Applications of weakly consistent distributed data structures over shared logs

Pranav Maddi

Contents

Abstract

The purpose of this is to help understand how the weaker consistency models allowed through shared-log environemnts could be useful for software defined networking. Describe the types of state maintained by the ONOS and OpenDaylight SDN controllers.

1 Background

The purpose of this report is to document the current progress in SDN systems and discuss the types and requirements of state in these systems. The paper introduces the idea of the shared-log architcture for providing causal consistency. Ultimately, we would like to motivate the idea that causal consistency would be useful in providing weaker consistency guarantees for certain types of shared state among SDN controllers.

1.1 Shared logs

A set of totally ordered shared logs can provide weaker consistency on the data structure built on top of it. Each element in the logs have a set of dependencies. That is, to know element `e`, all of the elements of `e.dependencies` must be loaded. By allowing dependencies to be specified between elements of various shared logs, arbitrary partial orderings of nodes can be constructed while preserving the convenience of having totally ordered logs. By carefully selecting what operations on the higher level data structure require global reads and writes, causal ordering can be guaranteed across the network.

Tango [3] aims to provide a way to store metadata by allowing transactional access to arbitrary data structures with persistence and high availability. The Tango paper notes that some services such as ZooKeeper [8] provide persistence, high availability, and strong consistency—but only for specific data structures.

Tango objects are a class of data structures stored in memory that built on the shared log. The history of the object is stored directly in the log, and each object can have multiple views which are copies of the data structure stored by the clients. The source of truth in the system is the shared log, and the views are "soft state" that can be reconstructed by replaying the shared history. Modifications to Tango objects are made by appending updates to the history, and accesses are made by syncing the view with the history.

Using a shared log makes strongly consistent operations easier than they would be in a conventional distributed system. Such applications include remote mirroring, coordinated rollbacks, and consistent snapshots across objects. These options are available since views can be synced to any offset in the log and replayed as needed.

The shared log abstraction represents a total ordering of the updates added to the log. This abstraction is a common way to organize changes to a file system, for instance [16]. By definition, this means that it is a set with a relation that satisfies reflexivity, transitivity and symmetry, where any two items can be compared. On the other hand, we want to represent a partial ordering of the updates so that we can allow the user the flexibility of not relying on other users writes. A directed acyclic graph is a data structure that can represent partial ordering relations. Initially, we considered using a shared DAG as a representation of state in the system. However, this seemed to leave the system too unconstrained since it would be difficult to maintain ordering across nodes and read forward to get new data.

The shared log runtime exposes the following API:

- `runtime.append(id, {dependencies}, data) -> index` Adds an entry to column id with containing data with dependency set `{dependencies}` and returns the index at which the entry was written
- `runtime.read_next(id) -> Option index` Attempts to read one unread entry from column id, along with all unseen dependencies of said entry, and if an entry was read returns its index
- `runtime.play_foward(id) > Option (index, index)` Reads all unseen en- tries for a given column, along with their dependencies, returning the first and last indices read.

1.2 Consistency

Not all applications require a strong consistency guarantee. Depending on the particular application, lower latency reads or the ability to write immediately could be preferred. We would like to enable our users to select what level of consistency works for them. Rather than have the consistency level be specified at development time, we would like to support this at the API level. By having something like a Service Level Agreement specified within the update_helper and query_helper, the client could configure the guarantee that works best in the situation. The SLA specified would then inform where in the DAG to place the update. The motivation for such a system is that different procedures in an application can have different guarantee requirements.

SLA requirements we may support include: [17]

strong returns the value of the last preceding Put(key) performed by any client

read-my-writes returns the value written by the last preceding Put(key) in the same session or returns a later version; if no Puts have been performed to this key in this session, then the Get may return any previous value as in eventual consistency.

bounded(t) returns a value that is stale by at most t seconds. Specifically, it returns the value of the latest Put(key) that completed more than t seconds ago or some more recent version.

causal returns the value of a latest Put(key) that causally precedes it or returns some later version. The causal precedence relation $<$ is defined such that op1 $<$ op2 if either (a) op1 occurs before op2 in the same session, (b) op1 is a Put(key) and op2 is a Get(key) that returns the version put in op1, or (c) for some op3, op1 $<$ op3 and op3 $<$ op2

Any of these consistencies could be supported with the multiple shared log structure, as long we carefully implement the distributed data structure on top of it.

