Unanimous

In Pursuit of Consensus at the Internet Edge

Heidi Howard, University of Cambridge

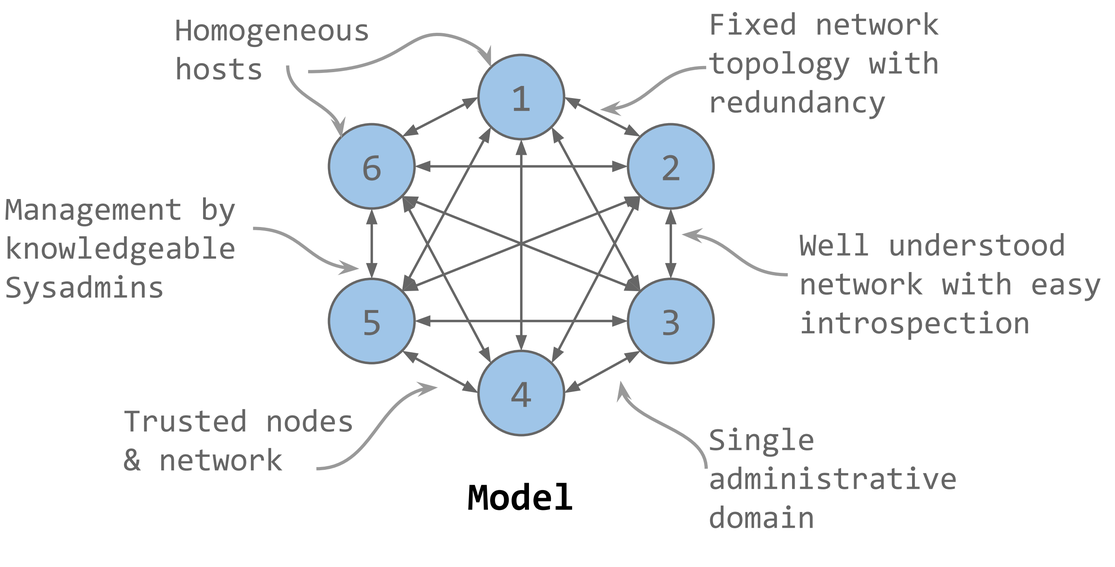
Problem

Consensus algorithms are built upon either Paxos's decade old assumptions or assume a datacenter like environment, neither approach can tolerate many of the common failures on the internet edge.

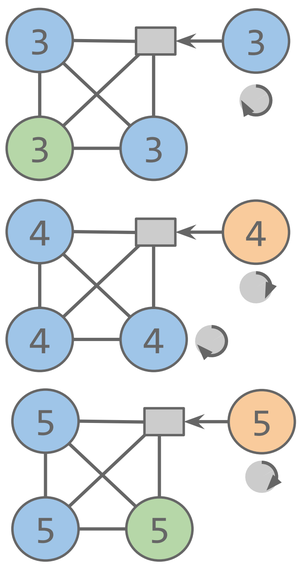
Consensus algorithms like Multi-Paxos are famously underspecified and even modern consensus algorithms remain underspecified in the pursuit of understandability.

Underspecification and outdated assumptions are just two examples of many issues with consensus at the internet edge. This poster introduces Unanimous, a new consensus algorithm for the internet edge.

Let's take a look at three examples to illustrate this problem. Here we have three sample executions from the recently developed Raft algorithm, which uses terms numbers and strong leadership to achieve consensus:

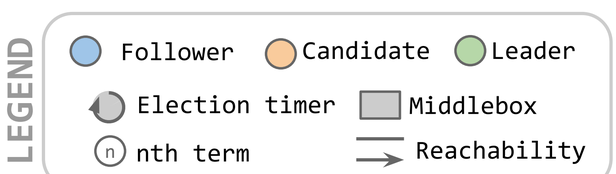


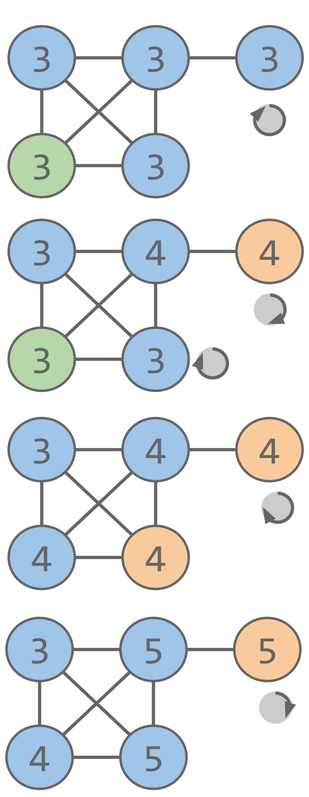
Model



Example 1: Symmetric Reachability

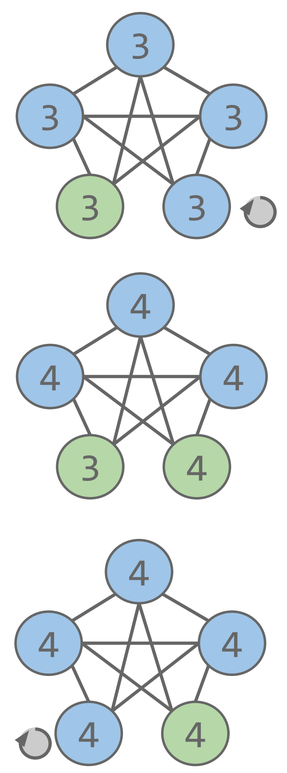
In this example (to the left), the node in the top right is behind a misconfigured middlebox, it is able to transmit messages to other nodes but not hear the responses. This node will never be elected leader nor hear the current leader, thus it will continuously timeout, incrementing its term and terminating the current leader. This demonstrates that in the protocol a leader is too quick to step down.

