THROUGHPUT MAXIMIZATION BY IMPROVED BOTTLENECK CONTROL

Elvira Marie B. Aske [*,**] **Sigurd Skogestad** [*,1]
Stig Strand [**]

[*] *Department of Chemical Engineering, Norwegian University of Science and Technology, N-7491 Trondheim, Norway*
[**] *Statoil R&D, Process Control, N-7005 Trondheim, Norway*

Abstract:
In many cases, optimal operation for a plant is the same as maximum throughput. In this case a rigorous model for the plant is not necessary if we are able to identify the bottleneck. Optimal operation is the same as maximum throughput in the bottleneck. If the bottleneck does not move, this can be realized with single-loop controller from the throughput manipulator to the bottleneck. However, if the bottleneck moves, single-loop control would require reassignment of loops which is undesirable. A better approach is then to use a multivariable coordinator controller since input and output constrains are directly included in the problem formulation.

Keywords: throughput, multivariable control, networks

1. INTRODUCTION

When disturbances of economic importance have a dynamic character and especially if they occur frequently compared to the controlled plant responses, steady state real-time optimization (RTO) will be inadequate to follow the optimal operation point in periods. However, in many cases the prices and market conditions are such that real-time optimization of the plant is the same as maximizing plant throughput. From network theory, the *max-flow min-cut* theorem states that the maximum throughput in a plant (network) is limited by the "bottleneck" of the network. In order to maximize the throughput, the flow through the bottleneck should be at its maximum flow. In particular, if the actual flow at the bottleneck is not at its maximum at any given time, then this gives a loss in production which can never be recovered (sometimes referred to as a "lost opportunity"). Maximize throughput in a network is a common problem in several settings (Phillips *et al.* 1976, Ahuja *et al.* 1993). In the special but important case of a linear network, optimal operation is the same as maintaining maximum flow through the bottleneck(s). A detailed nonlinear network model is not necessary in this simple case because the objective is to identify the active "bottleneck" constraint and implement maximum throughput at the bottleneck. If the bottleneck is fixed, a single-loop controller manipulating the throughput can be used (Skogestad 2004*a*). If the bottleneck(s) moves, a coordinator MPC (Aske *et al.* 2006) is suitable because of its ability to handle constraints in its process inputs and outputs and because the

[1] Corresponding author: e-mail: skoge@chemeng.ntnu.no, Tel: +47-73594154, Fax: +47-73594080

local MPCs can be used to estimate the available capacity in each unit.

2. BACKGROUND

2.1 Inventory control

Inventory control deals with how the mass balance is maintained in the plant. A chemical plant has usually a single "throughput manipulator" (TPM) which indirectly through the process and product requirements determines all the feed and product rates. The main exception where we have more than one TPM, is if there are several parallel trains from feed to product.

Definition 1. (Price and Georgakis 1993, Price *et al.* 1994). Throughput manipulator (TPM). The TPM is the degree of freedom used to set the throughput in the primary process path (from the major feed streams to the major products). Systems with parallel trains from feed to product have one TPM for each train.

Price and Georgakis (1993) describe two kinds of TPMs:

- Explicit TPM - a flow on a primary path (from feed to product)
- Implicit TPM - not a flow on an primary path, (e.g. a heat duty may be a TPM with flow set by a temperature controller)

There are three basic schemes for inventory control (see Figure 1), depending on where in the process the TPM is located (Buckley 1964, Price *et al.* 1994):

- *Scheme 1.* Feed as TPM (given feed): Inventory control system in the direction of flow (conventional approach)
- *Scheme 2.* Product as TPM ("on demand"): Inventory control system opposite to flow
- *Scheme 3.* TPM inside plant: Radiating inventory control

The selection of throughput manipulator is important, because the chain of level controls need to be constructed to radiate outward from the throughput manipulator to obtain self-consistency (Price and Georgakis 1993), which means that the flow is maintained through the plant by use of the local inventory loops only.

