Chapter 1

Capacity Planning and Management of IT Incident Management Services based on Queuing Models

Ta-Hsin Li
IBM T. J. Watson Research Center, USA

Juhnyoung Lee
IBM T. J. Watson Research Center, USA

ABSTRACT

Incident management or resolution services for the information technology (IT) infrastructure and software of large enterprises are labor-intensive operations. Because incidents have to be resolved in timely manner, performance targets are often set regarding the time to respond and the time to resolve for incident tickets. To meet these targets, adequate staffing is critical. At the same time, utilization rate of the staff must also be taken into account, because extra cost is often associated with an underutilized workforce. The management of IT incident resolution services always faces the question: how to properly staff a given operation, especially when the volume of service requests is expected to rise. Queuing models can be used to help address such questions. This chapter reviews the basic concepts in queuing models and discusses some practical issues in the application of queuing models to the capacity planning and management of IT incident resolution services.

INTRODUCTION

The objective of incident management or resolution for the information technology (IT) infrastructure and software of an enterprise is to restore the service of the infrastructure and software as quickly as possible in order to minimize the negative impact on business operations and ensure the best possible levels of service quality and availability (Orand, 2013). Normal service operation is often defined as
service operation within a service-level agreement (SLA). An incident is any event which is not part of the standard operation of a service and which causes, or may cause, an interruption to or a reduction in, the quality of that service. Incidents are the result of failures or errors in the IT infrastructure and software. Incident management becomes more important as the contribution of information technology to the business is ever growing. Incident management also faces increasing challenges because an enterprise often maintains many applications in a shared IT environment comprising thousands of interdependent IT components, e.g., network, hardware, software, etc. Incident diagnosis often requires investigation on complicated causes aggregated from this environment.

Requests for incident resolution, also known as tickets, can be issued through multiple channels of communication, including phone calls, emails, and Internet portals. Unlike the requests in call centers, the requests for incident resolution that we are concerned with are deemed too complicated to be handled online by the helpdesk support team. Instead, in the so-called level-2 technical support, all requests are processed offline by technical support specialists, which we call the agents for simplicity. Typically, the agent who is assigned to the task contacts the requestor (or client) at a later time, identifies the cause of the incident, and works out a solution which is agreed upon by the client. The entire process may take hours or days, depending on the complexity of the problem.

Incident management or resolution services for information technology are labor-intensive operations. Because incidents have to be resolved in timely manner, performance targets are often set regarding the time to respond and the time to resolve an incident ticket. To meet these targets, adequate staffing is critical. At the same time, the utilization rate of the staff must also be taken into account to minimize the cost associated with an underutilized workforce. The management of IT incident resolution services always faces the question: how to staff a given operation properly, especially when the volume of service requests is expected to rise. Queuing models can be used to help address such questions.

In this chapter, we first describe a case of IT incident resolution service and a general process of incident ticket analysis; then, we review some basic concepts in queuing models, illustrate with examples the application of queuing models to the capacity planning and management of IT incident management services, and discuss some technical challenges of the queuing-model-based approach in practice. We end the chapter with some additional remarks.

**BACKGROUND**

**Incident Resolution**

Before the discussion of queuing models and their application to capacity planning, let us first have a brief discussion on the practice of IT incident resolution services.

Incident resolution for information technology is labor-intensive and requires deep expertise. To offer a glimpse at the complexity of such services, consider, as an example, a large software company that has over a thousand IT applications to support its business. These applications range from large-scale software packages (e.g., SAP) to small proprietary systems, which run in a shared, dynamic IT infrastructure. A critical objective of the incident resolution services is to maintain high-level availability of these applications and reduce outages.