IPV6 DUAL-STACK DEPLOYMENT FOR THE DISTRIBUTED COMPUTING CENTER

A. Kotliar , V. Kotliar[a]

Institute for High Energy Physics named by A.A. Logunov of National Research Center "Kurchatov Institute", Nauki Square 1, Protvino, Moscow region, Russia, 142281

E-mail: [a] viktor.kotliar@ihep.ru

Computing center of the Institute for High Energy Physics in Protvino provides computing and storage resources for various HEP experiments (Atlas, CMS, Alice, LHCb) and currently operates more than 150 working nodes with around 3000 cores and provides near 2.5 PB of disk space. All these resources are connected through two 10 GB/s links to LHCONE and other research networks. IHEP computing center has IPv4 address space limited to one C-sized network and all computing nodes working behind the NAT that has some drawbacks for production use. To optimize routing, switching and to get higher network throughput for data transfer the IPv6 dual-stack was deployed on the computing farm. This work shows the full cycle of the real IPv6 dual-stack deployment from zero to production.

Keywords: IPv6, GRID, distributed computing

Anna Kotliar, Viktor Kotliar

1. Introduction

Computing center of the Institute for High Energy Physics in Protvino provides computing facilities and storage resources for various HEP experiments like Atlas, CMS, Alice, LHCb. It operates with more than 150 working nodes (around 3000 cores) and it provides near 2.5 PB of disk space. Main computations are done by using the distributed computing paradigm and GRID [1]. To connect all resources there are two 10 GB/s links to LHCONE and other research networks. IHEP uses its own AS2643 autonomous system to operate networks. Through many years IHEP computing cluster had IPv4 address space limited to one C-sized network and all computing nodes worked behind the NAT. Such configuration worked well on a small sized cluster but with the computing farm growing some drawbacks for production use were discovered. Main of them are IPv4 addresses costs and increasing the number of NAT devices and NAT traffic. To optimize routing, switching and to get higher network throughput for big data transfer (main traffic for IHEP) the IPv6 dual-stack was deployed at the computing center and on the computing farm. Due to used hardware and software stacks, it was done in seamless way without any disturb for the computations and operations of the computing center. The full cycle of the real IPv6 dual-stack deployment from zero to production is described below.

2. IPv6 features to know

As soon as IPv6 is a completely different protocol from IPv4 there are some basic things and features which you need to know before starting the deployment [2]. IPv4 addresses are sized at 32-bits and expressed in decimal octets like:

- 192.168.1.1 - address;
- 192.168.1.1/255.255.255.0 - address and netmask;
- 192.168.1.1/24 - address and prefix length.

IPv6 addresses are sized at 128-bits in hexadecimal words and they have opposite to IPv4 meaning for prefix length:

- 2001:0678:07d8:0000:0000:0000:0000:0001 – address;
- 2001:0678:07d8:0000:0000:0000:0000:0001/64 - address and prefix length;
- 2001:678:7d8::1/64 - short form with skipped zeros.

There is a completely different way for dynamically assigning IPv6 addresses to devices. Basics things to understand and configure are following:

- IPv6 Router Advertisements
  - IPv6 routers can be configured to send advertisements, both periodically and on request, with the following information: the IPv6 prefix/prefix length in use on this link; the IPv6 address of the router; various ‘flags’ (or hints) which tell hosts how to behave;
- IPv6 Dynamic Address Assignment
  - hosts can auto-generate their own addresses using SLAAC (State-Less Auto-Address Configuration) or via DHCPv6 through use of the ‘A’ or ‘M’ flags;
- Device Unique IDentifier [DUID]
  - DHCPv6 does not necessarily identify machines by MAC address and DUIDs can be one of: Link-layer address plus time [DUID-LLT], vendor-assigned UID based on Enterprise Number [DUID-EN], link-layer address [DUID-LL]UUID-based DUID [DUID-UUID];
  - Not all DHCPv6 clients use the same DUID type by default.

One more thing to take into consideration with IPv6 deployment is a packer fragmentation. IPv4 supports packet fragmentation and an intermediate router can break a large IPv4 packet into several smaller ones prior to forwarding. In opposite, IPv6 does not support packet fragmentation. This is not permitted in IPv6 and a router drops the packet and responds to the sender with ICMPv6 “Message Too Big” along with the size of packet that it will accept and forward. The sender receives the response and transmits a smaller packet. This repeats ad-infinitum along the traffic path until packet successfully reaches the destination. Packet fragmentation hides broken network paths.

3. Step by step deployment of IPv6

The real deployment of IPv6 in the production environment is a complex task. It consists of many sub tasks, which are difficult to know before the real deployment begins. There are some recommendations to do test setup on preproduction or testing environment. The article, in contrast, describes deployment in a working environment with minimal services downtimes for IPv6 dual-stack in stateful mode.

