OpenVMS Cluster over IP

Nilakantan Mahadevan
nilakantan.mahadevan@hp.com

OpenVMS Technical Journal V18

Table of contents

Introduction

This article describes OpenVMS cluster over IP (aka IP cluster interconnect), one of the major milestone project for OpenVMS. The article highlights the solution and techniques adopted toward using IP as a transport for OpenVMS cluster[2] and the customer advantage using OpenVMS cluster over IP solution.

Customers deploy OpenVMS clusters within their data centers and between their disaster tolerant (DT) sites. They use LAN bridging and Extended LAN technology provided by IP switches for LAN-based cluster traffic. Non availability of extended LAN services and restricted scope for LAN based protocols coupled with market factors have made it necessary for using IP as a transport for OpenVMS clusters.

This new support is added in a manner that helps discover new nodes in cluster using IP, smooth and transparent migration without a cluster reboot. It provides interoperability with servers running prior versions which are LAN based, ability to dynamically load balance, and failover between available connections/path and usage of path with least cost. Some of the features are unique to OpenVMS clustering solution. The ability to add support for IP as transport after overcoming the challenges of retaining the "gold" features of OpenVMS clusters is a major milestone.

Importance of OpenVMS clusters

OpenVMS clustering technology, known as "gold standard" for DT, hailed for its security and reliability, is adopted by HP major customers for whom downtime is never an option. OpenVMS clusters demonstrated its DT capabilities by failing over in less than 13 seconds without any manual intervention in the disaster tolerant demonstration by HP[3]. OpenVMS clusters are qualified to operate safely over a distance of 500 miles[4]. Customers can take advantage of the DT capabilities provided by OpenVMS clusters for their mission-critical environment.

Need for IP Cluster Interconnect (IPCI)

Technical background of the problem

Network interconnect is one of the preferred interconnect for OpenVMS cluster. Cluster protocol (system communication architecture [SCA aka SCS]) over LAN is provided by port emulator driver (PE driver), refer Figure 1. It implements the Network interconnect system Communication Architecture (NISCA) in OpenVMS server. Several major customers use OpenVMS clusters within their data centers and between their DT sites.

All nodes of an OpenVMS cluster should be in same VLAN for cluster communication. Customers use LAN bridging and Extended LAN technology provided by IP switches in order to transmit LAN NISCS traffic across sites. In the past, there have been few instances where network outages in data centers have impacted OpenVMS clusters. The cluster instability was diagnosed as only occurring during periods of heavy IP usage[1]. On a closer examination, it was discovered that network switch during higher loads give priority to IP traffic than SCS traffic resulting in SCS packets being dropped. In addition, router CPU utilization remains high when router is specially configured with ability to transport SCS traffic.

[1] "Customers deploy dedicated network infrastructures for Cluster traffic separating from IP traffic to prevent such issues"
[2] HP OpenVMS Systems
[3] Disaster-proof solutions from HP
[4] Add Software Product Description

Restricted scope of non-IP based protocols

Network managers found that extended LANs had stability and security issues that required a lot of skill, specialized knowledge, and also effort. Today, multiprotocol routing is very much out of fashion, along with bridging LAN protocols beyond the workgroup level. IP has become the de-facto industry standard.

Currently, majority of network managers are only now thoroughly trained on and comfortable with managing IP networks. As a result, the skills and knowledge required to manage extended LANs have become very scarce. The scalability, security, and network management complexity issues with extended LANs have led network managers to restrict the scope of non-IP protocols.

This, in turn, has led to corporate IT policies for network infrastructure management that strongly discourage or prohibit the use of non-IP protocols. Configuring an extended LAN beyond a workgroup or single switch requires policy exceptions and rare-extended LAN management skills.

