Running BGP in Data Centers at Scale

Anubhavnidhi Abhashkumar, Kausik Subramanian, Alexey Andreyev, Hyojeong Kim, Nanda Kishore Salem, Jingyi Yang, Petr Lapukhov, Aditya Akella, Hongyi Zeng

University of Wisconsin – Madison, Facebook
April 2021
Motivation
Facebook Data Center Growth
Facebook Data Center Growth

Tremendous user growth

500 million in 2010

2.5 billion in 2020
Facebook Data Center Growth

Tremendous user growth

500 million in 2010

2.5 billion in 2020

Our data center networks have grown to meet with user demand

Source: FBOSS: Building Switch Software at Scale, SIGCOMM 18
Data Center Routing Requirements
Data Center Routing Requirements

Scalable design to support quick growth
Data Center Routing Requirements

Scalable design to support quick growth

Reliability, expecting failures to be common
Data Center Routing Requirements

Scalable design to support quick growth

Reliability, expecting failures to be common

Operational efficiency to innovate rapidly while providing production reliability
BGP as the DC routing protocol
BGP as the DC routing protocol

BGP (Border Gateway Protocol) is chosen for scalability, extensive policy control, and long-term stability.
BGP as the DC routing protocol

BGP (Border Gateway Protocol) is chosen for scalability, extensive policy control, and long-term stability.

BGP on vendor devices helped grow our networks rapidly initially, compared to developing an SDN routing solution from scratch.
BGP Challenges
BGP Challenges

Long convergence, routing instabilities and misconfigurations in the Internet
BGP Challenges

Long convergence, routing instabilities and misconfigurations in the Internet

Problematic in the DC as well due to frequent failures and maintenance events
BGP Challenges

Long convergence, routing instabilities and misconfigurations in the Internet

Problematic in the DC as well due to frequent failures and maintenance events

Vendor BGP software not able to support evolving routing requirements fast
Our Solution
Our Solution

Scalable BGP routing design adhering to the principles of configuration uniformity and operational simplicity
Our Solution

Scalable BGP routing design adhering to the principles of configuration uniformity and operational simplicity

Routing policies to achieve reliability of the network under failures
Our Solution

Scalable BGP routing design adhering to the principles of configuration uniformity and operational simplicity

Routing policies to achieve reliability of the network under failures

In-house BGP agent coupled with automated testing and incremental deployment framework to achieve operational efficiency
DC Routing Design with BGP
Scalable Network Topology

Spine Plane 1

Spine Plane 4

Spine Switches (SSW)

Fabric Switches (FSW)

Rack Switches (RSW)

Server Pod 1

Server Pod N

Server
Scalable Network Topology

- Spine Plane 1
- Spine Plane 4
- Spine Switches (SSW)
- Fabric Switches (FSW)
- Rack Switches (RSW)
- Server Pod 1
- Server Pod N

Network components:
- Scalable Network Topology
- Spine Switches (SSW)
- Fabric Switches (FSW)
- Rack Switches (RSW)
- Server Pod 1
- Server Pod N
Scalable Network Topology

- **Spine Plane 1**
- **Spine Plane 4**
- **Spine Switches (SSW)**
- **Fabric Switches (FSW)**
- **Rack Switches (RSW)**
- **Server Pod 1**
- **Server Pod N**

The diagram illustrates a scalable network topology with Spine Switches connecting to Spine Planes, which in turn connect to Fabric Switches and Rack Switches, forming a network with multiple server pods.
Scalable Network Topology

Spine Plane 1

Spine Plane 4

Spine Switches (SSW)

Fabric Switches (FSW)

Rack Switches (RSW)

Server Pod 1

Server Pod N

Server
Reusable BGP ASN Allocation

Server Pod 65101

Confed AS: 65101

Spine Plane 65001

Plane AS: 65001
Route Summarization

Control route scale
Route Summarization

Control route scale

• Minimize HW FIB size requirements
Route Summarization

Control route scale

• Minimize HW FIB size requirements
• Reduce control plane processing
Route Summarization

