THROUGHPUT MAXIMIZATION BY IMPROVED BOTTLENECK CONTROL

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

[*] *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

Real-time optimization (RTO) offers a direct method of maximizing an economic objective function. Typically, RTO systems are model-based, closed-loop systems whose objective is to maintain the process operation as nearly as possible to the optimum plant operation (Zhang and Forbes, 2000). Lu (2003) claims that the wide use of MPC establishes a solid foundation for large-scale optimization. Dynamic coordination among MPC controllers is a key to tight integration between advanced process control and plant wide optimization.

In many cases the prices and market conditions are such that real-time optimization of the plant is the same as maximizing plant throughput. 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) (*max-flow min-cut theorem*). 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 based on manipulating the

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

throughput can be used (Skogestad, 2004). If the bottleneck moves, a coordinator MPC (Aske *et al.*, 2006) is suitable because of its ability to handle constraints in its process inputs and outputs (Qin and Badgwell, 2003) and because the local MPCs can be used to estimate the available capacity in each unit. This paper discusses these two approaches in more detail.

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 - flow on a primary path (from feed to product)
- Implicit TPM - flow not an primary path, e.g. a heat duty

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 is that the flow is maintained through the plant by use of the inventory loops only.

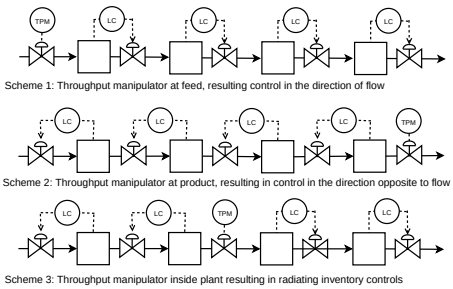

Fig. 1. Basic schemes for inventory control

2.2 Modes of optimal operation

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

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

Mode 2. Maximum throughput. The maximum throughput is the largest possible given throughput and is independent of cost data. This mode is optimal when the product prices are high and feed is available.

This paper focuses on mode 2. There is also a third, but less common mode:

Mode 3. Optimized throughput. This mode of operation occurs when feed is available (feed rate is a degree of freedom), but where the increase in production cost becomes large so that is not optimal to go all the way to maximum throughput.

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 ($g \leq 0$):

$$\begin{aligned} \min_u \; & J(x,u,d) \\ \text{s. t. } & f(x,u,d) = 0 \\ & g(x,u,d) \leq 0 \end{aligned} \tag{1}$$

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

$$-J = \sum_i p_{P_i} \cdot Pi - \sum_i p_{F_i} \cdot F_i - \sum_i p_{Q_i} \cdot Q_i \tag{2}$$

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 could be given).

- In mode 2 (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 optimal (maximum) value gives infeasible operation.
- In mode 3 (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 1a (given feed), scheme 2 is the natural choice for mode 1b (given product), whereas scheme 3 is the best choice for modes 2 and 3 where the optimal throughput is determined by some conditions internally in the plant.

2.3 Maximum Throughput (mode 2)

In mode 2 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 Equation (1). In the general case with multiple (independent) feeds, the throughput may be defined as the sum of the weighted feeds, and we have $F_w = \sum_i w_i F_i$. The maximum throughput is then the solution to the problem

$$\begin{aligned} \max_u \ & F_w \\ \text{s.t. } & f = 0 \\ & g \leq 0 \end{aligned} \tag{3}$$

Remark 1: For the case with a single feed this may be written on the form $J = -F$ in Equation (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 equally 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 (2).

3. BOTTLENECK

We consider here maximum throughput (mode 2), 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 that the unit can accept subject to achieving feasible operation. Mathematically, this corresponds to solving the maximum flow problem in Equation 3 for a given unit i, that is, to find the maximum value of F_i 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 manipulator.

Definition 5. Back off. Back off is the deviation between set point and optimal value (constraint) which is primarily introduced to avoid infeasibility dynamically in the presence of disturbances (Govatsmark and Skogestad, 2005). Typically, back off is required for output variables like pressure or temperature. The requirement of stable operation may require back off in input variable. For example, in a reactor, cooling may be a bottleneck constraint. If the cooling is used to stabilize the reactor temperature, then some back off from the maximum cooling rate is required to avoid saturation.

At maximum throughput there is one bottleneck for each TPM along the primary path. The TPM should be used to keep (control) the flow through the bottleneck as close as possible to its maximum.

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* 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 assumption:

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.

The flow should be at maximum at the bottleneck(s). This has implications for control of the bottleneck unit (Rule 3), and in particular for use of the throughput manipulator (Rules 4, 5 and 6).

Rule 3. Focus on the bottleneck unit. 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). 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 bottleneck manipulator (if there is one; see Definition 4) is a good choice as it directly maximizes the flow through the bottleneck.

Rule 6. Bottlneck 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.

Rule 7. Self-consistency of the inventory control system. For the material balance to be maintained, inventory control must be in the direction of flow downstream of TPM, and in direction opposite to flow upstream of TPM (see Figure 1).

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. 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. However in the general case model based control is proposed.