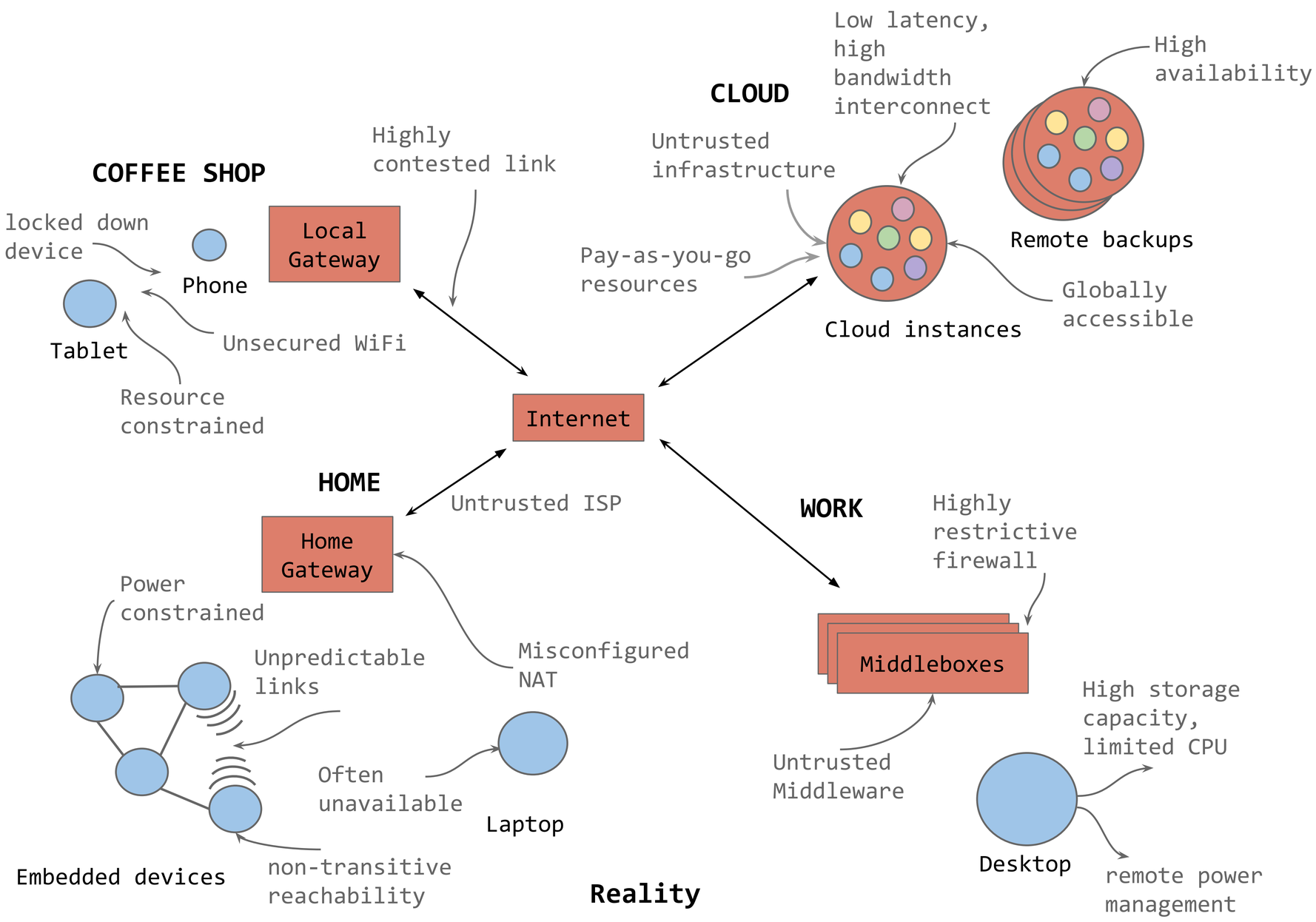




Examples 2 & 3: Transitive Reachability

In these examples (left & right), a node cannot hear the current leader thus becomes a candidate, increasing the term and terminating the current leader. The figure on the right demonstrates that the first node to detect a failure may not be the best next leader, as leadership bounces between the two nodes either side of the failure. The figure on the left shows that the protocol is too quick to terminate leadership.





Reality

Approach

1. Designed for developer usability and performance, even in the hostile internet edge.
2. Based on the reality of the modern internet, not Paxos's model assumptions.
3. Conservative leader election with smart failure detectors, converging towards the most reliable and highly connected nodes
4. A complete modular specification with extensions such as dynamic membership, byzantine fault tolerance, load balancing and address discovery.
5. Fine-grained approach to participation including various degrees of passive participation.

Our Related Projects

Signposts - authenticated identities and transitive reachability for the edge network *[FOCI'13]*

Databox - manifesto for an alternative to third party centralised services *[*arXiv: 1501.04737*]*

Raft Refloated - reproduction study of the Raft consensus paper *[SIGOPS OSR Jan'15]*

let's continue the discussion:

email: heidi.howard@cl.cam.ac.uk
homepage: www.cl.cam.ac.uk/~hh360
twitter: @heidiann360