Scaling the Edge: Approaches to Application Load Balancing

A Panel Discussion

Key Questions

What approaches to application load balancing provide the best value, for whom, and when?

- The ALB space offers a wide array of strategies and options. This leads to design indecision and dependence on general-purpose solutions. *In most cases this is fine*, until…
- As applications scale, the shortcomings of existing solutions become apparent if the solution is not carefully chosen from the outset.
- However, different solutions **DO** make sense at different scaling points. A reasonable solution at launch stage may no longer perform as the service approaches “internet scale”.

The current landscape?

A matrix of options:

	Appliance-based	Software/Cloud
Commercial	A10 AX series, F5 BIGIP, Netscaler	Riverbed Stingray, SW editions of appliance solutions
Open Source	n/a	LVS/keepalived, Varnish, mod_proxy

A number of base “Styles”:

- Layer 7 (Application Proxy)
- Layer 4 Inline
- Layer 4 DSR (L2 and L3)

A Load Balancing Primer

Layer 7 Load Balancing

- The ALB is an application proxy
- Can handle decryption/SSL offload, application-specific request routing, connection coalescing
- More “high-touch” and CPU intensive than alternatives
- Supported by all major LB vendors, multiple open-source software solutions (Varnish, Apache mod_proxy, …) but not limited to HTTP/HTTPS services
- CDNs are a flavor of L7 load balancing as a service.

A Load Balancing Primer Pt. 2

Layer 4 Load Balancing

- The ALB is a TCP/UDP router/NAT device
- Application agnostic, but often application-level health checking is desired
- Less resource intensive (bring your own SSL)
- If Direct Server Return is set up, LB only has to process inbound traffic for even better scalability
- Supported by all major LB vendors, although DSR implementations may vary. OSS solutions as well (LVS)

Scaling to multiple endpoints?

Eventually, one VIP isn’t going to be enough. What now?

- DNS-based (GLB software, Neustar/Dynect, etc.)
- Active-Active HA configurations
- ECMP balancing—takes advantage of upstream flow-hashing
- Anycast (not just for UDP anymore?)
- Different approaches have different failover scenarios.

Today’s Panel:

Moderator:

Chris Woodfield, Twitter

Panelists:

Leslie Carr, Wikimedia

Jamie Dahl, Yahoo!

Mike Thompson, A10 Networks

Sridhar Devarapalli, Citrix Systems

Questions for Panel:

- How do app and network designs inform LB scaling strategies? What are the risks and rewards of different approaches?
- What application services does the ALB layer need to provide to your application?

Questions for Panel (Cont’d):

- What are the drivers for multi tenancy and administrative partitioning features in current load balancing products? How does this affect the scaling challenge?
- At what scale does automation resources become a requirement? What is the role of automation in your success?

Audience Questions?