Optimal deployment for critical applications in Infrastructure as a Service

Imre Kocsis[1], Zoltán Ádám Mann[2], Dávid Zilahi[1]
[1]Department of Measurement and Information Systems
[2]Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Budapest, Hungary
ikocsis@mit.bme.hu, zoltan.mann@gmail.com, zilahia@gmail.com

Infrastructure as a Service (IaaS) offers tenants virtualized resources; most importantly, virtual machines (VMs) that can be created and destroyed on demand. However, tenants have only limited and indirect influence over the deployment of their reservations onto physical resources, making critical applications vulnerable to a variety of faults (e.g. common mode hardware and capacity faults or poor performance isolation). Operators and tenants strive to optimize their operations with conflicting goals in this setting. Operators aim at consolidating tenant VMs to as few hypervisors as possible to save power and air conditioning costs by switching off unused hosts [1]. Reliability, availability, performance stability and homogeneity are secondary concerns.

In contrast, a number of emerging critical cloud application categories require tenants to be able to formulate explicit requirements on the VM allocation in order to maintain strong Service Level Agreements (SLAs). An example is Network Function Virtualization (NFV) [2]: the current push in the telco domain to migrate network functions from dedicated appliances to IaaS. Carrier clouds as well as the cloud backends of cyber-physical systems (CPSs) are further examples. Most critical functions in these domains are real-time, heavily latency and throughput sensitive services; their deployment has to adhere to rules as CPU core affinity and antiaffinity policies (to avoid ”noisy neighbor” effects), the need to colocate components of a function on the same hypervisor (for performance) or explicit CPU time and network bandwidth allowance minimums.

Our paper contrasts VM placement optimization in this new setting with a unified view of ”classic” IaaS allocation optimization problem models [1]. NFV is used as a representative example. We show that at its core, placement remains a multi-aspect deployment optimization problem – much of the mathematical modelling *approaches* can be reused. However, adaptations as well as entirely new specific constraints and objectives are required. Our treatment focuses on environments serving critical applications – as NFV –, where cooperative allocation planning is a must. However, our approach has applications in general purpose IaaS, too – e.g. for soft real-time applications as virtual desktops.

References

[1] Z. Á. Mann, “Allocation of virtual machines in cloud data centers – a survey of problem models and optimization algorithms,” http://www.cs.bme.hu/~mann/publications/Preprints/Mann_VM_Allocation_Survey.pdf

[2] European Telecommunications Standards Institute, “Network functions virtualisation – introductory white paper,” https://portal.etsi.org/NFV/NFV_White_Paper.pdf