The Beginning of a Cognitive Software Engineering Era with Self-Managing Applications

Giovanni Morana
C3DNA Inc
Cupertino,CA, USA
giovanni@c3dna.com

ABSTRACT

The recent explosion of data and the resurgence of AI, Machine Learning and Deep Learning, and the emergence of unbounded cloud computing resources are stretching current software engineering practices to meet business application development, deployment and management requirements. As consumers demand communication, collaboration and commerce almost at the speed of light without interruption, businesses are looking for information technologies that keep up the pace in delivering faster time to market and real-time data processing to meet rapid fluctuations in both workload demands and available computing resources. While the performance of server, network and storage resources have dramatically improved by orders of magnitude in the past decade, software engineering practices and IT operations are evolving at a slow pace. This paper explores a new approach that will provide a path to self-managing software systems with fluctuation tolerance to both workload demands and available resource pools. The infusion of a cognitive control overlay enables an advanced management of application workloads in a distributed multi-cloud computing infrastructure. Resulting architecture provides a uniform framework for managing workload non-functional requirements such as availability, performance, security, data compliance and cost independent of the execution venue for functional requirement workflows.

CCS CONCEPTS

• **Software and its engineering** → **Software development techniques**; • **Theory of computation** → *Semantics and reasoning*; • **Applied computing** → IT governance;

KEYWORDS

Semantic Network, DIME Network Architecture, Dev-Ops, Connectivity, Modularity, Turing O-Machine, Cloud Computing

ACM Reference Format:

Giovanni Morana. 2018. The Beginning of a Cognitive Software Engineering Era with Self-Managing Applications. In *Proceedings of ACM Woodstock conference (WOODSTOCK'97),* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.475/123_4

WOODSTOCK'97, July 1997, El Paso, Texas USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

As we enter the third era of business process automation, each era has been punctuated by rise of complexity followed by simplification. These automation waves, fueled by hardware and software advances, drove human, computer interactivity and simplification of programming. High performance servers, network equipment and storage devices along with software technologies that made business process automation possible led the first wave. Virtualization of network, storage and computing services to address the efficiency, resiliency and scaling demanded by the global need for business services led the second wave. The success of these technologies [2, 3, 12] has in turn accelerated the demand for communication, collaboration and commerce conducted by both people and devices almost at the speed of light and created a new requirement for uninterrupted service delivery with high security and optimal performance without increasing the complexity in today's application management.

Recent trends [5, 8–11, 18–20, 23] in enterprise workload management have focused on leveraging multiple cloud infrastructures offered by different service providers and automating the provisioning of resources and configuring end-to-end application components, their dependencies and definable actions.

These advances are aimed at:

(1) Reducing time to market,
(2) Improve OPEX and CAPEX using commodity cloud resources and
(3) Improve developer agility by automating infrastructure provisioning and application deployment in public or private clouds

Once provisioned, all these approaches however, fall short in addressing the automation of run-time workload quality of service (QoS) assurance dealing with the availability, performance, security, data-compliance, and cost requirements. This is especially important in the face of fluctuations in workload demands and available computing resources, which vary with time and circumstance.

Examples are order of magnitude variation in pizza orders during a Super-Bowl game, Black Friday demand, increased processing of orders received by an on-line e-commerce application as the result of a promotion etc.

With the increased demand for always-on digital communication, collaboration and commerce services almost at the speed of light, real-time QoS assurance of applications/workloads without interrupting the user experience or disrupting service transactions in progress provides that extra competitive differentiation.

Current computing architecture delivering the services has evolved from server-centric, low-bandwidth, client-server origins to current distributed on-demand resilient infrastructure delivering business processes and workflows composed of applications connected with high-bandwidth networks. We can classify the evolution into three phases.