Midrange/high-end switch hardware design centre is now IP

The emergence of IP as the de-facto standard for computer networks has led switch vendors to optimize their hardware designs for IP rather than LAN by:

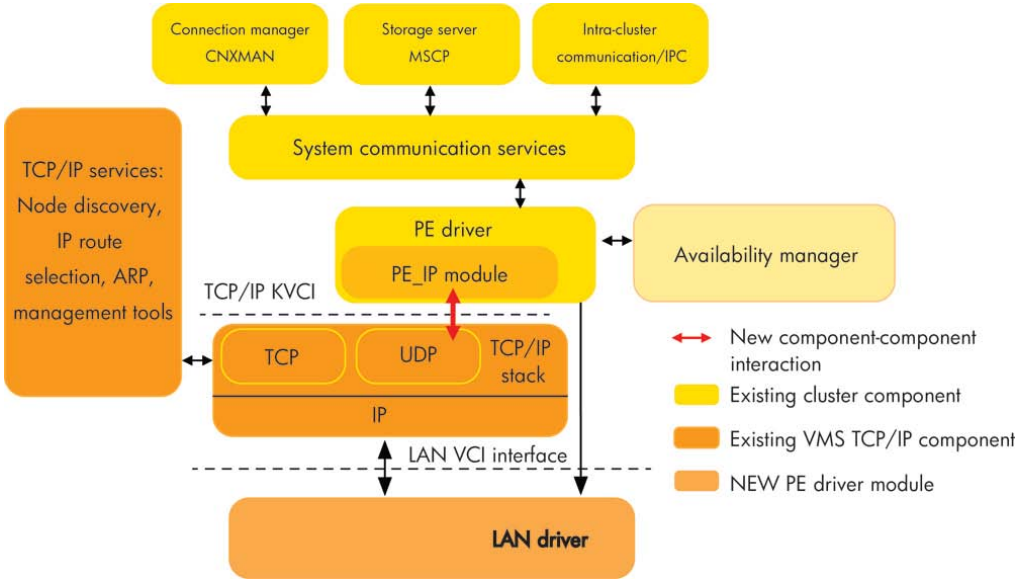
- Giving IP absolute hardware priority for IP routing over LAN bridging
- Providing additional license fee for routing non-IP packets

Business and market motivation

The market drivers for putting cluster traffic on IP network are:

- Non-availability of LAN bridging from all switch and Telco vendors
- Corporate policies restricting scope of non-IP protocols
- Specialized hardware resource costs to set up multisite DT cluster with LAN bridging
- Lower cost of high-speed IP services

Figure 1. OpenVMS Cluster Communication Stack



Solution

Alternatives investigated

A high-level OpenVMS technical team had initial discussion on the various solutions to the above problem. The various alternatives are:

- Software-based IP clusters alternatives:
  - PE driver over raw IP
  - PE driver over User Datagram Protocol (UDP)
  - TCP SCA port driver (aka: TCP driver)
- Hardware alternative:
  - RNIC SCA port driver

The above alternatives were analyzed from different dimensions like project scope, Manageability, data integrity, data encryption, load distribution, CPU cost, responsiveness to losses.

Alternatives analysis summary

All alternatives require these TCP/IP changes:

- TCP/IP BG driver and the TCP/IP services loadable executive images (execlet) must be loaded early in the life of system, just after the LAN drivers are loaded:
- A minimum TCP/IP "boot time database" must also be loaded for TCP/IP stack to initialize during boot:
  - Node name (optional, SCSNODE could be used)
  - IP address
  - Interface name
  - Preferred gateway
  - Multicast group (derived from cluster group #)
- TCP/IP management tools need to be extended to maintain the node specific 'boot time database'
- IPCI security model needs to be developed and provide new ways to turn security use on/off for cluster connections

Based on the investigation and analysis, it was decided that the best solution for the stated problem was PE driver over UDP.

Goals

The following were set as the high-level goals for the project:

- Support IP as a cluster interconnects in addition to LAN Supported Platform Architectures:
  This feature must be supported on Integrity and Alpha systems. It must also support Integrity-Alpha interoperability using IP. Common source modules permit both architectures to benefit from the new functionality.
- Facilitate a mixed LAN/IP environment
- Provide interoperability with the LAN-based cluster communication in mixed LAN/IP environment.
- Enable prior version interoperability:
  Must offer LAN interoperability with the target and subsequent releases, Migration and Warranty support OpenVMS Versions.

- Provide a security model.
- Improve performance and availability:
  IP support must not result in significant application performance degradation and availability. CPU cost, response time under load, and data throughput is the measures of performance.
- Provide management control of:
  - VMS cluster protocol utilization of IP source-destination paths and local IP enabled LAN adapters
  - PE driver use of only IP, only LAN, or both IP and LAN
  - The UDP Port used for IPCI
  - Support for management using HP Availability Manager Product

An implicit and important goal is to retain the OpenVMS Cluster flavor and strength as it is, and to make this feature integrated in a transparent manner. This was the guiding principle as we made key design choices and progress through the various stages in the project.