2 SDN Controllers

In general, SDN allows central control of a network. It aims to keep the complexity of the network manageable by providing abstractions for the underlying network devices. SDN provides four main advantages over traditional networks:[9]

- Separation of the control and data planes
- Centralization of the control plane
- Programmability of the control plane
- Standardization of API's

[10]

Networking protocols are often arranged into data and control planes. The data plane consists of all the messages that are generated by the network clients. The network must then decide how to transport the packets from source to destination, which requires some level of decision making by the control layer. In a traditional network, these decisions are made by the control layer of the router which determines which routing behavior to perform. This might include computing L3 or L2 routing protocols such as Open Shortest Path First or Spanning Tree. A network manager might keep track of traffic statistics and the network state.

With SDN, this control logic is both separate and centralized, away from the data plane. The data plane still forwards the packets using the forwarding table assigned by the control plane, but that logic is separated out into a controller. This has the added benefit of reducing the cost of the switches. SDN's centralization of control makes determining the network state faster than with standard distributed network practices, and policy changes can be propagated much faster. Centralized control in SDN networks also have scaling issues, and it is necessary to divide the network into subnets that can share control strategy. OpenFlow is the main southbound protocol for controllers to interact with routers and switches. It gives the controller access to the forwarding plane of the network, and provides the ability to add, modify, and remove packet matching rules and actions.[14]

2.1 ONIX

ONIX was the first of these distributed SDN controllers to be released with the goal of providing networking management view abstractions apart from NOX, which did not address reliability or scalibility concerns [7]. ONIX also provides a more general API than it's predicesors, along with the felxibility to connect to general datastores and southbound protocols [11].

ONIX consists of the unerlying physical infrastructure that needs to be managed, connectivity infrasctructure, Onix controllers, and the application logic. The system also introduces the idea of the Network Information Base (NIB) which extends the idea of the Routing Information Base to include a graph of the network topology. This topology is how applications primarily interact with the system, both for modifying and accessing network state. Therefore, much care is taken to allow the application to distribute this information and replicate it among the controllers.

The NIB provides methods for the applicaiton logic to modify the network elements through an index of the network elements, as well as a notification system to asynchronously alert the appliation of state changes of watched network elements. When making these state changes, the system doesn't allow locking across the system, but only on the local controller without guarantees on the global state by default.

To add some level of consistency to sets of updates that might have dependencies on one another, ONIX provides a synchronization primative. Though all operations are asynchronous, this primitive allows callbacks to be executed once a change propagated to avoid race conditions.

To improve the scalablity of the system, the controller logic can be is partitioned over multiple Onix controllers depending on the application requirements. This is handled in two ways: dividing the network traffic in case the forwarding state entries exceed the memory on single controllers, and dividing the NIB over the controllers too if its representaion cannot fit in memory. Generally, the full NIB can fit in memory and writes are rare, so it ist stored in the fully replicted datastore across all controllers. The link utilization, however, is always changing and is stored in a DHT with weaker consistency guarantees.

Another way the systems aims to improve performance is by aggregating each subset of network elements managed by each Onix node into single nodes. Then, the topology can be represented by the links amongnetwork subsets. This avoids spamming external CPUs with high rates of network events since there are fewer links to describe. A downside to this is that the DHT outside of the subset would only have link-utilization for the aggregated topology. To increase performance from this point, the systems would need partition the NIB over controllers.

To increase the reliablity of the sysetm, Onix deals with four types of network failure: forwarding failure, link failure, Onix instance failure, and network connectivity failure. To deal with forwarding and link failures, the controllers direct traffic around the failures. These failures are disseminated as updates to the topology and force recomputation of the forwarding failures. When an Onix controller fails, the Zookeeper coordination service sees that the failure happened by the lack of a keep-alive heartbeat. Then, the node responsibilities are pawned off as needed. The controllers must also resolve race conditions that may have occured from incompletely written state.