(1) The conventional datacenter consisting of a rigid network of servers optimized for delivering services. Adding or deleting servers in the cluster executing the service address fluctuations. This architecture is characterized by the large infrastructure provisioning and orchestration times often-involving labor intensive tasks accompanied by service interruptions and are only suited for services that can tolerate large service provisioning and orchestration times.
(2) As service workload demands increase with mobile, social, big data, and Internet of Things (IoT) technologies, the need for hyper-resilient infrastructures with shorter and shorter infrastructure provisioning times is met through virtualization and proliferation of on-demand cloud platforms. Resiliency and scale are improved by automating the administration of resources to meet the service scale and fluctuation demands.
This current state of the art provides hyper-resilient infrastructure islands with computing nodes that are mobile and scalable with shorter infrastructure provisioning and orchestration times. Public Cloud providers are examples of proprietary infrastructure islands with architectural and API dependency that lock-in the users to their infrastructure in the end. A new generation of infrastructure with open-source infrastructure management solutions based on OpenStack [24] is evolving with various infrastructure vendors embracing it as a common framework. Whether it is open or closed, the current approach heavily depends on moving the infrastructure by detecting the needs of application components to scale and address fluctuations.
When the infrastructure is distributed, often operated by multiple providers, use different architectures and management systems, conveying the service intent and orchestrating the resources using automation reaches a point of diminishing returns where complexity outweighs flexibility and increases the cost of coordination.
The choice then is between lock-in and management of complexity.
(3) The third phase is just emerging from a deep understanding of current limitations of von-Neumann stored program control implementation of the Turing machine.
It takes a different approach by extending the current computing, management and programming models to infuse cognition (service intent and its management) into computing by using a meta-model of the service workflow and manage the service by placing, and configuring dynamically, the component applications on the right infrastructure that provides the required resources.
The meta-model of the service and the service provisioning and orchestration software are decoupled from the infrastructure provisioning and orchestration software. The service provisioning and orchestration software negotiates with multiple service providers to provision a cluster of computing nodes with required resources to accommodate the speed and memory needs of service delivery. It monitors its own service QoS. When fluctuations demand scaling, based on RPO and RTO, it implements strategies based on policies to orchestrate the services by negotiating the resources required from infrastructure providers.
This model is more suitable to use a largely distributed resource pool and supports large scale computing with predictable application QoS assurance even when there are wild fluctuations in workloads, business priorities and latency constraints.
This approach puts survival and safety of service and its components first and allows isolation, diagnosis and fixing at leisure. The decoupling of services provisioning and orchestration from infrastructure provisioning and orchestration is accomplished by infusing cognition into distributed computing with a service control architecture. This allows reliable services to be deployed on even not so reliable infrastructure.

2 ADVANTAGES OF NEW APPROACH

The new approach [4, 6, 7, 13–17, 21, 22, 25] has many advantages going beyond the current state of the art providing unique benefits.

2.1 Infrastructure independence of application/workload run-time QoS

An application virtualization technology, with cognitive application management overlay, empowered with global knowledge, decouples the application management from the infrastructure management and assures the right resources (CPU, memory, network latency, bandwidth, storage IOPs, throughput and Storage capacity) in real-time to the right application component.
This enables applications to:

(1) Become Hypervisor agnostic and Operating System independent
(2) Become Agnostic to underlying IaaS and PaaS
(3) Be executed on either Virtual Machines, Containers[1], Bare Metal deployed in datacenters or Private/Public clouds
(4) Execute real time live Application/workload migration (Stateful), with zero transaction loss, between any infrastructures (Public or Private Cloud, datacenters with or without Hyper converged infrastructure anywhere) while providing end to end visibility and control to the application/workload owner, thus eliminating cloud/co-lo vendor lock-in to the customer.
(5) Provide application layer data replication across distributed infrastructures.
(6) Provide multi-cloud application/workflow orchestration and life cycle management during the run time by monitoring applicationâĂŹs vital signs and taking appropriate action based on its global knowledge of how to react based on pre-defined policies with the elimination of moving Virtual Machine and the associated complexity
(7) Migrate without requiring third-party tools leading to significant reduction of Op-ex to the enterprises.

(8) Migrate or live replicate to achieve continuous High Availability and Disaster Recovery without disturbing the service delivery (based on RTO/RPO policy setup) with resilience and with zero loss of transaction through multi-master, active-active, cross-cloud configuration.