PE driver over UDP

As PE driver was the best solution over UDP option, two parts of this solution had to be developed independently and integrated as a final solution.

PE driver over UDP support

This enhances PE driver to use the UDP protocol. Some of the features of this solution include

- The IP UDP service has the same packet delivery characteristics as 802 LANs. PE driver implements the transport layer of NISCA, which has inbuilt delay probing, retransmission, reliable delivery for sequenced messages, implement datagram service, and also variable buffer size for block transfers for I/O suitable for cluster traffic.
- The Kernel VMS Communication Interface (KVCI) is a highly efficient interface of the HP OpenVMS TCP/IP services stack. It is a variant of the VCI interface, which PE driver uses to communicate with OpenVMS LAN drivers. PE driver interfaces with UDP as if it were a LAN device. The current model involves a virtual circuit between two nodes consisting of LAN channels for cluster communication. With IPCI, a virtual circuit between two nodes consists of LAN and IP channels for cluster communication.
- Only the lowest layer of PE driver needs to be extended to support UDP. The PE driver changes are transparent to PE driver's upper layers. PE driver's use of UDP does not require changes to the TCP/IP stack other than early initialization of the stack during the boot. The result is a minimum risk, well bounded set of changes to the PE driver.
- Providing management interface ability to control and configure IP interfaces to PE driver

TCP/IP services boot time loading and initialization

In order to ensure that cluster communications are available in an IP-only network environment, it is essential to have TCP/IP stack loaded when the cluster formation starts. This also retains the existing functionality of cluster formation of OpenVMS clusters. This consists of the following:

- Add HP OpenVMS TCP/IP services execlet (kernel modules) to the list of images to be loaded early in boot.
- Enable the TCP/IP initialization routines to read a configuration file that supplies the key parameters (like routing information) necessary to minimally initialize the TCP/IP services

Node discovery and path keep-alive model

PE driver uses 802 multicast for discovering LAN connected nodes and for keep-alive (hello) packets. IP multicast maps 1:1 onto the existing LAN discovery, so it has been selected as the preferred node discovery mechanism. This auto-discovery mechanism based on IP multicast requires very less manual intervention when setting up or reconfiguring a cluster. It is also used for keep-alive mechanism. The default IP multicast address is selected from the administratively scoped IP multicast address range of 239.242.x.y. The last two octets x and y are generated based on the cluster group number. For example, if the cluster group number is 1985, then multicast address is calculated as follows. Cluster group number is 1985 and calculates as follows:

x= 1985/256

y= 1985 — (256 *x)

The ability to override the default multicast address by a unique address for their environment is another flexibility given to customers. An alternative and additional unicast mechanism for node discovery is also provided since not all IP networks are configured to allow multicast address.

Security model

The model for providing security in an IP and WAN environment is:

- Isolating IP subnets conveying cluster communications subnets from the WAN-IP environment
- Ensuring that cluster communications over insecure WAN links is encrypted and authenticated

Standard firewall technique would be applicable to IPCI. Customers whose intranet spans multiple sites must (and normally do) use secure private links or inter-site encryption such as VPN type tunneling between the firewalls. Firewall and dedicated network for IPCI along with encryption can be adopted for internal security. Implementation also has validation of IP source address and privileged port numbers to protect against denial of service attacks. In future IPsec could be enabled at boot time, thus all IP-based cluster communications would be secure without any other external units. System administrators are provided the flexibility to specify time to live (TTL) IP multicast packets. It specifies the number of hops allowed for IP multicast packets and increases control span.

Major project parts

Initial investigation was completed during the summer of 2007. Once the high-level way forward strategy was determined, the project was essentially driven in parallel streams with appropriate integration phase.

- Cluster
  - PE driver over UDP
  - Early loading of TCP/IP exec lets
  - Support for boot time database
  - SCA Control Program (SCACP) management support[5]
  - System Dump Analyzer (SDA) support (PE SDA extensions)
- TCP/IP service boot time support
- Availability manager support

A working prototype was developed that could successfully use IP to transmit SCS packets between geographically distributed nodes (for example, India and USA). PE driver over UDP had major functionalities built into it successfully and passed major functionality test with this initial prototype. It has successfully demonstrated usage of only IP transport for cluster communication in labs during spring 2007.