2.2 ONOS

ONOS is an open source network "operating system" that aims to have distributed control of the network. Like the other systems covered here, ONOS uses the OpenFlow control-data plane interface. The overall architecture of ONOS is composed of a network view that contains a network graph abstraction, OpenFlow managers, as well as a Zookeeper registry for coordination. The main goal of ONOS is to provide effective network virtualization.[4]

To interact with the OpenFlow interface, ONOS uses modules from Floodlight, which is another SDN implementation. This includes the switch manager, the IO loop, and link discovery. These modules as well as shared Floodlight module manager means that ONOS is slightly more compatible with other SDN controllers.

The network view data model was stored in an ONOS graph database built on top of a distributed and persistent key-value store called *RamCloud*. RamCloud exposes the Blueprint graph API which applications use to access the network state. The system guarantees eventual consistency in the graph view.[4]

The transition semantics of the graph database are maintained to correspond with structural nature of the graph, such as requiring that an edge connects two nodes. To justify that eventual consistency is good enough for the graph view, some of the ONOS experiments show that when using redundant and identical flow path computation on all instances, the computation modules on each instance converge to the same path. Also, legacy networks have embedded control and are not logically centralized, so it's eventual consistency doesn't impede it's normal operations. However, the authors of the paper admit that there could be some benefit to having some sequencing guarantees that provide deterministic state transition in the network.

Out of the box, ONOS has support for a distributed

controller architecture. It can run on multiple servers, where each is the master OpenFlow controller for some switches, and any switch has exactly one master controller. Therefore, an ONOS instance must propagate state changes from the switches in its domain to the network of ONOS controllers. The system also aims to be scalable so that as data plane capacity grows or control plane demand increases, more ONOS instances can be created.

ONOS uses Zookeeper as a coordination mechanism for general fault tolerance and master election for switches. Since ONOS is distributed, it has secondary controllers for each switch in case the primary controller goes down. That way, the load of a primary failure are distributed to the remaining controllers. During runtime failure of any nodes detected by the lack of a keep-alive message, a replacement leader or follower is elected to replace it.

Since ONOS assigns a switch to multiple instances, and exactly one master, it also uses a master election process for failure handling. The switch's master connection communicates all the relevant information about the switch to the network view, and the secondary controllers for the switch are only required to elect a new master. The authors emphasize that using Zookeeper to coordinate this election ensures that switches always have a unique master.

The data model behind the graph abstraction represents each network object in a different table. This was implemented as a means to reduce the number of updates, since a previous version stored all the network objects in a single table.

2.3 OpenDaylight

The OpenDaylight defines a model driven approach for managing software defined networks. It aims to create a more modular system for SDN, where subscribers and providers for services are plugins to the system. This Model Driven Software Engineering approach creates a framework for relationships between models, standard mappings and patterns that enable model generation. An advantage of the system, therefore, is that it can create code and API's directly from models.[15]

OpenDaylight controller uses YANG as a data modeling language to specify these cross platform portable data types, which can describe network operation. While YANG was initially developed for this purpose, it effectively describes other network elements, including policy, protocol, and subscriptions to services. Rather than being an object-oriented structure, YANG is tree-structured and includes complex data types like lists and unions.[13]

For example to represent a grouping of nodes in YANG:[13]

```
grouping target {
    leaf address {
        type inet:ip-address;
        description "Target IP address";
    }
    leaf port {
        type inet:port-number;
        description "Target port number";
    }
}
container peer {
    container destination {
        uses target;
    }
}
```

To modify the data defined by YANG, OpenDaylight uses the NETCONF and RESTCONF protocols. NETCONF is a network management protocol that defines the configuration and operations on a data store. It also has the ability to run remote procedure calls and send notifications about the underlying data. NETCONF encodes its messages and data in XML. RESTCONF is similar to NETCONF, except that it exports a REST interface over HTTP for use by external applications.[6]

An operation configuring such a peer with NETFCONF:[13]

```
<peer>
  <destination>
    <address>192.0.2.1</address>
    <port>830</port>
  </destination>
</peer>
```

We will focus on the specific application of network virtualization in OpenDaylight. The network virtualization libarary of OpenDaylight comes packaged with the ODL Controller Platform and provides the functionality to run multiple controller nodes in a logically centralized manner [1]. Just as with ONOS, each switch in the system is assigned a leader controller.This virtualization services allow machines to be grouped into virtual network segments (VNS) which can be used to logically segment a network acorrding to a specialized use case. Some advantages of doing this include sharing of resources, isolation, agregration of resources, and increased ease of management [9].

