Clouds are affected by malicious attacks and failures of the infrastructure (e.g., power failures). Such

events can affect Internet domain name servers and prevent access to a cloud or can directly affect

the clouds. For example, an attack at Akamai on June 15, 2004 caused a domain name outage and a

major blackout that affected Google, Yahoo!, and many other sites. In May 2009 Google was the target

of a serious denial-of-service (DoS) attack that took down services such Google News and Gmail for

several days.

Lightning caused a prolonged downtime at Amazon on June 29 and 30, 2012; the AWS cloud in the

Eastern region of the United States, which consists of 10 data centers across four availability zones,

was initially troubled by utility power fluctuations, probably caused by an electrical storm. A June 29,

2012 storm on the East Coast took down some Virginia-based Amazon facilities and affected companies

using systems exclusively in this region. Instagram, a photo-sharing service, was one of the victims of

this outage, according to http://mashable.com/2012/06/30/aws-instagram/.

The recovery from the failure took a very long time and exposed a range of problems. For example,

one of the 10 centers failed to switch to backup generators before exhausting the power that could be

supplied by uninterruptible power supply (UPS) units. AWS uses “control planes” to allow users to

switch to resources in a different region, and this software component also failed. The booting process

was faulty and extended the time to restart EC2 (Elastic Computing) and EBS (Elastic Block Store)

services. Another critical problem was a bug in the elastic load balancer (ELB), which is used to route

traffic to servers with available capacity. A similar bug affected the recovery process of the Relational

Database Service (RDS). This event brought to light “hidden” problems that occur only under special

circumstances.

Arecent paper [126] identifies stability risks due to interacting services.Acloud application provider,

a cloud storage provider, and a network provider could implement different policies, and the unpredictable

interactions between load-balancing and other reactive mechanisms could lead to dynamic

instabilities. The unintended coupling of independent controllers that manage the load, the power consumption, and the elements of the infrastructure could lead to undesirable feedback and instability

similar to the ones experienced by the policy-based routing in the Internet Border Gateway Protocol

(BGP). For example, the load balancer of an application provider could interact with the power optimizer

of the infrastructure provider. Some of these couplings may onlymanifest under extreme conditions and

be very hard to detect under normal operating conditions, but they could have disastrous consequences

when the system attempts to recover from a hard failure, as in the case of the AWS 2012 failure.

Clustering the resources in data centers located in different geographical areas is one of the means

used today to lower the probability of catastrophic failures. This geographic dispersion of resources could

have additional positive side effects; it can reduce communication traffic and energy costs by dispatching

the computations to sites where the electric energy is cheaper, and it can improve performance by an

intelligent and efficient load-balancing strategy. Sometimes a user has the option to decide where to

run an application; we shall see in Section 3.1 that an AWS user has the option to choose the regions

where the instances of his or her applications will run, as well as the regions of the storage sites. System

objectives (e.g., maximize throughput, resource utilization, and financial benefits) have to be carefully

balanced with user needs (e.g., low cost and response time and maximum availability).

The price to pay for any system optimization is increased system complexity, as we shall see in

Section 10.7. For example, the latency of communication over a wide area network (WAN) is considerably

larger than the one over a local area network (LAN) and requires the development of new

algorithms for global decision making.