[5] SCACP is used to monitor and manage cluster communications.

In order to test PE driver over UDP, cluster is first formed with traditional LAN and later force communication to failover to IP transport channels by disabling LAN channels. At this point, TCP/IP early initialization work was currently under progress by simulating an early boot environment for TCP/IP startup. Cluster formation using IP was not possible because TCP/IP early loading work was under progress. However, the ability to form the cluster with LAN initially and then switch to IP gave an initial advantage for the project. It was possible for the testing team to start their testing cycles.

Support for boot time database and early loading of execlets

Once the core protocol support was built, the next phase included the support for boot time database. OpenVMS cluster relied on the system parameter VAXCLUSTER and NISCS_LOAD_PEA0 for loading the necessary cluster execlets. A new "sysgen" parameter NISCS_USE_UDP was introduced for enabling IPCI.

During system startup, LAN drivers are loaded first, and then the PE driver was loaded. PE driver opens a port with LAN driver and sends the broadcast packet in a LAN cluster. In case of IPCI, (that is, NISCS_USE_UDP is set) once LAN drivers are loaded, TCP/IP execlets are loaded followed by PE driver. It opens the ports with both LAN and TCP/IP stack and sends the multicast packets. The flowcharts given in Appendix A and B describe the cluster formation in LAN and IPCI.

The boot time database consists of two different files:

SYS$SYSTEM: PE$IP_CONFIG.DAT

- Generated by CLUSTER_CONFIG_LAN.COM
- Read early in the boot sequence
- Provide information to PE driver
- Can be common throughout cluster
- Remote node IP address should be present in local node PE$IP_CONFIG.DAT in order to allow remote node join the cluster using IP unicast
- Best practice for IP unicast: Include all IP address and have one copy of the file throughout the cluster
- "$MC SCACP reload" to be used to refresh IP unicast list on a live system

SYS$SYSTEM: TCPIP$CLUSTER.DAT

- Generates TCPIP$CONFIG which is invoked by CLUSTER_CONFIG_LAN.COM
- Read early in the boot sequence
- Provides information to PE driver to use the correct TCP/IP interface (WE0 OR WE1) for cluster traffic
- Provides information to TCP/IP stack to initialize the interface with IP address and default route

One of the challenges faced in the support of boot time database is non-availability of regular file system to load and read these files. The technique adopted was to use the primitive file system support available with OpenVMS in order to load the contents of the files into system memory. Here again, teams were working in parallel. The configuration utilities (CLUSTER_CONFIG_LAN.COM and TCPIP$CONFIG.COM) were being changed to create these files. At the same time, the usage of primitive file system support was incorporated and tested using template files created by hand.

Cluster formation in an IP only network

Once PE driver over UDP and early loading of TCP/IP services pieces were complete, the next phase was to integrate and test the formation of cluster in an IP-only environment. This was accomplished by forming a cluster between India (Bangalore) and Nashua (USA). Though the geographical distance was beyond the supported distance, the formation of cluster was a significant milestone. This was a key milestone for the project in summer 2008. The cluster formation between nodes in two different continents was successfully demonstrated in OpenVMS boot camp that was held on May 2008 in Nashua.

Cluster formation in IP-only environment is described as flowchart in Appendix B of this article. The modified process for IP-only environment is highlighted in the flowchart.

Availability manager support

HP Availability Manager is a system management tool, which monitors OpenVMS nodes in a cluster. The ability to view and manage OpenVMS cluster using IP for communication was built by having the necessary support built into both PE driver and Availability Manager.

Performance

A key challenge is to keep up performance levels when using IP for cluster traffic. Focus on performance aspects involves analyzing and benchmarking the performance of PE driver over UDP against traditional LAN and to identify areas of improvement. For long distance cluster, the speed-of-light delay when dealing with geographically distant sites quickly becomes the dominant factor for latency, overshadowing any delays associated with traversing the IP stacks within the cluster member hosts.