2.2 Composed application workflows and their management on distributed infrastructures

Workflow QoS assurance automation decouples the application workflow management from the run-time infrastructure management systems (except for provisioning of resources) and assures run-time QoS based on policies, set up by the application/workload owner, without disrupting user experience or interruption to run-time service transactions. This enables applications/workflows to:

(1) Avoid vendor lock-in, complexity, tool fatigue, and innovation chaos while providing application/workflow self-repair, auto-scaling and live-migration on an interoperable network of choice of clouds connected using public or private networks.
(2) Migrate based on business owner-established policy containing the quality of service to decide what to do to address performance, security and compliance issues and to know where to get resources and do it.
(3) Allow to be managed by the owner of the applications with end to end service visibility and control across any infrastructure independent of native infrastructure management systems.
(4) Allow to be managed with blueprint based workflow configuration and multiple migration policies which can be set up off-line to suite different environmental conditions and user requirements.
(5) Legacy Application Migration: Make legacy apps/workloads cloud ready and on-boarding them to a cloud of choice without vendor lock-in or the complexity of âĂIJLift and shiftâĂİ. A discovery based workflow profiling and blueprint creation process allows easy migration of legacy applications and workflows.
(6) Decoupling Development and Operations of applications, and workflows: Developers develop applications without regard to how they are deployed at run-time and on what infrastructure. Operators define run-time QoS based on business priorities, deploy blue-print based workflows and assure their run-time behavior.
(7) Allow to deploy >99.99 application workflows on even a 99.99 infrastructure
(8) Enable workflow level security in addition to infrastructure based security.
(9) Enable real-time data switching and control at the application layer independent of infrastructure allowing geo-fencing and other data-compliance features

In summary, the business workflows, applications and workload evolution are rendered cognitive, and capable of self-provisioning, self-healing, self-monitoring, self-protecting, and self-controlling to adjust their structure and maintain desired quality of service all the time. It allows workload resiliency and tolerance to fluctuations both in demand and available resource pools at scale while lowering OPEX significantly.

The video $https://youtu.be/tu7EpD_bbyk$ demonstrates the application of the cognitive control overlay to manage workload quality of service in a multi-cloud environment.

3 CONCLUSIONS

In summary, the cognitive control overlay introduced by the new approach allows managing application workloads at run-time independent of their execution venue.

Current state of the art consists of providing secure mobility of virtual machines to deliver self-repair, auto-scaling and live-migration of applications and services. There are some recent attempts (Docker, LXC [1]) using containers to provide fine grain mobility of applications without moving the Virtual Machine itself. These containers still provide hardware abstraction and suffer same issues that are faced by virtual image mobility requiring infrastructure management connectivity and control.

The cognitive control overlay approach uses the knowledge of both the workload requirements and available resource pools to configure monitor and control their evolution in real-time.