Of the network functions served by the ODL controller, which include a statistics manager, forwarding rules manager, and an inventory manager, the topology service is of particular importance for network virtualization becuase it is required to translate virtual routing descisions to network flows [9]. To achieve this, the system stores the messages from BGP-LS and passes them

to the Topology Exporter that generates that makes available the topology of the controller's switchces via RESTCONF. The topology includes the connections between nodes, routers, switches, hosts and various appliances of the netowork. This system produces a reliable network view in the case of a single controller, but becomes more complicated if more than one controller is used, since the sub-network views must be constructed into the global view to make flow descisions [12].

3 State Maintainence

We examine the types of state maintained by software defined networks, and how they are maintained. Each of the systems stores state differently, with varying levels of flexibility depending on the application and with different default configurations. While the controllers differ in this regard, all have the ability to provide the following servies for storing and disseminating state:

- Datastore that provides transactions over the network with arbitrarily large fields
- Coordination system that prefers small fields and is optimized for reads over writes
- Distributed storage mechanisim that provides weaker guarantees but can provide higher availibility
- Notificaion system where controller logic can subscribe to state change events from particular elements of data

With these services, all of these systems aim to manage the state of the network through the network information base (by various names depending on the controller.) These include the representation fo the network topology as well as the state shared by the other controller instances. While this idea unites the various controllers, they behave differently in how they divide these abstractions among the data plane, state managment layer, and the application control logic. Newer systems such as OpenDaylight and ONOS tend to give the application higher level abstraction as to how to store state, and in the case of OpenDaylight completely abstracts the datastore as dynamically loadable resource that can be selected by the API. Older systems such as ONIX theoretically proivde the same functionality but simply define an interface for managing the resources.

Inventory

The main objects that the SDN controllers must track are the topological network elements and the controller nodes [2]. These entities include the routers, switches and hosts of the system.

Depending on how the controller aggregates network elements, this state can be represented in a few ways. ODL maintains an a set of network layers, and each one includes a list of nodes relative to its network layer. In ODL, there are two components of the inventory and topology management model, one of which defines the layers of the network along with inventories, and another which adds the links between the nodes in each layer to reflect how it is connected. Therefore, the system enforces that the inventories of nodes and their layering are stored separately from the link information to allows varying consitency guarantees as needed by the application.

How the inventory is managed depends primarily on the scale of the network since applciations prefer full access to the inventory when possible. When the NIB is small enough that it fits in memory, it can be kept in either a transactional database or in a coordination service. In ONOS, the node ownership tree is kept in ZooKeeper to maximize the consistency and availibility of the hosts. When the system scales to the point where the NIB cannot be kept in memory, the local set of nodes under its control can be kept in memory, while the remaining state is in the evetually consistent store.

Topology and link-state

The topology of the system includes the inventory of nodes, as well as the link-states between them. In general, the consistency requirements are weaker when dealing with link-state, since they can always be dynamically generated from the network. Applications tend to prefer availibility, and no applications should require persistence of the link-state. ONIX stores these eventually consistent state in a DHT with one-hop routing similar to Dynamo [**?**].