2.2 Modes of optimal operation

Most process plants have two main modes in terms of optimal operation:

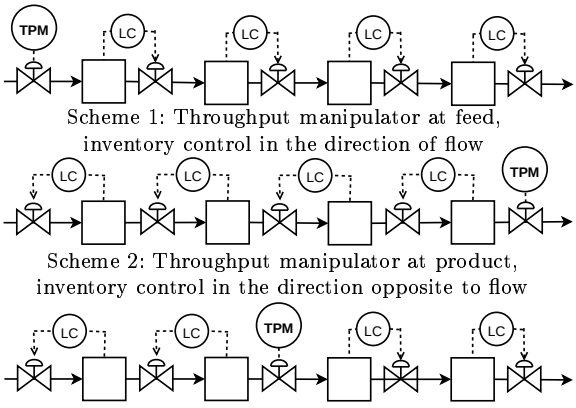

Fig. 1. Basic schemes for inventory control

Mode 1. Given throughput. The objective is then maximum efficiency, that is, minimize utility (energy) consumption for the given throughput. This mode of operation typically occurs when the feed rate is given (or limited) or the product rate is given (or limited, for example, by market conditions).

Mode 2. Feed is available and throughput is a degree of freedom. We have here two cases:

a Maximum throughput. The maximum throughput is the largest given throughput that achieves feasible operation, i.e. satisfies Equation 2.
b Optimized throughput. This mode of operation occurs when feed is available, but where the increase in production cost becomes large so that is not optimal to go all the way to maximum throughput.

This paper focuses on mode 2a, which is optimal when feed is available and the product prices are high.

Mathematically, optimal operation in all three cases is to minimize the cost J (maximize the profit $-J$), subject to satisfying given specifications and model equations ($f = 0$) and operational constraints ($g \leq 0$):

$$\min_u J(x,u,d) \tag{1}$$

$$\text{s. t. } f(x,u,d) = 0 \tag{2}$$

$$g(x,u,d) \leq 0 \tag{3}$$

Here u are the manipulated variables (including the feed rates), d the disturbances and x the (dependent) state variables. A typical profit function is

$$-J = \sum_i p_{P_i} \cdot Pi - \sum_i p_{F_i} \cdot F_i - \sum_i p_{Q_i} \cdot Q_i \tag{4}$$

where P_i are products, F_i are feeds, Q_i are utilities (heating, cooling, power), and p indicates the price for each of the element.

- In mode 1, the feed rates F_i are given and the optimization problem is modified by adding a

set of constraint, $F_i = F_{i0}$ (alternatively, the product rates or flow inside the plant could be given).

- In mode 2a (maximum throughput), the feed rates F_i are degrees of freedom, and the cost data are such that we have an constrained optimum with respect to the feed rates (i.e. $dJ/dF_i < 0$). Increasing F_i above its maximum (optimal) value gives infeasible operation.
- In mode 2b (optimized throughput), the feed rates F_i are degrees of freedom, and the cost data are such that we have a unconstrained optimum with respect to the feed rates (i.e. $dJ/dF_i = 0$). Increasing F_i above its optimal value is feasible, but gives a higher cost J.

In terms of location of the TPM, scheme 1 (in Figure 1) is the natural choice for mode 1 with given feed, scheme 2 is the natural choice for mode 1 with given product, whereas scheme 3 is the best choice for modes 2a and 2b where the optimal throughput is determined by some conditions internally in the plant.

2.3 Maximum Throughput (mode 2a)

In mode 2a the objective is to find a feasible solution with maximum throughput, and since the maximum throughput is independent of cost data, we can simplify the cost function J in Problem (1). In the general case with multiple (independent) feeds, the throughput may be defined as the sum of the weighted external feeds, and we have $F_w = \sum_i w_i F_i$, where F_i denotes external ("fresh") feeds. The maximum throughput is then the solution to the problem

$$\begin{aligned} \max_u \ & F_w \\ \text{s.t. } & f = 0 \\ & g \leq 0 \end{aligned} \tag{5}$$

Remark 1: For the case with a single feed this may be written on the form $J = -F$ in (1) and we note that $dJ/dF = -1$ (also at the optimum). *Remark 2:* For multiple feeds, the simplest case is were all the feeds have equal weight, $w_i = 1$. *Remark 3:* More generally, w_i should express the relative value of processing the various feeds, but this value may be difficult to find. Thus, to find the maximum throughput for the case with multiple feeds, it may be better to use the economic cost function in Equation (4).

