

August 20, 2001

Just-In-Time Capacity Management: Another Dimension For Improving Web Sites

Jean-Pierre Garbani

Catalyst

A client inquiry

Question

What are the advantages of active service-level management?

Answer

Active service-level management is similar in concept to one of the great revolutions in manufacturing, just-in-time (JIT) inventory management. The basic idea is to provision only what is needed for a certain level of operation and putting in place a number of management functions that will trigger the provisioning of inventory is one of the key elements that allowed the manufacturing of goods to contain production costs. The same concept can be applied to IT and it is starting to gain momentum, especially in the management of Web applications.

Infrastructure sizing in Web applications — be it e-trade, e-commerce or simply Web hosting — is a difficult exercise, yet it is the key to the profitability of the site. It is a difficult exercise because IP over a wide area network (WAN), the Internet, suffers from a number of ailments in terms of predictability. These mostly stem from the fact that the traditional distributions are not applicable to Internet traffic (see “Where Mathematics Meet the Internet,” Walter Willinger and Vern Paxson, September 1998, www.ams.org/notices/199808/paxson.pdf), which makes the modeling of Internet traffic difficult but also from the lack of control over bursts of visits to a given site.

The unpredictability of the workload applied to a public site leads the way to two types of reactions:

1. Overprovisioning of infrastructure capacity. It is common to see Web sites running at five to 10 times the required “normal” capacity in order to absorb traffic bursts
2. The creation of a number of potential corrective measures: load balancing, traffic shaping, fast reconfiguration of servers, etc.

The first type of reaction is, of course, very detrimental to the potential profitability of a site. The second type is better in terms of profitability but requires the constant monitoring and understanding of the behavior of the global infrastructure and suffers from the latency due to human intervention.

The answer lies in the combination of monitoring, problem resolution and automated corrective actions. Intelligent performance management requires building an automated control loop of the infrastructure by understanding the behavior of the different components, understanding the root cause and acting, in real-time, on the corrective measures available. Effectively, this type of technique is equivalent to a just-in-time capacity management concept. It does provision, by optimizing the infrastructure, extra capacity when only when it is needed.

Resonate, known for its load balancing product, is entering the market with Commander, a product that exemplifies the concept: it monitors, analyzes and acts directly on the different JIT capacity systems, such as **Cisco** and **Nortel** switches. **Peakstone** is another example of the same idea.

What this brings to the market is another dimension in the performance management of Web sites. At a time when everyone with public-facing Web sites knows that performance is one of the keys to business success, it is paramount to reach the performance goals by optimizing the costs, not by overengineering the infrastructure. Optimizing costs while providing the necessary level of service is what we expect from just-in-time capacity management.