There may be a tradeoff between the latency of failover and steady-state performance. We consider localization of cluster traffic in the normal (non-failover) case as vital to optimizing system performance as the distance between sites is stretched to supported limits (and well beyond). OpenVMS engineering conducted in house test by having 32 nodes in a cluster with IP for cluster communication as proof point to showcase the ability to scale to a higher number of nodes. The tests were conducted in winter 2008. The target for initial release was to have acceptable performance and to improve the same in subsequent releases.

Testing and validation

OpenVMS Cluster test manager CTM (Cluster Test Manager) was extensively used during the testing phase. The emphasis was given on several aspects, including data integrity, node failover detection, and cluster reformation. Several cycles of tests were conducted in order to help ensure data integrity even under extreme disturbances caused to cluster simulating a real-world scenario.

Tests were conducted with Mass Storage Control Protocol (MSCP)[6] serving of disks using IP cluster interconnect. The MSCP disks were used for system disk shadow set successfully. A large intercontinental cluster consisting of nodes in Asia, Europe, Australia, and America which also included an HP virtual machine guest as member of the cluster was formed using Cluster over IP (Appendix C). This was demonstrated in HP tech forum at Las Vegas, USA, in 2009.

Comparative study with other clustering technology

There are other clustering solutions from HP which includes HP Serviceguard for HP-UX and other clustering software as well as from IBM and SUN. IBM High Availability Cluster Multi processing (HACMP) uses UDP data grams to send its heartbeat messages. HP Serviceguard clustering software uses TCP/IP network services for reliable communication, while VERITAS cluster has a special transport as replacement for IP stack. Some of the salient features described below are unique to IPCI software. IPCI successfully retains the gold features of VMS clustering software.

Documentation

OpenVMS is known for the excellent documentation describing the features and functionalities for users. Several updates were done to cluster manuals to include the necessary information about cluster over IP. The documents were updated with clear-cut examples and usage scenarios to enable users to benefit from the feature.

[6] MSCP protocol is used to serve devices to the another node in the cluster

Salient features of IPCI solution

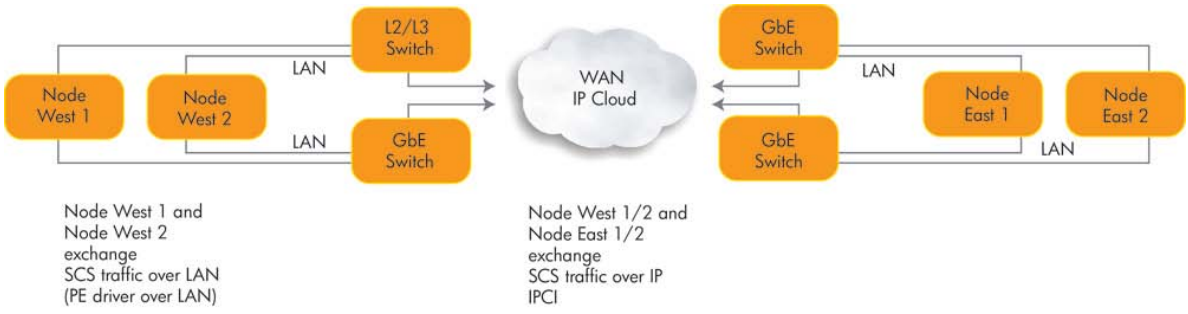
- Discover nodes beyond LAN segment in an IP-only network.
- Add and remove nodes dynamically without any disruption to cluster; this functionality exists currently with OpenVMS clusters and continues with IPCI.
- Perform cluster state transitions and failover with minimal latency
- Retain OpenVMS Cluster feature of rolling upgrades to the new version without a cluster reboot
- Facilitate interoperability with servers running prior versions of OpenVMS clusters, which are LAN-based
- Load balances dynamically between all the available healthy interfaces
- Provide PE driver delay probing that helps to reduce latency in IP network
- Detect and remove faulty interface from the set of healthy interfaces transparent to higher layers
- Provide IPv6 support in future releases with minimal effort
- Enable the software to operate on low-level LAN-based protocol and IP protocol
- Load TCP/IP stack using a minimal infrastructure provided by operating system early in boot.

Customer advantage with IPCI

- Use OpenVMS clusters when only IP services are provided by Telco Vendors. Refer Figure 2.
- Low infrastructural and management costs
- Reduce operational costs
- Leverage the benefits from the improvements in IP technology

The complete implementation with all the features was rolled out as a part of OpenVMS 8.4 in June 2010. This feature enables customers to have long distance clusters using IP backbone and without specialized hardware for extended LAN. Refer Figure 2.