Control route scale

• Minimize HW FIB size requirements
• Reduce control plane processing

Example of Production routes, reflecting topology hierarchy
Route Summarization

Control route scale

- Minimize HW FIB size requirements
- Reduce control plane processing

Example of Production routes, reflecting topology hierarchy

Server IP addresses
face:b00c::1:0/128
Route Summarization

Control route scale

• Minimize HW FIB size requirements
• Reduce control plane processing

Example of Production routes, reflecting topology hierarchy
Route Summarization

Control route scale

- Minimize HW FIB size requirements
- Reduce control plane processing

Example of Production routes, reflecting topology hierarchy
DC Policy Goals

In the Internet, BGP polices implement business relationships.
DC Policy Goals

In the Internet, BGP polices implement business relationships.

<table>
<thead>
<tr>
<th>Goal</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>Enforce route propagation scopes. Predefine backup paths for failure.</td>
</tr>
</tbody>
</table>
## DC Policy Goals

In the Internet, BGP polices implement business relationships.

<table>
<thead>
<tr>
<th>Goal</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>Enforce route propagation scopes. Predefine backup paths for failure.</td>
</tr>
<tr>
<td>Scalability</td>
<td>Enforce route summarization. Avoid backup path explosion</td>
</tr>
</tbody>
</table>
## DC Policy Goals

In the Internet, BGP polices implement business relationships.

<table>
<thead>
<tr>
<th>Goal</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>Enforce route propagation scopes. Predefine backup paths for failure.</td>
</tr>
<tr>
<td>Scalability</td>
<td>Enforce route summarization. Avoid backup path explosion</td>
</tr>
<tr>
<td>Maintainability</td>
<td>Isolate/remediate problematic nodes without disrupting traffic</td>
</tr>
</tbody>
</table>
## DC Policy Goals

In the Internet, BGP polices implement business relationships.

<table>
<thead>
<tr>
<th>Goal</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>Enforce route propagation scopes. Predefine backup paths for failure.</td>
</tr>
<tr>
<td>Scalability</td>
<td>Enforce route summarization. Avoid backup path explosion</td>
</tr>
<tr>
<td>Maintainability</td>
<td>Isolate/remediate problematic nodes without disrupting traffic</td>
</tr>
<tr>
<td>Service reachability</td>
<td>Avoid service disruptions when instances of services are added/removed/migrated</td>
</tr>
</tbody>
</table>
Reliability

Predefine backup paths for link failure

During BGP Route Convergence
Reliability

Predefine backup paths for link failure

RSW1’s rack_prefix: 
-> RSW1 (best)

**Action:** add tag ‘rack_prefix’

During BGP Route Convergence
Reliability

Predefine backup paths for link failure

RSW1’s rack_prefix:
- RSW1 (best)
- RSW2--FSW2--RSW1 (backup)

During BGP Route Convergence

match: ‘rack_prefix’
action: add tag ‘backup_path’

match: ‘backup_path’
action: allow

action: add tag ‘rack_prefix’

RSW1’s rack_prefix:
- RSW1 (best)
- RSW2--FSW2--RSW1 (backup)
Reliability

Predefine backup paths for link failure

Traffic flow on link failure
Reliability

Predefine backup paths for link failure

Traffic flow on link failure
Scalability

Enforce route summarization and avoid backup path explosion

During BGP Route Convergence

match: ‘backup_path’
action: add tag
‘completed_backup_path’

completed_backup_path

RSW1

FSW2

RSW2

Routes with tag ‘completed_backup_path’ won’t be propagated to other devices.
Scalability

Enforce route summarization and avoid backup path explosion

Convergence happens locally within a server pod

During BGP Route Convergence
Operational Experience
Built Operational Pipeline

Support BGP like any other SW, enabling fast incremental updates!
Conclusion
Conclusion

Scalable BGP-based routing design: Reusable ASN allocation and route summarization coupled with our scalable topology design

Reliable under failures: Routing policies such as predefined backup path

Operationally Efficient: Automated testing and incremental deployment framework for our in-house BGP implementation
facebook