Different applications tolerate varying levels of inconsistency in the link-state. Ethane [11], which provides a flow managment interface, prefers the availibility of the link-state to better install flows on the switches.

igp topo, l3 unicast, ospf, isis generalized aggregetation as overlay networks onix nib etc typed entries consistency / durability depending on requirements. replicated transactional database and one hop DHT for more tolerant things distribute nib net policy declarations change slowly and high durability link load weak status flags of adjacency easier to resolve so kinda in the middle app designers decide where to put nib. can even roll ya own transactional thing / dynamoish DHT

mainly interact with NIB return openflow events and stuff with network state mgmt map swtich database to NIB entries any southbound interface

Flow policies and forwarding behavior

(Partitioning the flow-processing state requires that all controllers be able to set up paths in the network, end to end. Therefore, each Onix instance needs to know the location of all end-points as well as the link state of the network. However, it is not particularly important that this information be strongly consistent between controllers. At worst, a flow is routed to an old location of the host over a failed link, which is impossible to avoid during network element failures. It is also unnecessary for the link state to be persistent, since this information is obtained dynamically. Therefore, the controllers can use the DHT for storing link-state, which allows tens of thousands of updates per second (see Section 7).)

user rules

Notifications

data change notification

4 Proposed SDN extensions

There are two main representations of state in the systems we have discussed. There is a coordination service ONOS that maintains the naming and other information that must be stored by the controllers. Also, there is a representaion of the graph topology that each controller must learn from the physical switches under its domain. Then, to generate the full topology of the network, these subnetwork topologies must be assembled in the full network view.

Consistency, durability, isolation etc requirements for each element of data

4.1 Isolation

In both ONOS, the graph topology information stored by the controllers is eventually consistent. The ONOS paper suggests that this is "good enough" for traditional netowrking, and presumably good enough for their applications [4].

Similarly, with OpenDaylight, the graph topology of each controller is generated by BGP and disseminated to the remaining controllers. To generate the full nework view, these sub-network topologies are passed to the master controller-node of the network [1]. However, these network change events are assumed to be rare, and can only gurarantee eventual consistency over the topology link-state graph [12].

However, I would argue that this can lead to problems with advanced SDN applications, primarily with network virtualization, since effective netowork virtualzation must be able to reliably isolate resources.

Consider a network configuration with two groups that must be kept separate—Imagine that the engineering and accunting machines of an organization are on the same SDN network, and we want no packets to flow between the Engineering Virtual Network Segement and the Accounting VNS. In this case, imagine the following occurs:

Event 1 Alice connects to a controller in the Engineering VNS.

Event 2 Alice connects to a controller in the Accounting VNS.

Event 3 Alice disconnects from the Engineering VNS controller.

Instead, with eventual consistency over the topology, the following ordering of events could be observed after those events:

Event 1 Alice connects to a controller in the Engineering VNS.

Event 3 Alice disconnects from the Engineering VNS controller.

Event 2 Alice connects to a controller in the Accounting VNS.

Which the virtualization service would deem valid after Event 3.

Therefore, before it is possble for the network virtualizaton application to observe after the disconnection event that Alice is disconnected from the Engineerring VNS. In reality, however, Alice is connected to both the Engineering and Accounting VNS's. Clearly, in this case, some security measures could be installed to prevent such attacks on the network such as requiring authenticiton or ACL's—But it does illustrate that only eventual consistency leaves something to be desired when virtualzing a network. Such problems compound when we try to build larger scale geo-distrubuted virtualized networks, since correlated network splits and latency issues are compounded.

To enforce the global invariants required of the VNS isolaton, a simple solution is to simply require strong consistency over the link-state graph views [12]. This could be implemented by having a Paxos-like service that requires majority consensus to make link-state changes. Then, installed isolation rules could be guaranteed over the network view. This is not an unreasonable consideration—ONOS already maintains global coordination state about the in controller-switch relationships in ZooKeeper. Such a coordination service is not included in the OpenDaylight network virtualization system. So adding this would increase the system complexity and reduce the efficency of the network.

What's more problematic is that the CAP theorem [5] shows that we cannot have consistency over the whole virtual network as well as complete availibility and network partition tolerance. This means that to gain strong consistency over the global network view and gain the ability to enforce the global invariants required for isolation, we need to sacrifice avalibility or partition tolerance.

(Could using a shared-log approach to causal ordering guarantees be useful? Perhaps we could use this to enforce strong consistency in each controller's sub-network and define dependencies between them.)

