this leads to an increase in WIP) and thus have a positive “price of flexibility”. Alternatively, if we adjust target utilization and fixed throughput time downwards for some edges, e.g. by investing into more powerful machines or accepting lower target utilization levels, we will incur higher investment cost. However this does not change the latency of the original Nash flow distribution, thus again leading to additional cost for flexibility.

6.2. The Case for M/M/1 Cost Functions

We observe that utilization-attaining flow-distributions with more than the minimal set of machines defined in Sec. 5.4 are possible: As the simplest example, consider a single-commodity flow networks with \( n \) parallel worksystems. Setting equal target-utilization levels and capacities at all machines, a flow-distribution that equally uses all alternatives will have the same latency across all path alternatives and hence also be a Nash-flow. Given same flows and \( u^*_e \forall e \), the flow is also utilization-attaining and has \( n \) times the social cost (WIP) as compared to a single machine with identical \( u^*_e \) (c.f. Table 1). This is similar to the result by Korilis, Lazar, and Orda (1997).

Beyond such simple examples, we are not able to provide guarantees for the nature or existence of utilization-attaining Nash equilibria. As with affine cost functions, changing the target utilization levels (and hence the constant term per \( u^*_e \) edge) can give the designer additional degrees of freedom to attain such flow distributions.

With this additional degrees of freedom, the designer may wish to design a utilization-attaining flow distribution that is as close as possible to some desired flow distribution \( \lambda \), which (through \( \Phi \)) also yields a given edge flow distribution \( f_e \). For such a reasonable case we can express the edge latency as a function of \( u_e \), finding that \( c_e(f_e) \propto \frac{u^*_e}{1-u^*_e} \) for a given edge flow \( f_e \). Also note that, given \( \lambda \), we can subdivide the set of path alternatives \( P \) into two mutually exclusive sets \( P_{\text{in}} \) and \( P_{\text{out}} \), depending on whether the designer wishes a certain path to be part of the flow distribution or not. The latency of a M/M/1 worksystem, given utilization level \( u_e \) and flow \( f_e \) then is \( \frac{1}{1-u_e} \cdot f_e^{-1} \cdot d(u) \cdot f_e^{-1} \), where \( d(u) = \frac{1}{1-u} \). The problem is hence linear in \( d(u) \) and we can find a vector \( \Delta \) such that \( \lambda \) is a utilization attaining-flow distribution for the adjusted target utilizations. In particular, we can minimize the largest absolute adjustment to \( d(u^*) \) necessary through the following linear program:
\[
\begin{align*}
\min_{A_{e,i}} & \quad h \\
\text{subject to} & \quad h \geq \Delta_e \quad \forall e \in E \quad \text{(8b)} \\
& \quad h \geq -\Delta_e \quad \forall e \in E \quad \text{(8c)} \\
& \quad l_i - \sum_{e \in P} \Phi_{e,P} \cdot \frac{d_e(u^*_e)}{f_e} \Delta_e = \sum_{e \in P} \Phi_{e,P} \cdot \frac{d_e(u^*_e)}{f_e} \quad \forall i, P \in \mathcal{P}_{\text{in}} \quad \text{(8d)} \\
& \quad \sum_{e \in P} \Phi_{e,P} \cdot \frac{d_e(u^*_e) + \Delta_e}{f_e} \geq l_i \quad \forall i, P \in \mathcal{P}_{\text{out}} \quad \text{(8e)}
\end{align*}
\]

with
\[
d_e(u) = \frac{u}{1 - u}
\]
\[
l_i = \text{Latency for commodity } i \quad \forall i
\]

Eq. (8d) enforces that all “desired” paths (where \(\lambda_P > 0\)) have the same latency \(l_i\) per commodity, whereas Eq. (8e) ensures that unused paths have higher latency. Together, the constraints ensure that \(\lambda\) is a utilization-attaining Nash equilibrium under the adjusted target-utilization values. Note that, given the non-linearity of \(d_e(u)\), minimizing \(h\) is not identical to minimizing the absolute deviation from \(u^*_e\). Assuming similar values of \(u^*_e\) across all \(e\), this may be ignored. In case of larger differences, we may substitute \(\Delta_e\) by the linear approximation of \(d_e(u)\) at \(u^*_e\).

7. Conclusion & Further Research

This contribution has discussed the possibility to transfer widely held assumptions during the manufacturing system design process towards the new regime of intelligent, selfish-acting products, as a hallmark of Industry 4.0 and other new paradigms of production.

Using models from game-theory we were able to investigate the simultaneous distribution of flow and capacity analytically and show that for both affine and M/M/1 cost function classes, utilization-attaining and socially optimal Nash flows can be found, albeit with low levels of flexibility. We have discussed opportunities to purposely increase the flexibility of the flow distributions, seeing a trade-off between flexibility and cost of inventory (WIP) as well as equipment.

Future research should address the price of flexibility in more detail. In addition, other forms of agent-influencing, e.g. through tolls (c.f. e.g. Gibbens and Kelly 1999), or stigmergy (Van Dyke Parunak 1997; Armbruster et al. 2006) in the context of manufacturing system design should be analyzed.

We have not discussed the “price of predictability” (Corollary 2) and can say little about it, except that it is, by the nature of the problem, subject to the same upper bounds for the price of anarchy that others have shown (c.f. Sec. 2).

Acknowledgements

The authors acknowledge Mirja Meyer for her input on capacity dimensioning literature in manufacturing system design.

References

Methods and Applications

Game Theory: Numerical Methods and Applications.