3. BOTTLENECK

We consider here maximum throughput (mode 2a), which in practice is achieved by maximizing the flow through the bottleneck.

Definition 2. Maximum flow for a unit. The maximum flow (capacity) of a unit is the maximum feed rate F_i^u that the unit can accept subject to achieving feasible operation. Mathematically, this corresponds to solving the maximum flow problem stated in (5) for a given unit i, that is, to find the maximum value of F_i^u that satisfies the constraints $f_i = 0$ and $g_i \leq 0$ for the unit.

Definition 3. Bottleneck (operation). A unit is a bottleneck if maximum throughput (maximum network flow for the system) is obtained by operating this unit at maximum flow (with no available capacity left). In some cases the bottleneck can not be located to a specific unit, but rather to a system of units ("system bottleneck").

Definition 4. Bottleneck constraints (operation). The active constraints at maximum flow in a bottleneck unit are called the bottleneck constraints. If one of the active constraints is a flow on the primary path, then this is called a direct bottleneck constraints and the corresponding flow is called a direct bottleneck manipulator.

Definition 5. Back off. Back off is the deviation between the actual value and the constraints (optimal value) needed to avoid infeasibility dynamically in the presence of disturbances (Govatsmark and Skogestad 2005, Narraway *et al.* 1991, Narraway and Perkins 1993).

These concepts are closely related to the problem of maximum flow in networks considered in the operations research community, (e.g. Phillips *et al.* (1976)). Such a network consists of sources, arcs, nodes and sinks. An arc is like a pipeline with given (maximum) capacity, and the nodes may be used to add or split streams. The main restriction is that the flow must satisfy conservation at the nodes. This may be written as a linear programming problem, and the trivial but important solution is that the maximum flow is dictated by the network bottleneck. To see this, one introduces "cuts" through the network, and the capacity of a cut is the sum of the capacity of the forward arcs that it cuts through. The *max-flow min-cut theorem* (Ford and Fulkerson 1962) says that the maximum flow through the network is equal to the minimum capacity of all cuts (the minimal cut). We then reach the important insight that maximum network flow (maximum throughput) requires that all arcs in some cut have maximum flow, that is, they must all be bottlenecks (with no available capacity left).

In terms of process engineering systems, a unit with a single product is an arc, and flow splits and flow junctions are nodes. In network theory, the flow splits in nodes are free variables,

like crossovers between parallel trains in "our" processes. A unit with several products (e.g. a distillation column) is a combination of an arc and a node, but there is usually a limited degree of freedom to adjust the split because of product constraints. To get a linear network the split factor must either be constant or a free variable.

To apply network theory to process engineering systems, we first need to obtain the capacity (maximum flow) of each unit (arc). This is quite straightforward, and involves solving a (nonlinear) feasibility problem for each unit (see Definition 2). The capacity may also be computed on-line, for example, by using local MPC implementations as proposed in the next section.

Assumption: The mass flow through the network is represented by a set of units (where each unit capacity is obtained locally) with linear flow connections.

Note that the nonlinearity of the equations within a unit is not a problem, but rather the possible nonlinearity in terms of flows between units. The main problem of applying linear network theory to process engineering systems is therefore that the flow split in a unit, e.g. a distillation column, is not constant, but depends on the state of its feed, and, in particular, of its feed composition. The main process unit to change composition is a reactor, so decisions in the reactor may strongly influence the flow in downstream units and recycles. Another important decision that affects composition, and thus flows, is the amount of recycle. One solution to avoid these sources of nonlinearity is to treat certain combinations of units, like a reactor-recycle system, as a single combined unit as seen from maximum throughput (bottleneck) point of view.

In summary, we derive from the max-flow min-cut theorem the following useful insights (rules) about the maximum flow solution for a linear network which satisfies the above assumptions:

Rule 1. At maximum throughput the network must have at least one bottleneck unit.

