GAIA: Strengthening the Reliability of Datacenter Computing with Fast and Scalable Consensus

Abstract:

To handle the rapidly increasing volume of data, more and more software applications run within a datacenter containing numerous computers. To harness the massive computing resources, applications are deployed by two major types of infrastructures: schedulers and virtual machines (VM). These infrastructures have brought many benefits, including improving resource utilization, balancing computer computer loads, and saving energy. Unfortunately, as an application runs on more computers, minor computer failures will occur more likely and can turn down the entire application, causing disasters such as the 2015 NYSE trading halts and recent Facebook outages. Existing infrastructures lack a high availability support for applications.

This GAIA project takes a holistic methodology to greatly improve application availability via three objectives. First, we will create a fast, scalable distributed consensus protocol for general applications. Distributed consensus, a fault-tolerance theory, runs multiple replications of the same application and makes these replications behave consistently as long as a majority of them still work normally. An open challenge is that traditional consensus protocols are too slow mainly because their messages go through various software layers (e.g., OS kernels). We will create a fast consensus protocol called FALCON with an ultra-fast OS kernel bypassing technique called Remote Direct Memory Access. Preliminary results published in [SOSP '15] show that FALCON supports unmodified real-world applications, and it is 11.2X faster than traditional protocols even when running on 35X more computers than these protocols.

Second, we will construct TRIPOD, the first fault-tolerant scheduler by integrating Falcon with popular datacenter schedulers. To seamlessly manage computing resources and replications, we propose a novel replication-aware resource allocation workflow. Preliminary results published in [APSys '16] show that TRIPOD supports a real-world key-value store application efficiently.

Third, we will make our protocol and VMs form a mutual-beneficial eco-system. This eco-system not only leverages the VM hypervisor to achieve automatic replication, but it also nurtures a new lightweight VM live migration approach for balancing computer loads. To migrate a VM to a remote computer, prior approaches incur substantial VM downtime or resource consumption on the local computer. Our new approach needs only migrate a consensus leadership, which consumes little time and resource.

The success of this GAIA project will greatly strengthen the reliability of many datacenter applications and will benefit almost all computer users and software vendors, including many financial platforms in HK. GAIA will also advance broad datacenter techniques (e.g., VM migration) and attract researchers to build more reliable infrastructures.

Long term impact:

To deal with the rapidly increasing volume of data, more and more software applications run within a datacenter containing numerous computers. Many applications are mission-critical (e.g., financial platforms, social network platforms, military services, and medical services), so they naturally demand both high reliability and performance. To harness the massive computing resources, applications are deployed with two major types of datacenter infrastructures: schedulers and virtual machines (VM). These infrastructures have brought many benefits, including improving resource utilization, balancing computer loads (e.g., VM live migrations), and saving energy.

Unfortunately, as an application runs on more computers, minor computer failures will occur more likely during application execution. If the failure computer runs a critical component of this application, the entire application can be turned down. For instance, due to minor computer errors, New York Stock Exchange (NYSE) has experienced a whole-site outage in 2015, and Nasdaq encountered a one-hour IPO delay in 2012. Moreover, social network applications such as Facebook tend to be online 24-7, but minor computer failures have turned down the whole Facebook site for several times in recent years. All these disasters have caused huge money lost.

A key problem of these disasters is that datacenter infrastructures lack a high availability support for general applications. Therefore, although some companies have built replication approaches to improve the availability of individual applications, most other applications still suffer.

The proposed GAIA project tackles this problem with a holistic methodology: it first builds a fast, scalable consensus protocol, it then integrates this protocol into two major infrastructures. This methodology has two main benefits. First, we can greatly improve the fault-tolerance of mission-critical applications. Distributed consensus is recognized a strong fault-tolerance theory because it maintains multiple consistent replications of the same application to overcome minor computer failures. Although consensus uses extra computing resources for fault-tolerance, in practice, it has been widely considered worthwhile by academy and industry, because the computing resource capacity is often not a bottleneck for mission-critical applications.

A second benefit of our methodology is that it is easy to make the fault-tolerance itself robust. Consensus is notoriously difficult to be robust because its distributed protocol is extremely hard to understand, build, or test. For instance, although some top companies (e.g., Microsoft) have built their own consensus protocols for individual applications, ironically, recent research works have detected numerous bugs in these protocols. Building one consensus protocol for each individual application could be a nightmare for all application developers. The methodology of GAIA is to build a general consensus protocol and to integrate it with two major infrastructures.

Therefore, state-of-the-art only needs to focus on testing our protocol and infrastructures, then we can make many applications enjoy robust fault-tolerance.

We envision significant impacts of this GAIA project in three different terms.

In the near term, Objective 1 can largely improve both the scale and performance of many replication systems. For instance, a notable key-value store system called Scatter deploys 8~12 replicas in each consensus group, and now it can deploy hundreds of replicas in each group and achieves much better performance. Overall, a fast, scalable, and general consensus protocol will significantly improve both the performance and fault-tolerance of various software systems in datacenters.

In the intermediate term, by realizing Objective 2 and 3, we will greatly strengthen the reliability of datacenter applications and benefit almost all computer users and software vendors. For instance, HK has lots of financial platforms which naturally desire high availability in their operational hours, and this GAIA project can bring practical reliability benefits to these platforms and avoid horrible outages such as the NYSE trading halts in 2015.

In the long term, we anticipate that this GAIA project will advance broad datacenter techniques (e.g., VM live migration) and attract researchers to build more reliable datacenter infrastructures. As datacenter emerges to be a "giant computer", a future datacenter OS for such a computer will gradually come up. Therefore, consensus protocols, schedulers, and virtual machines will become essential components for this OS, and the outcomes of GAIA will eventually be adopted in a future datacenter OS.