REFERENCES

[1] David Bernstein. 2014. Containers and cloud: From lxc to docker to kubernetes. *IEEE Cloud Computing* 1, 3 (2014), 81–84.
[2] A Binu and G Santhosh Kumar. 2011. Virtualization techniques: a methodical review of XEN and KVM. In *International Conference on Advances in Computing and Communications*. Springer, 399–410.
[3] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic. 2009. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems* 25, 6 (2009), 599–616.
[4] Giovanni Cammarata, Rao Mikkilineni, Giovanni Morana, and Riccardo Nocita. 2018. Improving Security with Cognitive Workflows. In *International Conference on Emerging Internetworking, Data & Web Technologies*. Springer, 429–436.
[5] Antonella Di Stefano, Giovanni Morana, and Daniele Zito. 2013. Scalable and Configurable Monitoring System for Cloud Environments. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2013 IEEE 22nd International Workshop on*. IEEE, 134–139.
[6] Eugene Eberbach and Rao Mikkilineni. 2014. Cloud Computing with DNA Cognitive Architecture in the Context of Turing's" Unsinkable" Titanic Machine. In *WETICE Conference (WETICE), 2014 IEEE 23rd International*. IEEE, 125–130.
[7] Eugene Eberbach, Rao Mikkilineni, and Giovanni Morana. 2012. Computing models for distributed autonomic clouds and grids in the context of the DIME network architecture. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2012 IEEE 21st International Workshop on*. IEEE, 125–130.
[8] Pankaj Goyal. 2010. Enterprise usability of cloud computing environments: issues and challenges. In *Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE), 2010 19th IEEE International Workshop on*. IEEE, 54–59.
[9] Pankaj Goyal and Rao Mikkilineni. 2013. Implementing managed loosely-coupled distributed business processes: a new approach using dime networks. *International Journal of Collaborative Enterprise* 3, 2/3 (2013), 93–109.
[10] Pankaj Goyal, Rao Mikkilineni, and Murthy Ganti. 2009. FCAPS in the business services fabric model. In *Enabling Technologies: Infrastructures for Collaborative Enterprises, 2009. WETICE'09. 18th IEEE International Workshops on*. IEEE, 45–51.
[11] Pankaj Goyal, Rao Mikkilineni, and Murthy Ganti. 2009. Manageability and operability in the business services fabric. In *Enabling Technologies: Infrastructures for Collaborative Enterprises, 2009. WETICE'09. 18th IEEE International Workshops on*. IEEE, 39–44.
[12] Jon Matias, Jokin Garay, Nerea Toledo, Juanjo Unzilla, and Eduardo Jacob. 2015. Toward an SDN-enabled NFV architecture. *IEEE Communications Magazine* 53, 4 (2015), 187–193.
[13] Rao Mikkilineni. 2011. *Designing a new class of distributed systems*. Springer Science & Business Media.
[14] Rao Mikkilineni, Albert Comparini, and Giovanni Morana. 2012. The Turing O-Machine and the DIME Network Architecture: Injecting the Architectural Resiliency into Distributed Computing.. In *Turing-100*. Citeseer, 239–251.
[15] Rao Mikkilineni and Giovanni Morana. 2014. Infusing Cognition into Distributed Computing: A New Approach to Distributed Datacenters with Self-Managing

Services on Commodity Hardware (Virtualized or Not). In *WETICE Conference (WETICE), 2014 IEEE 23rd International*. IEEE, 131–136.

[16] Rao Mikkilineni, Giovanni Morana, and Daniele Zito. 2015. Cognitive Application Area Networks: A New Paradigm for Distributed Computing and Intelligent Service Orchestration. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2015 IEEE 24th International Conference on*. IEEE, 51–56.

[17] Rao Mikkilineni, Giovanni Morana, Daniele Zito, and Surendra Keshan. 2017. Cognitive application area networks. *International Journal of Grid and Utility Computing* 8, 2 (2017), 74–81.

[18] Rao Mikkilineni and Vijay Sarathy. 2009. Cloud Computing and the Lessons from the Past. In *Enabling Technologies: Infrastructures for Collaborative Enterprises, 2009. WETICE'09. 18th IEEE International Workshops on*. IEEE, 57–62.

[19] Rao Mikkilineni and Ian Seyler. 2011. Parallax-A new operating system prototype demonstrating service scaling and service self-repair in multi-core servers. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2011 20th IEEE International Workshops on*. IEEE, 104–109.

[20] Mohamed Mohamed, Djamel Belaïd, and Samir Tata. 2013. Self-Managed Micro-containers for Service-Based Applications in the Cloud. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2013 IEEE 22nd International Workshop on*. IEEE, 140–145.

[21] Giovanni Morana. 2017. Self-Managing Distributed Systems and Globally Interoperable Network of Clouds. In *Multidisciplinary Digital Publishing Institute Proceedings*, Vol. 1. 188.

[22] GIovanni Morana and Rao Mikkilineni. 2011. Scaling and self-repair of linux based services using a novel distributed computing model exploiting parallelism. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2011 20th IEEE International Workshops on*. IEEE, 98–103.

[23] Vijay Sarathy, Purnendu Narayan, and Rao Mikkilineni. 2010. Next generation cloud computing architecture: Enabling real-time dynamism for shared distributed physical infrastructure. In *Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE), 2010 19th IEEE International Workshop on*. IEEE, 48–53.

[24] Omar Sefraoui, Mohammed Aissaoui, and Mohsine Eleuldj. 2012. OpenStack: toward an open-source solution for cloud computing. *International Journal of Computer Applications* 55, 3 (2012).

[25] Francesco Tusa, Antonio Celesti, and Rao Mikkilineni. 2011. AAA in a cloud-based virtual dime network architecture (dna). In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2011 20th IEEE International Workshops on*. IEEE, 110–115.