Job Concentration Model
Simulation Report of Claim 1 of patents
US 11,204,810 & GB 2558063

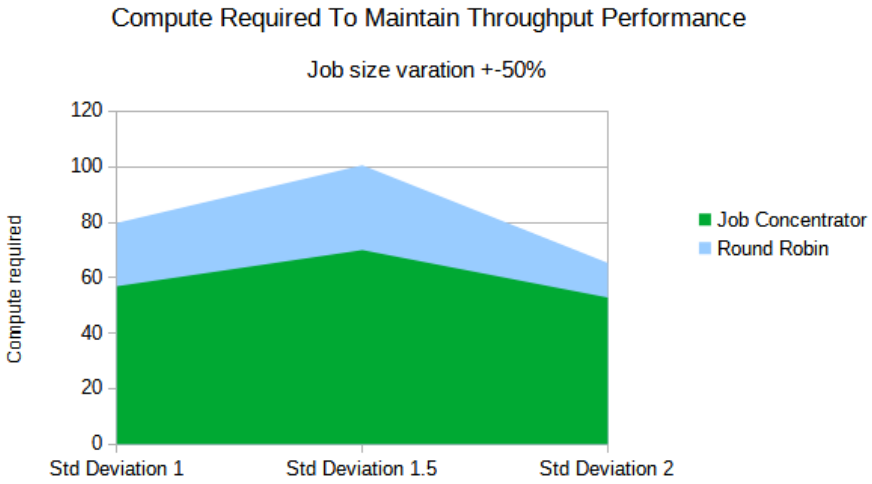
Will job concentration save compute overheads for compression, encryption, AI, document and image processing or serverless computing? Modeling indicates savings could be significant, but should largely depend upon the volumes and types of workloads involved. The following job-based scenarios may have reduced overall projected compute costs of up to 30%:

- Compression*
- Encryption
- Document processing
- Image processing*
- Artificial intelligence (on trained models)
- 'Serverless' computing

**Often required in artificial intelligence applications.*

Reducing wasted IT

Virtual computing common in today's data centers is typically set up in layers: At the base there is hardware such as racks of computers and networks. Upon this runs hypervisors to provide software-controlled access for various operating systems. In the next layer often come operating systems and optionally, application containers. Conceptually, on top of this stack runs applications which perform the data processing.



Data processing 'jobs' may consist of data sent to a program for processing or a mixture of data and programs sent to a virtual computing environment for execution. Such jobs are distributed to different virtual machines using a load balancer – like dealing out playing cards. They may be further distributed to different processors by an operating system using a task scheduler.

Before the Internet, jobs were often run in batches according to data center resource availability. But now more than ever, people expect jobs to be executed in real-time: Is the motion detected by the camera an intruder or is it authorized? What about new activity on the network? Can I spend my digital coin immediately? How might I eliminate "please try again later" for my customers? How can I maximize expensive higher-security infrastructure where computing jobs are isolated?

The connectivity of 5G means scenarios in which virtual computing environments are required to be 'always on' for fluctuating real-time processing, will likely continue to grow. Somehow, IT management is expected to predict in advance what all the demand curves are and will be, while simultaneously juggling systems maintenance. This is usually achieved by wasteful over-provisioning, such as "auto-scaling" capacity "on-demand". Such over-provisioning is often done according to preset or machine-learned network or processor resource usage thresholds, which can reduce effective average server utilization to 70% or less. Could job concentration eliminate some of this waste?

Job concentration is about replacing conventional load balancing with a mechanism for matching jobs to virtual machines, according to their available processor cores, also taking into account when a virtual machine was started. This is intended to reduce the inefficiency of having jobs dispersed throughout a virtual computing environment using unintelligent aggregate performance metrics and preset ratios or timetables, which may keep under-utilized capacity online for longer than required. Job concentration also tends to cycle applications through a population of virtual machines – with classes of virtual machines which have been running the longest being shut down first – keeping data processing updated and refreshed. This is intended to reduce the complexity and cost of managing virtual computing environments.

Model results

Modeling suggests job concentration can reduce over-provisioning of virtual computing by 50% with no degradation in performance or throughput. This was based on a normal distribution curve of jobs using a standard deviation of 1.5 and comparing job concentration with round-robin load balancing. Similar savings were also indicated by the simulation at a standard deviation of 2, with significant savings still indicated with a simulated job flow curve having a standard deviation of 1. The results are interpreted to represent compute cost savings under a wide range of load curves (in this case with a variation of +/- 50% in job size), thanks to fewer overheads thrown away by over-provisioning.

JOB SLOT OVERLOAD TEST			
Distribution type	**Approx SD Load Curve**	**% Over-provision**	**Average Slots Used**
Round Robbin LB	1.0	36	79.7
Job Concentrator	1.0	15	56.9
Round Robbin LB	1.5	32	103.6
Job Concentrator	1.5	13	71.1
Round Robbin LB	2.0	33	65.3
Job Concentrator	2.0	13	52.9

The modeling was only designed to emulate the job concentration of claim 1 of US11,204,810 & GB2558063. There may be scope for further gains if aspects of the other claims were also included in the model. However the model does not include an estimate of any additional savings in labour costs through the anticipated simplification of virtual computing environment operations. The working model including its source code is available on application.

Possible implications

The use of job concentration could help offset rising costs due to anticipated data center decentralization, security-by-isolation, and energy shortages, in a deteriorating economic environment – often without modifying existing applications. And as data nationalism/sovereignty drives adoption of higher security clouds, job concentration may also provide greater efficiencies in smaller, more private infrastructures. By replacing traditional load balancing, the technology may benefit existing applications with little or no modification, implemented perhaps as a highly visible feature for data-center users to select.

Further information:

Eric Wilson
DocBlaster Pty. Ltd. (A.B.N. 45 616 534 724)
ericw@mail.docblaster.com
www.docblaster.com