Rule 2. Additional independent feeds and flows splits ("independent" means that they are not indirectly determined by other flows in the process, e.g. a crossover flow between processing trains) may give rise to additional bottlenecks, and the idea of "minimal cut" may be used to identify the location of the corresponding bottleneck units.

Rule 3. Focus on the bottleneck unit(s). To maximize throughput, the flow through the bottleneck should be as close as possible to its maximum at any given time. This requires "tight" control of the bottleneck unit, as any deviation from optimal operation in the bottleneck unit due to poor control (including any deviation or back off from the bottleneck constraints) implies a loss in throughput (which can never be recovered).

Rule 4. Use TPM for control of the bottleneck unit. This follows because TPM is the degree of freedom for throughput which according to Rule 3 should be maximized at the bottleneck. In practice, TPM is often used to control one of the bottleneck constraints (see Definition 4).

Further refinements of Rules 3 and 4 are given by Rules 5 and 6.

Rule 5. TPM should be located so that controllability of the bottleneck unit is good (Skogestad 2004*a*). This is to reduce the throughput loss due to imperfect control. For example, if TPM is used to control one of the bottleneck constraints then the effective time delay from TPM to its bottleneck constraint should be small. Selecting TPM as a direct bottleneck manipulator (if there is one; see Definition 4) is a good choice as it directly maximizes the flow through the bottleneck.

Rule 6. Bottleneck unit: focus on tight control on the variable with the most costly back off in terms of loss in throughput. This follows because back off is needed on the constraint variables in the presence of disturbances.

The ideas of linear network theory may be very useful for "our" systems. Although the linearity assumptions will not hold exactly in most of "our" systems, the bottleneck result is nevertheless likely to be optimal in most cases.

4. ESTIMATION OF BACK OFF (LOSS)

The back off is the distance to the active constraint needed to obtain feasible operation due to unmeasured disturbances, model errors, delay and other sources for imperfect control. Back off results in a loss and should be as small as possible Narraway and Perkins (1993). The back off is a "safety factor" and should be obtained based on information about the disturbances and the expected control performance. We here assume that the control of the active constraints y (for which we want to estimate the back off) is based on feedback control. For the linearized system the transfer function from a disturbance d to y is

$$y = (I + GK)^{-1} \cdot G_d d = SG_d d \qquad (6)$$

where G is the process model, K is the controller, S is the sensitivity function and G_d is the disturbance model. Assume that the disturbances are sinusoidal $d(t) = d_0 \sin(\omega t)$ and that $\|d_0\|_2$

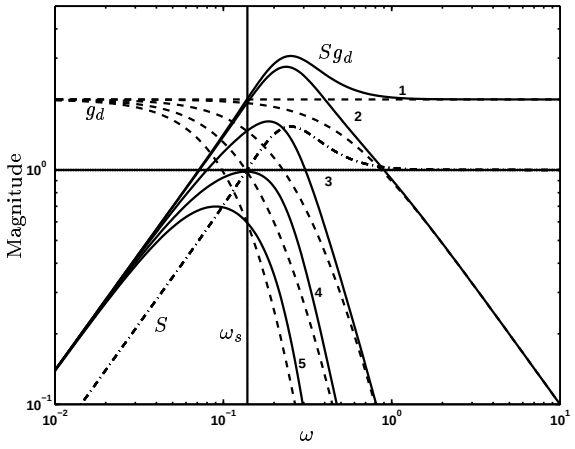

Fig. 3. $|Sg_d|$ (solid), $|g_d|$ (dashed) and $|S|$ (dash-dotted) for disturbances $1-5$ entering the process, see Figure 2

is bounded. The worst case amplification from d to y is then given by

$$\text{Back off} = \max \|y\|_2 = \|SG_d\|_\infty \cdot \|d_0\|_2 \qquad (7)$$

The $\max_{\omega,d} |y|$ represents the effect of the worst case disturbance over all frequencies and directions and therefore represents the minimum back off.

Let us consider the scalar case. Then

$$\|SG_d\|_\infty = \max_w |SG_d(j\omega)| \geq |SG_d(j\omega_0)| \qquad (8)$$

where ω_0 is any specific frequency.