3.1 Planning and preparations

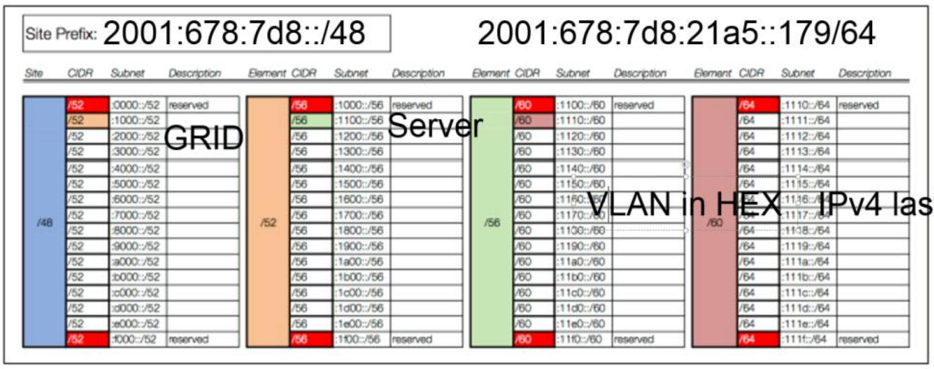
Before starting the setup IPv6, it should be clear understanding what is the reason for the deployment and the main goal. Then there is a checklist for the readiness of the current environment to IPv6:

- Network devices support for IPv6;
- Used operating system support for IPv6;
- Used system applications and services support for IPv6 (DNS, mail, DHCP, https, netflow…);
- Used application software support for IPv6 (decide on default IP protocol for dual-stack);
- Possible security issues and firewalls readiness for IPv6.

Moreover, important decision should be made on IPv6 addressing. There are many recommendations and books written about this and it depends on the used addressing scheme for IPv4 and network topology. Usually IPv6 prefix /64 is used for IPv4 with prefix /24 and the addressing depends on used VLANs, IPv4 and services roles. Here is an example [fig. 1] for IHEP addressing plan based on recommendations from Tom Coffeen’s book [3]. For the site we use /48 IPv6 network and split this space across different levels:

1. Network type - GRID, IHEP general, special devices;
2. Role of the device – server, working node for GRID;
3. VLAN ID from the L2 network layer;
4. Last IPv4 byte (it has meaning because IPv4 addresses at IHEP based on VLANs).

So for IPv4 GRID storage server 194.190.165.179/24 (se0003.m45.ihep.su) there is IPv6 address 2001:678:7d8:21a5::179/64 where 2001:678:7d8::/48 is an address space for IHEP.



Site Prefix: 2001:678:7d8::/48

2001:678:7d8:21a5::179/64

Site	CIDR	Subnet	Description	Element	CIDR	Subnet	Description	Element	CIDR	Subnet	Description	Element	CIDR	Subnet	Description
/48	/52	:0000::/52	reserved	/52	/56	:1000::/56	reserved	/56	/60	:1100::/60	reserved	/60	/64	:1110::/64	reserved
	/52	:1000::/52			/56	:1100::/56			/60	:1110::/60			/64	:1111::/64	
	/52	:2000::/52			/56	:1200::/56			/60	:1120::/60			/64	:1112::/64	
	/52	:3000::/52			/56	:1300::/56			/60	:1130::/60			/64	:1113::/64	
	/52	:4000::/52			/56	:1400::/56			/60	:1140::/60			/64	:1114::/64	
	/52	:5000::/52			/56	:1500::/56			/60	:1150::/60			/64	:1115::/64	
	/52	:6000::/52			/56	:1600::/56			/60	:1160::/60			/64	:1116::/64	
	/52	:7000::/52			/56	:1700::/56			/60	:1170::/60			/64	:1117::/64	
	/52	:8000::/52			/56	:1800::/56			/60	:1180::/60			/64	:1118::/64	
	/52	:9000::/52			/56	:1900::/56			/60	:1190::/60			/64	:1119::/64	
	/52	:a000::/52			/56	:1a00::/56			/60	:11a0::/60			/64	:111a::/64	
	/52	:b000::/52			/56	:1b00::/56			/60	:11b0::/60			/64	:111b::/64	
	/52	:c000::/52			/56	:1c00::/56			/60	:11c0::/60			/64	:111c::/64	
	/52	:d000::/52			/56	:1d00::/56			/60	:11d0::/60			/64	:111d::/64	
	/52	:e000::/52			/56	:1e00::/56			/60	:11e0::/60			/64	:111e::/64	
	/52	:f000::/52	reserved		/56	:1f00::/56	reserved		/60	:11f0::/60	reserved		/64	:111f::/64	reserved

Figure 1. IHEP IPv6 addressing plan

3.2 One computer setup

It is possible to start just with one dedicated computer in the network with IPv6 dual-stack. Such setup touches the whole network stack with local access and Internet necessaries (routing, DNS). To setup one computer:

1. Get IPv6 address space from provider or LIR [4] (PI address space recommended);
2. Configure internal network infrastructure and connect one host to IPv6 network (ping -6 should work internally between routers and IPv6 testing computer);
3. Setup external routing for IPv6 in RIPE DB and with providers (ping -6 should work with Internet in both directions);
4. Setup DNS to resolve IPv6 addresses for direct and back zones (ping -6 by name should work from Internet).

