Yandex

Yandex DC Design Evolution

Dmitry Afanasiev, fl0w@yandex-team.ru

Network Architect

Yandex

- We're rather typical MSDC
- Monthly user audience of over 90 million worldwide.
- ~Services: search, music, video, cloud storage, news, weather, maps, traffic, email, ads ...
- Several DCs in Russia and abroad + peering and traffic exchange points + MPLS backbone to connect them
- Workloads: interactive request processing, object storage, map-reduce-like, data streaming, large scale replication, machine learning...

What we need?

- Cheap and abundant bandwidth
- Scalable forwarding with minimal state
- Multitenancy / network virtualization - for historical reasons
- Efficient resource pooling
- InterDC traffic engineering
- Stable routing system and reasonably fast convergence
- Function chaining: load balancing, FW, etc.
- Automation at scale

What we don’t need

We are trying to keep design really simple. Don’t need many functions often perceived as desireable:

- L2 (but nodes can use overlays)
- VM mobility
  - In scale-out applications nodes coming and going is a norm, no need to move them around while preserving state and identity
  - VM mobility increases complexity as it depends on other features
- Multicast
- We don't have too many changes in topology

Our Infrastructure

- About 100k servers and growing fast
- Mostly IPv6 internally, need to serve external IPv4 - tunnels
- 2 WANs - for interactive and bulk traffic
- 10GE to the server, Nx100GE inter-switch in DC, Nx100GE WAN, looking at 25GE to the server
- Eliminated L2 in new DC designs -> L3 to the ToR (VPN or multi-VRF), smaller L3 domains in some locations (L3/port and eventually to server)
- Eliminated multi-hop multicast
- /64 per server (for virtualization, also removes most ND from ToRs)
- Still need FW (technical debt), moving to hosts (HBF), some tricks with host part of IPv6 addr

Our Infrastructure (2)

- Need to support 10k+ nodes clusters, recent DC design scales to 25-30k nodes
- Clos fabrics, 2 spine layers
  - modular spines but also looking at fixed boxes (need radix >= 64 to stay with 2 spine layers)
- 1k-4k ECMP routes per DC, 4x-16x ECMP, can be 32x in future
  - one of the limits is power
  - another is ECMP table(s) size with MPLS on ToRs - need separate rewrite entries for each next hop, can be improved with global labels

Our Infrastructure (3)

- BGP in DC fabrics - 2 flavors
  - iBGP and per-hop RR+NHS, similar to RFC 7938
  - iBGP with off-path route servers (some modular routers don't work well with 100s of BGP sessions)
- OSPF + TE in WANs, considering SR-TE in future
- DC borders are starting to look like small fabrics

Challenges and Future Work

- Diagnostics, measurements and monitoring - need to look at fast processes and transient events - buffering, convergence
- Balance between reducing control traffic and aggregating routing information and disseminating enough information to achieve
  - granular enough traffic manipulation - drain, steering, TE between DCs
  - adjusting load balancing in presence of failures - need to look beyond 1 hop even in highly regular topologies
- Combining programmability/centralized control with local reaction to failures
  - BGP is really useful here - a lot can be done with controller that looks just like RR from protocol PoV but implements more complex logic

Yandex

Questions?