Army of digital "air traffic controllers" makes clouds more efficient, scalable

May 1 2014, by Jay Cadmus

IBM inventors have developed a method for managing how resources are used and work is done within a cloud by distributing control throughout the interconnected systems, reducing bottlenecks and increasing efficiency.

The IBM cloud computing invention, U.S. Patent #8645745: Distributed Job Scheduling In A Multi-Nodal Environment, was originally designed to help manage resources in high performance computing systems used for government and academic research. These systems consist of hundreds or thousands of computing resources that are connected to perform complex tasks where demands for system resources can rise and fall dynamically—similar to the model for cloud computing.

Tracking and prioritizing the many requests for service across these systems, as well as the resources needed to fulfill them—processors, storage, network bandwidth, input-output operations—has typically been performed at one central point, which can cause congestion and slowdowns as demand increases. This invention distributes this function across the various systems in the cloud, each managing the traffic within its own system, but also able to see and borrow resources from other systems to ensure speed and efficiency.

"Try to imagine the tens of thousands of airplanes that fly across the United States each day being managed by a single air traffic controller, who must keep track of each one and determine what runways, gates, maintenance facilities, etc. are available at every airport across the country," said IBM inventor Eric Barsness. "The best way to complete that task efficiently and effectively is to break it up to thousands of air traffic controllers, each of which has a view of all resources available and can prioritize and direct traffic accordingly. That's effectively what this invention can do within a cloud computing environment."

Applications for this invention in a cloud computing environment where demands for service can vary significantly include:

- A bank or financial institution, which experiences peaks in demand for processing and reporting information to meet daily, weekly, monthly or annual deadlines;
- A trading system, where requests to buy and sell stocks, bonds or commodities are volatile and must be completed very quickly;
- An online or brick-and-mortar retailer, which must address increased demand to complete transactions, manage inventory, and control shipping and logistics during peak periods such as Black Friday and Cyber Monday;
- An online gaming site, which may see substantial increases in traffic after school, at night or on weekends when more people have time to play;
- A medical research organization tracking the spread of an epidemic, gathering information in a variety forms, quantities and volumes from various and disparate sources across many geographies.

An army of "air traffic controllers"

Within cloud systems and other computer networks the function that determines what resources are available and decides what applications or requests for services have priority is normally centralized in one place. This control function—called a job scheduler—must evaluate what resources are available for use by various requesters of services.

It also must analyze which requests for services should get a higher priority based on diverse criteria. In some cases, the decision will be made based on how much processing power they require or how bandwidth intensive they are, how critical they are to business operations, or even which department it comes from or the status of the individual requester. During peak periods of activity, this function can become a bottleneck, slowing down the processing of requests and the performance of the tasks performed within the system.

With this invention, the job scheduling function is broken up into pieces and distributed throughout the system, with a job scheduler assigned to each node of a system or cloud. Each job scheduler can manage resources and prioritize requests within its own node. But if it determines that it needs more resources it can dynamically join with other nodes. One of the nodes becomes the primary node, controlling the shared resources within the node group.

The invention can also be useful in a hybrid cloud environment, where it can help determine when enough resources are available within a private cloud to meet service requirements and when it's necessary to burst out to a public cloud for additional resources. This can help control cost, as private cloud resources generally have lower operational costs but are finite in their capabilities, but also allows the flexibility to use a public cloud when necessary to ensure additional services are available when needed.

The patent for this software-based invention was awarded in February, and can be applied across IBM's portfolio of servers. IBM Systems and Technology Group offers a full range of offerings supporting public, private and hybrid cloud implementations that integrate with IBM's cloud software and services. This Systems portfolio includes IBM

System x racks and BladeCenter, NeXtScale, PureFlex, Power Systems and System z servers, and IBM Storage solutions.

Provided by IBM

Citation: Army of digital "air traffic controllers" makes clouds more efficient, scalable (2014, May 1) retrieved 29 April 2024 from https://phys.org/news/2014-05-army-digital-air-traffic-clouds.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.