Normally, the worst case frequency is around the closed loop bandwidth and let us consider two particular frequencies, ω_s and $\omega_{resonance}$ This gives

$$\text{Back off} \geq \begin{cases} \|G_d(j\omega_s)\|_\infty \cdot \|d\|_2 \\ M_S \cdot \|G_d(j\omega_{resonance})\|_\infty \cdot \|d\|_2 \end{cases} \qquad (9)$$

where by definition $|S(j\omega_{resonance})| = M_S$ and $|S(j\omega_s)| = 1$. Typically $M_S \leq 2$ and robust tunings make M_S smaller.

Consider a process string similar to Figure 2 where we want to maximize throughput. The plant has a bottleneck and the TPM is placed inside the plant and disturbances entering the process.

Assume that the disturbances $1-5$ enters as displayed in Figure 2. The G_d is of first order and the feedback controller K is tuned by using Skogestad's tuning rules (SIMC) (Skogestad 2004*b*).

From Figure 3 we see that the peak of $|Sg_d|$ occurs at different frequencies, dependent on the distance from the bottleneck. For disturbances that enters the process near the TPM (2 and 3), ω_s is a good approximation of the peak frequency. However, for disturbances that enters the process at a distance from the TPM (1,4 and 5), the ω_s is not a good estimate of the peak frequency.

5. REALIZE MAXIMUM THROUGHPUT

In terms of realizing maximum throughput there are two problems:

(1) Identify the bottleneck(s)
(2) Implement maximum flow at the bottleneck

In the simplest case, the bottleneck is fixed and we can use single-loop control (Skogestad 2004*a*).

If the bottleneck moves in the plant, then single-loop control requires reassignment of loops, both the TPM to control the bottleneck unit (Rule 5), and the inventory loops to ensure self-consistency in the plant In addition, the moving bottleneck needs to be identified. A better approach is to use a multivariable controller were input and output constraints are included directly in the problem formulation (e.g. MPC).

A case study using this method is described in Aske *et al.* (2006). The conventional inventory control (scheme 1 in Figure 1) is considered in the case study, using feeds, feed splits and crossover are used to maximize the plant throughput.

Also when using MPC, there may be a "long" loop between the bottleneck and the TPM (feeds). Therefore is back off needed on the constraint variables in the presence of unmeasured disturbances and model errors (Rule 6). There are two ways of avoiding this long loop. First, the TPM can be moved closer the the expected bottleneck unit. This requires (permanent) reassignment of the level loops. The idea is that the "mean" distance to the bottleneck unit will be smaller, which is similar to the arguments of Price and Georgakis (1993). However, if the bottleneck moves the distance to the loop may still be longer than desired. Also, moving the TPM inside the plant requires that the inventory control will be in both direction of flow and direction opposite to flow in the plant, to ensure self-consistency in the plant, and this may be undesirable, for example because it adds confusion for engineers and operators.

Another way for reducing the long loop is to use inventories (buffer tanks etc.) as dynamic degrees of freedom to maximize the flow through the bottleneck unit. Buffer volumes placed upstream the bottleneck leads to smaller back off, whether a buffer volume downstream the bottleneck has no effect. However, hold-up volumes between the TPM and the bottleneck increases the effective delay, and tight control of the bottleneck unit becomes more difficult. With large buffer volumes inside the plant and with the TPM placed at the feed, it become therefore even more important to exploit the buffer volumes between the inlet and the bottleneck to obtain tight control of the bottleneck unit. An MPC controller can manipulate on the level control set points in the

