*IIE Transactions* (1995) **27**, 339–347

# Mean–variance interactions in process improvement and capacity design

ARNOLD H. BUSS[1] and STEPHEN R. LAWRENCE[2]

[1]*Operations Research Department, Naval Postgraduate School, Monterey, CA 93943-5000, USA.*
[2]*College of Business Administration, University of Colorado, Boulder, CO 80309-0419, USA*

Received May 1991 and accepted July 1993

We investigate mean–variance interactions of processing time as applied to process improvement and capacity design. For general capacity cost and flowcost functions, we demonstrate that production processes fall into one of six regions on the mean–variance interaction plane, each with its own policy implications. The general model is specialized to the case of an M/G/1 queue with linear and separable mean and variance costs, and with flowcosts proportional to mean queue length. Optimal solutions for processing-time mean and variance are derived, and easily obtained operating parameters are used to identify appropriate process improvement policies. A simulation example of a production network taken from industry verifies the efficacy of the linear M/G/1 model in a more general setting. We conclude that intelligent management of both processing capacity (i.e. mean processing time) and processing-time variances can be powerful tools for both capacity design and process improvement.

## Introduction

In this paper we investigate mean–variance interactions of processing times as applied to capacity planning and process improvement. Making reasonable assumptions about the costs of increasing capacity (*i.e.*, mean processing time) and reducing processing variance, we apply results from queueing theory to develop solutions for both the optimal capacity and the optimal processing variance of a simple production system. In addition, we characterize circumstances for which reductions in processing variances are most beneficial and those for which increases in processing capacity are preferred.

The motivation for this research comes from anecdotal evidence that many well-managed production facilities (particularly Japanese facilities) do not operate at high levels of utilization (see, for example, Hayes, 1981; *The Economist*, 1990). Instead, capacity cushions are maintained to buffer against inherent production uncertainties, and considerable managerial and engineering attention is devoted to reducing processing variances (see, for example, Schonberger, 1986). Such policies are reported to reduce work-in-process inventories, decrease throughput times, and increase delivery reliability. The question naturally arises as to the optimality of such policies: What levels of production capacity and processing variance minimize relevant costs? The issue we address is the determination of capacity and variance levels that are economically advantageous to the steady-state operation of the facility. These capacity design issues are the focus of this paper; we do not address the shorter-term problems of controlling production.

Processing time variances are typically not considered as decision variables by Western production managers, but are assumed to be inherent in the production technology, and therefore beyond management control. Consequently, managerial attention has been focused on production capacity, and particularly on utilizing capacity as fully as possible (see standard cost accounting texts; for example, Horngren (1991)). Recent research has called into question this emphasis on capacity utilization (Banker *et al.*, 1988; Buss *et al.*, 1994; Lawrence and Buss, 1995).

There is a small but growing literature addressing process-time variability as a control variable. Sarkar and Zangwill (1991) studied the effects of arrival rate, service rate, and setup variances on inventory levels, waiting times, and cycle times in cyclic production systems. Erlebacher (1992) examined the allocation of processing-time variance among the stations of a paced assembly line with the objective of minimizing the incidence of over-cycle occurrences. Tangentially related to this paper is the work of Karmarkar (1987), and Karmarkar *et al.* (1985), which derive optimal lot sizing policies in stochastic M/G/1 and M/G/k production settings. Because lot size adjustments are one way of reducing effective processing time variance (where one lot is considered to be a single job), these papers implicitly manage variance through lot size control. Wacker (1987) used simulation analysis to compare several operating policies such as just-in-time and repetitive lots to demonstrate that reduced move-time and processing-time var-