4.2 Fault tolerance

ONOS maintains a ZooKeeper

Is this a global invariant that any switch cannot be part of the South and North Korea VNS's? If so, supporting causal consistency of the graph operations wouldn't do much to prevent such an attack on the system.

Different netowrk elements only work with their own data.

The shared-log model of ZooKeeper could be very useful in the case of nework separation. If the durable store were maintained at each VNS, then inventory managment could be logically segmented into

4.3 Sharing of data (Avalibility)

5 Conclusion

References

[1] Opendaylight network virtualization (onv).

[2] A. CLEMM, J. MEDVED, R. V. T. T. N. B. H. A. A data model for network topologies.

[3] BALAKRISHNAN, M., MALKHI, D., WOBBER, T., WU, M., PRABHAKARAN, V., WEI, M., DAVIS, J. D., RAO, S., ZOU, T., AND ZUCK, A. Tango: Distributed data structures over a shared log. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles* (New York, NY, USA, 2013), SOSP '13, ACM, pp. 325–340.

[4] BERDE, P., GEROLA, M., HART, J., HIGUCHI, Y., KOBAYASHI, M., KOIDE, T., LANTZ, B., O'CONNOR, B., RADOSLAVOV, P., SNOW, W., ET AL. Onos: towards an open, distributed sdn os. In *Proceedings of the third workshop on Hot topics in software defined networking* (2014), ACM, pp. 1–6.

[5] BREWER, E. A. Towards robust distributed systems. In *PODC* (2000), vol. 7.

[6] ENNS, R., BJORKLUND, M., AND SCHOENWAELDER, J. Netconf configuration protocol. *Network* (2011).

[7] FEAMSTER, N., REXFORD, J., AND ZEGURA, E. The road to sdn: an intellectual history of programmable networks. *ACM SIGCOMM Computer Communication Review 44*, 2 (2014), 87–98.

[8] HUNT, P., KONAR, M., JUNQUEIRA, F. P., AND REED, B. Zookeeper: Wait-free coordination for internet-scale systems. In *USENIX Annual Technical Conference* (2010), vol. 8, p. 9.

[9] JAIN, R., AND PAUL, S. Network virtualization and software defined networking for cloud computing: a survey. *Communications Magazine, IEEE 51*, 11 (2013), 24–31.

[10] JAIN, R., AND PAUL, S. Network virtualization and software defined networking for cloud computing: a survey. *Communications Magazine, IEEE 51*, 11 (November 2013), 24–31.

[11] KOPONEN, T., CASADO, M., GUDE, N., STRIBLING, J., POUTIEVSKI, L., ZHU, M., RAMANATHAN, R., IWATA, Y., INOUE, H., HAMA, T., ET AL. Onix: A distributed control platform for large-scale production networks. In *OSDI* (2010), vol. 10, pp. 1–6.

[12] LAPUKHOV, P. Centralized routing control in bgp networks using link-state abstraction.

[13] M. BJORKLUND, E. Yang - a data modeling language for the network configuration protocol (netconf). *Internet Engineering Task Force (IETF)* (2010).

[14] MCKEOWN, N., ANDERSON, T., BALAKRISHNAN, H., PARULKAR, G., PETERSON, L., REXFORD, J., SHENKER, S., AND TURNER, J. Openflow: enabling innovation in campus networks. *ACM SIGCOMM Computer Communication Review 38*, 2 (2008), 69–74.

[15] MEDVED, J., TKACIK, A., VARGA, R., AND GRAY, K. Opendaylight: Towards a model-driven sdn controller architecture. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2014 IEEE 15th International Symposium on a* (June 2014), pp. 1–6.

[16] ROSENBLUM, M., AND OUSTERHOUT, J. K. The design and implementation of a log-structured file

system. *ACM Trans. Comput. Syst. 10*, 1 (Feb. 1992), 26–52.

[17] TERRY, D. B., PRABHAKARAN, V., KOTLA, R., BALAKRISHNAN, M., AGUILERA, M. K., AND ABU-LIBDEH, H. Consistency-based service level agreements for cloud storage. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles* (New York, NY, USA, 2013), SOSP '13, ACM, pp. 309–324.