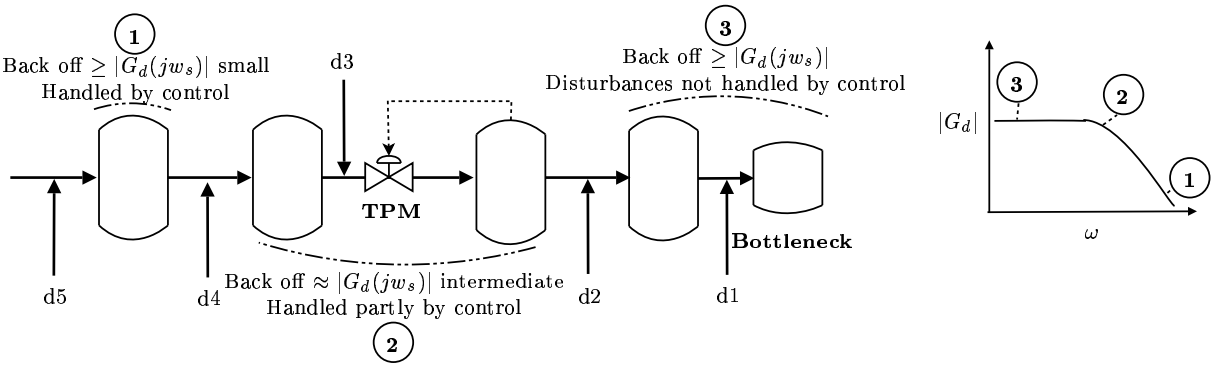

Fig. 2. Process example with disturbances

hold-up volume or directly manipulate the bias on the flow. The first approach has the disadvantage that it depends strongly on the tuning of the inventory loops. For example, if the hold-up level control is tuned to work as a buffer volume, the response from the level control set point to the flow change will be slow. The second approach where the MPC controller can manipulate directly on the level control valve depends less on the tuning of the level loop in the regulator control layer. The disadvantage that it requires the bias for the inputs is available in the MPC layer.

By using inventories, the flow rate through a bottleneck can be corrected faster due to shorter dead time and settling time in the plant, compare to using only the TPMs, which reduces the necessary back off.

6. CONCLUSION

In many cases, the optimal operation for a plant is the same as maximum throughput. The concepts are closely related to maximum flow in networks considered in operation research community. The max-flow min-cut theorem can be used in process engineering systems under the assumption of the mass flow through the network can be represented as a set of units with linear flow connections. The most important rule derived from the max-flow min-cut theorem is the TPM should be used for control of the bottleneck unit to obtain maximum flow. However, back off is needed to avoid infeasibility dynamically in the presence of disturbances. With some knowledge of where the disturbance enters and the shape of the disturbance model and the controller, frequency analysis can be used to obtain an estimate of the back off.

REFERENCES

Ahuja, Ravindra K., Thomas L. Magnanti and James B. Orlin (1993). *Network Flows*. Prentice Hall.

Aske, E.M.B, S. Strand and S Skogestad (2006). Coordinator mpc with focus on maximizing throughput. *ESCAPE+PSE, July 9-13, 2006, Garmisch-Partenkirchen, Germany.*

Buckley, Page S. (1964). *Techniques of Process Control*. John Wiley & Sons, Inc., NY, USA.

Ford, L.R. and D.R. Fulkerson (1962). *Flows in Networks*. Princeton University Press.

Govatsmark, M.S. and S. Skogestad (2005). Selection of controlled variables and robust setpoints. *Ind. Eng. Chem. Res* **44**, 2207–2217.

Narraway, L.T. and J.D. Perkins (1993). Selection of process control structure based on linear dynamic economics. *Ind. Eng. Chem. Res.* **32**(11), 2681–2692.

Narraway, L.T., J.D. Perkins and G.W. Barton (1991). Interaction between process design and process control: Economic analysis of process dynamics. *J. Proc. Cont.* **1**, 243–250.

Phillips, Don T., A Ravindran and James. J. Solberg (1976). *Operations Research Principles and Practice*. John Wiley.

Price, Randel M. and Christos Georgakis (1993). Plantwide regulatory control design procedure using a tiered framework. *Ind. Eng. Chem. Res* **32**, 2693–2705.

Price, Randel M., Philip R. Lyman and Christos Georgakis (1994). Throughput manipulation in plantwide control structures. *Ind. Eng. Chem. Res.* **33**, 1197–1207.

Skogestad, S. (2004*a*). Control structure design for complete chemical plants. *Computers & Chemical Engineering* **28**, 219–234.

Skogestad, S. (2004*b*). Simple analytic rules for model reduction and pid controller tuning. *Modeling, Identification and Control* **25**(2), 85–120.