4.1 Fixed bottleneck: single-loop control

Assume that the bottleneck is always located in the same unit. In this case single-loop control is sufficient and the following discussion is based on Skogestad (2004).

As we increase the throughput we reach a point where the bottleneck reaches its maximum capacity and becomes a bottleneck for further increase in production. In addition, as we reach the bottleneck constraint, we lose a degree of freedom for control, and to compensate for this we have several options:

(1) Manual control: Reduce the throughput and "back off" from the constraint in the bottleneck unit (gives economic loss).
(2) Use the throughput manipulator as a degree of freedom for control of the (new) bottleneck constraint. For example, if the new constraint is a manipulated variable that saturates, then there are two options:
    (a) Fix the manipulated variable at its maximum and use the bottleneck manipulator to take over its "lost" control task. Problem: May get poor control due to "long" loop (long physical distance).
    (b) Keep the existing loop in place, and introduce an outer flow control loop where TPM is adjusted on a longer time scale to reset the inner loop set point. This is known as input resetting (Skogestad and Postlethwaite, 2005), mid-ranging control (Allison and Isaksson, 1998) or a valve positioning scheme (Shinskey, 1988). Problem: Large back-off in flow (and loss in throughput) if "long" loop, that is, large effective delay from TPM to bottleneck unit.

To avoid this slow ("long") loop one may either:

(3) Install a surge tank upstream of the bottleneck, and reassign its outflow to take over the lost control task, and use the throughput manipulator to reset the level of the surge tank, or
(4) Move the throughput manipulator to the bottleneck and reassign all inventory control loops between the bottleneck and the original throughput manipulator. The reassignment is to ensure self-consistency (Rule 7) and may involve many loops.

All these options are undesirable. A better solution is to *permanently* move the throughput manipulator to the bottleneck unit. The justification for this rule is that the economic benefits of increasing the production are usually very large (when the market conditions are such), so that it is important to maximize flow at the bottleneck. On the other hand, if market conditions are such that we are operating with a given feedrate or given product rate, then the economic loss imposed by using a outer cascade loop to adjust the production rate at the bottleneck (somewhere inside the plant) is usually zero, as deviations from the desired feed or production rate can be averaged out over time, provided we have storage tanks for feeds or products.

However, one should be careful when reassigning, as also other considerations may be important, such as the control of the individual units (e.g. distillation column) which may be affected by whether inflow or outflow is used for level control.

4.2 Moving bottleneck: multivariable control

If the bottleneck may move in the plant, then single-loop control requires reassignment of loops. In addition, the 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).

The bottleneck can be identified (Problem 1) by using a detailed steady-state model of the plant. However, there are two drawbacks here. First, a detailed model for the whole plant is expensive to obtained, time consuming to solve and second, the models may not be accurate.

A better approach is to use a multivariable coordinator controller (Aske *et al.*, 2006), a separate MPC installed at the plant. The local MPC on each unit is slightly extended to estimate remaining feed capacity, based on the models and constraints already available. The estimate of remaining feed capacity is solved in the steady state part of the MPC at each sample. The calculations are executed at each sample and the estimate will tend to be better when operating closer to the bottleneck, and this is what is important for maximizing throughput.

Using the estimate of the remaining capacity in each unit, the coordinator MPC can use the throughput manipulators (usually feeds) and also any internal crossovers and splits to maximize the throughput subject to keeping the units within its capacity. A case study using this method is described in Aske *et al.* (2006). In the case the conventional inventory control (scheme 1 in Figure

1) is considered and feeds, feed splits and crossover are used to maximize the plant throughput.

Also when using coordinator 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 disturbances (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 (Rule 7), 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. The coordinator can manipulate on the level control set points in the hold-up volume or directly manipulate the bias on the flow. The first approach has the disadvantage that is 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 coordinator MPC 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. This leads so a reduction in back off and then again an increase in the production.

5. 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 theorem says that maximum throughput is obtained with maximum flow in the bottleneck. If the bottleneck is fixed in the same unit, single-loop control from the bottleneck to the TPM can be used to maximize throughput. If the bottleneck move, MPC to maximize throughput is an alternative approach. In the latter case, the dynamic response can be improved by using inventories as additional dynamic degrees of freedom.

REFERENCES

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

Allison, B.J. and A.J. Isaksson (1998). Design and performance of mid-ranging controllers. *Journal of Process Control* **8**(5-6), 469–474.

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.

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

Lu, J.Z. (2003). Challenging control problems and emerging technologies in enterprise optimization. *Control Engineering Practice* **11**, 847–858.

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.

Qin, S.J. and T.A. Badgwell (2003). A survey of industrial model predictive control technology. *Control Engineering Practice* **11**, 733–764.

Shinskey, F.G. (1988). *Process Control System - Application, Design, and Tuning 3rd ed.*. McGraw-Hill Inc., New York, USA.

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

Skogestad, Sigurd and Ian Postlethwaite (2005). *Multivariable Feedback Control: Analysis and Design*. John Wiley & Sons.

Zhang, Y. and J.F. Forbes (2000). Extended design cost: a performance criterion for real-time optimization systems. *Computers and Chemical Engineering* **24**, 1829–1841.