3.3 Monitoring, accounting and firewall

As soon as there is already one computer with IPv6 and full stack deployed it is possible to configure and adjust monitoring, accounting and firewall for the new protocol. At IHEP, netflow is used for traffic accounting and pmacct [5] installed as a collector for netflow v9 from routers. Cacti plug-in added to show and analyze pmacct netflow DB based on postgresql. Also flox tool for pmacct adapted for the new setup. To monitor IPv6 traffic on network device interfaces some changes were made to the local monitoring system based on mrtg. To address security basic IPv6 rules were added on the IHEP border firewall and some of IPv4 restrictions were adjusted. As soon as at IHEP installed next generation firewall based on applications and not on IP addresses, initial IPv6 setup and configuration almost repeats IPv4 rules and is very simplified.

3.4 Adding the cluster for distributed computing

After IPv6 stack started to work in general, the distributed computing cluster was reconfigured to use it. At IHEP, it consists of storage nodes, compute nodes, infrastructure and management servers. Dual-stack stateful IPv6 configuration is used on all of them. For distributed storage servers IPv6 static configuration is applied through cluster management system [6] and static DNS records are added to the cluster DNS zones. The used storage software (dCache, xrootd) is configured to work with IPv4 and IPv6 where IPv6 is made as preferable protocol for data transfers to and from outside. For computer nodes was added special DHCPv6 server [7] that is allowed to use the same MAC to IPv6 binding as it was done for IPv4. It needs to be mentioned that such behavior is not a part of IPv6 protocol standart but worked well for the computing cluster and allowed fully control IP addresses assignment for both IPv4 and IPv6 in the same manner. In addition, a network topology was changed to allow direct connections for IPv6 from nodes to outside where, alongside, IPv4 still works behind NAT. As the last step all computer nodes were configured to prefer IPv6 to IPv4 where it is possible. For management and infrastructure servers static IPv6 configuration was applied and IPv6 protocol was set as default for all services after carefully inspecting server by server and service by service.

Following the last configuration step the firewall rules were updated on the main IHEP firewall and on all the servers and nodes connected to IPv6 network.

4. Conclusion

Current setup for IPv6 deployment is shown on figure 2. Only a computing cluster is fully connected to IPv6 in dual-stack deployment and there is still a lot of work that should be done in connecting IHEP generic public network to IPv6.

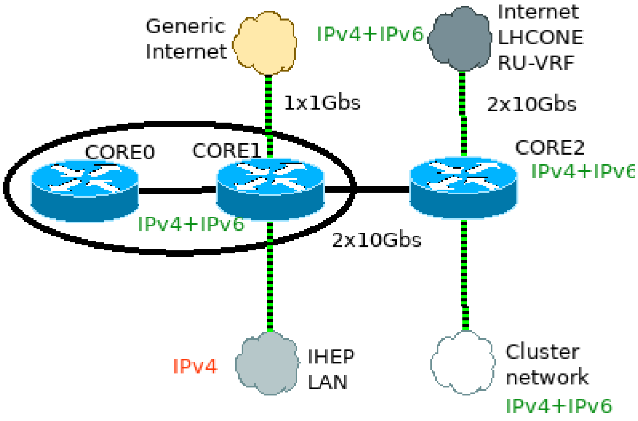


Figure 2. IHEP IPv6 deployment

The usage of IPv6 with end-to-end hosts connectivity to all devices allowed to increase cluster efficiency for data transfers and efficiency of using CPU on the compute nodes. It allows to grow a compute center by adding more storage and compute nodes by removing IPv4 addresses restrictions (too expensive addresses) for Internet connections. All new big network setups for projects at IHEP are going to use IPv6 as primary protocol in IPv6 only or in IPv6 with IPv4 dual-stack deployments meanwhile migrating IHEP GPN from IPv4 only to IPv4 with IPv6.

References

[1] Kotliar V. IHEP cluster for Grid and distributed computing // CEUR Workshop Proceedings. February 2017: Vol. 1787.- pp. 312-316

[2] Hagen S. IPv6 Essentials, 3rd Edition // O'Reilly Media, Inc. June 2014. ISBN: 9781449319212

[3] Coffeen T. IPv6 Address Planning // O'Reilly Media, Inc. November 2014. ISBN: 9781491902769

[4] IPv6 Address Allocation and Assignment Policy [ripe-738]. Available at: https://www.ripe.net/publications/docs/ripe-738 (accessed 22.09.2021)

[5] pmacct a small set of multi-purpose passive network monitoring tools [pmact github]. Available at: https://github.com/pmacct/pmacct (accessed 22.09.2021)

[6] Ezhova, V., Kotliar, A., Kotliar, V., Popova, E. Multicomponent cluster management system for the computing center at IHEP // CEUR Workshop Proceedings. 2018: Vol. 2267. - pp. 37-43.

[7] dhcpy6d MAC address aware DHCPv6 server [dhcpy6d github]. Available at: https://github.com/HenriWahl/dhcpy6d (accessed 22.09.2021)