Data Center Reference Architectures

Manish Karir

Merit Network Inc.

Outline

- Background
- Generalized data center architecture
- Data center variations
- Factors affecting data design
- Generalizing scale and workload
- Discussion

Goal

- To distill from mailing list and other discussions a common architecture for use by ARMD working group
- Data center designs are highly customized based on assumptions regarding use, traffic, platforms, and desired performance
  - Causes problems where no on knows what particular scenario someone else has in mind
  - Problems in terminology
  - Difficulties in identifying problems
- *Data center design is a result of balancing various trade-offs to minimize *your* particular issues. If done well the result is that in your design there are no issues left that matter to you*

A Generalized 3 Layer Data Center Network Architecture

```
 +-----+-----+     +-----+-----+
 |   Core0   |     |    Core1  |      Core
 +-----+-----+     +-----+-----+
       /    \        /       /
      /      \----------\   /
     /    /---------/    \ /
   +-------+           +------+
 +/------+ |         +/-----+ |
 | Aggr11| + --------|AggrN1| +      Aggregation Layer
 +---+---+/          +------+/
   /     \            /      \
  /       \          /        \
+---+    +---+      +---+     +---+
|T11|... |T1x|      |T21|     |T2y|  Access Layer
+---+    +---+      +---+     +---+
|   |    |   |      |   |     |   |
+---+    +---+      +---+     +---+
|   |... |   |      |   |     |   |
+---+    +---+      +---+     +---+  Server racks
|   |... |   |      |   |     |   |
+---+    +---+      +---+     +---+
|   |... |   |      |   |     |   |
+---+    +---+      +---+     +---+
```

Generalized Data Center Components

- Servers - Racks of equipment that require network access
- Access Layer – Equipment directly connected to servers either in the same rack (ToR) or at the end of the row (EoR)
- Aggregation Layer – Equipment that aggregates access layer devices to provide connectivity among Access Layer domains
- Core – Equipment that interconnects multiple aggregation layer devices either within a data center or across geographic locations with outside world
- Note: No mention of Layer 2 / Layer 3 boundaries

Data Center Design Variations/ Topology

- Layer 2/Layer 3 boundary can vary greatly from one data center to next
- Layer 3 to access switches – Each rack enclosure/row is a single Layer 2 domain – extensive virtualization may result in potentially large L2 domain
- Layer 3 to aggregation switches – most common middle of the road solution – flexibility in L2 domain size and VM mobility
- Layer 3 in the core only – large multi-site data centers – good for applications that require high VM mobility
- Overlays – L2 or L3 can be used to move the L2/L3 boundary around

Factors affecting Data Center Design

- Data center purpose and anticipated traffic patterns:
  - Large virtualized web farm - high in/out traffic, low volume of local traffic
  - Large compute cluster – large volume of local traffic, little in/out traffic
  - Multi-tenant data center – customer traffic segregation requirements
- Potential complications of Virtualization:
  - Higher server densities
  - Additional VLANs for HA beacons/migrations

Impact of Data Center Design on L2 protocols

- L2/L3 boundary is the critical pain point
- Crossing L3/L2 boundary involves ARP/ND processing - the larger the L2 the larger the potential load
- Bi-directional traffic crossing multiple VLANs internal to the data center can cause twice the load as ARP/ND is involved in both directions
- Dual-stack servers in a data center have both ARP and ND traffic for the same number of devices

Problem of Generalizing

- Generalizing topology is not enough
- Need to account for different traffic patterns
- Need to account for differences in L2/L3 boundary designs
- Need to account for virtualization densities
- Need to account for scale variations

Defining Typical Topology

```
 +-----+-----+     +-----+-----+
 |   Core0   |     |    Core1  |      Core
 +-----+-----+     +-----+-----+
       /    \        /       /
      /      \----------\   /
     /    /---------/    \ /
   +-------+           +------+
 +/------+ |         +/-----+ |
 | Aggr11| + --------|AggrN1| +      Aggregation Layer
 +---+---+/          +------+/
   /     \            /      \
  /       \          /        \
+---+    +---+      +---+     +---+
|T11|... |T1x|      |T21|     |T2y|  Access Layer
+---+    +---+      +---+     +---+
|   |    |   |      |   |     |   |
+---+    +---+      +---+     +---+
|   |... |   |      |   |     |   |
+---+    +---+      +---+     +---+  Server racks
|   |... |   |      |   |     |   |
+---+    +---+      +---+     +---+
|   |... |   |      |   |     |   |
+---+    +---+      +---+     +---+
```

Defining Typical Scale

- What should the typical scale of the data model be?
  - Basic model (as suggested at previous mtg):
    - Container based data center – 8-20 racks
    - 4 dell chassis/rack = 64 blades/rack – assume dual socket hex-core per blade?
    - Results in 768 cores per rack oversubscribe 2:1 = 1500 VMs/ rack -> 12K-30K VM per container
    - 1 ToR per rack, 2 aggregation switches, 2 core switches- outbound traffic
- Small, medium, large, x-large categorization:
  - Small: <10K, Medium: 10-20K, Large: 30-50K, x-Large >50K+

Defining Typical Workload

- Different Workloads:
  - Web Farm/Data Serving Usage: A handful of VLANs heavy traffic in/out pattern little cross VLAN traffic except to data store and databases
  - Compute Farm: A handful of VLANs heavy cross VLAN traffic
  - Multi-tenant virtual colo: Large number of VLANs, little cross VLAN traffic except control plane VLANs
- Generalize workload by assigning a fraction of total VM population that requires ARP/ND lookups at any given time.
- Example:
  - web farm/data serving usage might result in 5% of VMs that require ARP/ND lookups at any given time = 1500 messages/second
  - Compute Farm might result in 1% of VMs that require ARP/ND lookups at any given time = 300 ARP/ND message/second
- Focus away from applications and traffic load as they vary drastically from one data center to the next, focus on percentage of VMs that require ARP/ND service at any given moment. – You figure out what that percentage is for your application

Conclusions

- ARMD should develop and use a generic data center physical topology in problem description to abstract away different variations but that is not enough:
- ARMD might consider the use of container scale data center in determining if performance and scale issues exist and are significant enough in a typical scenario – alternative is to use small, medium, and large with some concrete definitions (e.g.small < 10K, large > 50K)
- ARMD might consider the use of "fraction of nodes that require ARP/ND service" as a metric in attempting to describe performance or scaling issues
- Generalize to avoid talking about specific scenarios

Discussion

- ARMD mailing list discussion has been varied and difficult to frame in a single framework
- Data centers means different things to different people
- Our goal was to try to collate all feedback and discussion into a common umbrella from the perspective of impact on protocols such as ARP/ND
- Is there a better way to structure list feedback into a useful framework given that the goal of ARMD is to determine potential performance bottlenecks that might emerge in large scale data centers.
- Current solutions use a variety of design alternatives to avoid any bottlenecks but perhaps additional solutions might be possible if L2 performance bottlenecks were tackled more directly