Figure 2. Disaster Tolerant Clusters with IPCI

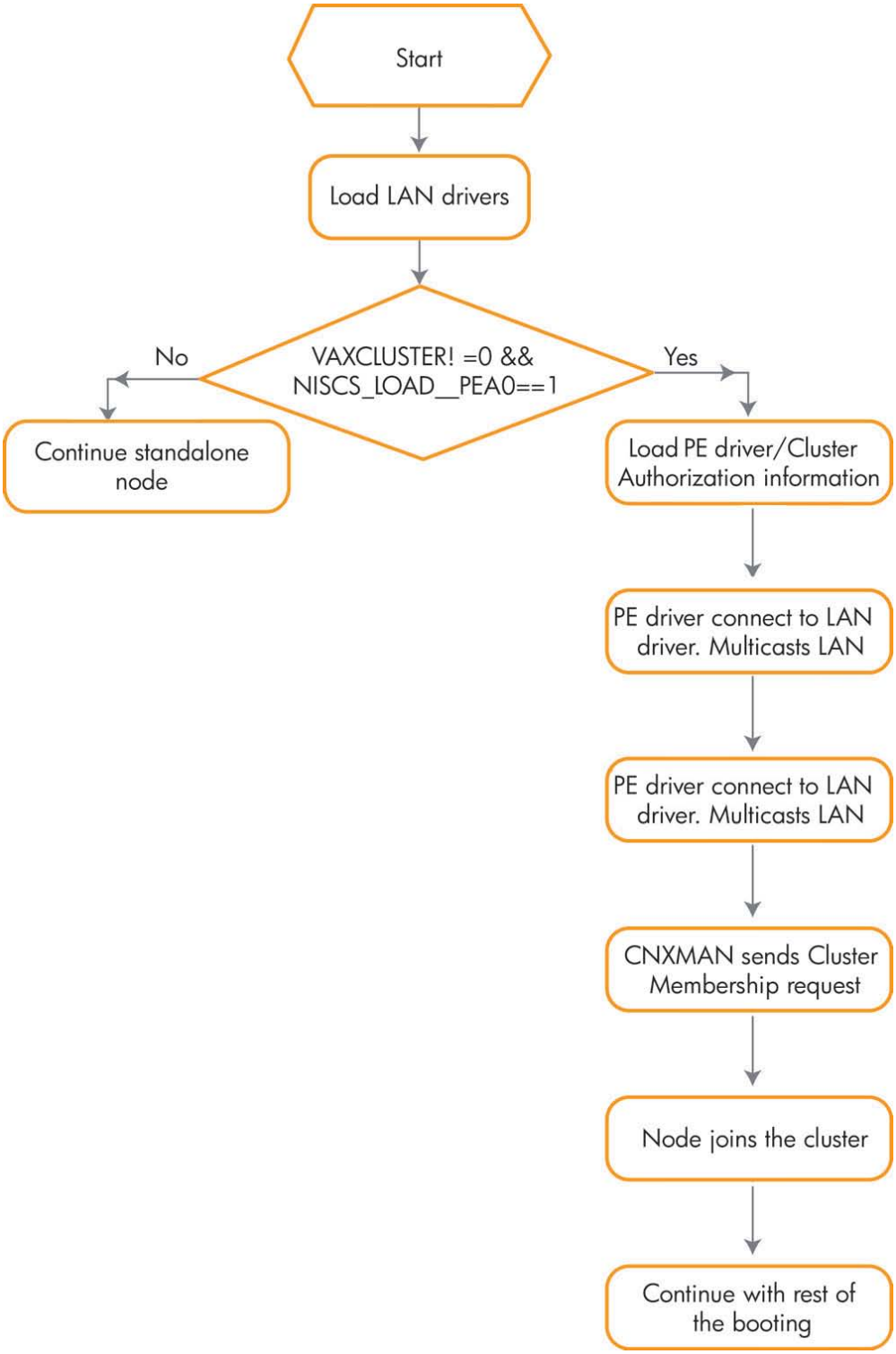


"...DT customers required a redundant pair of network connection.

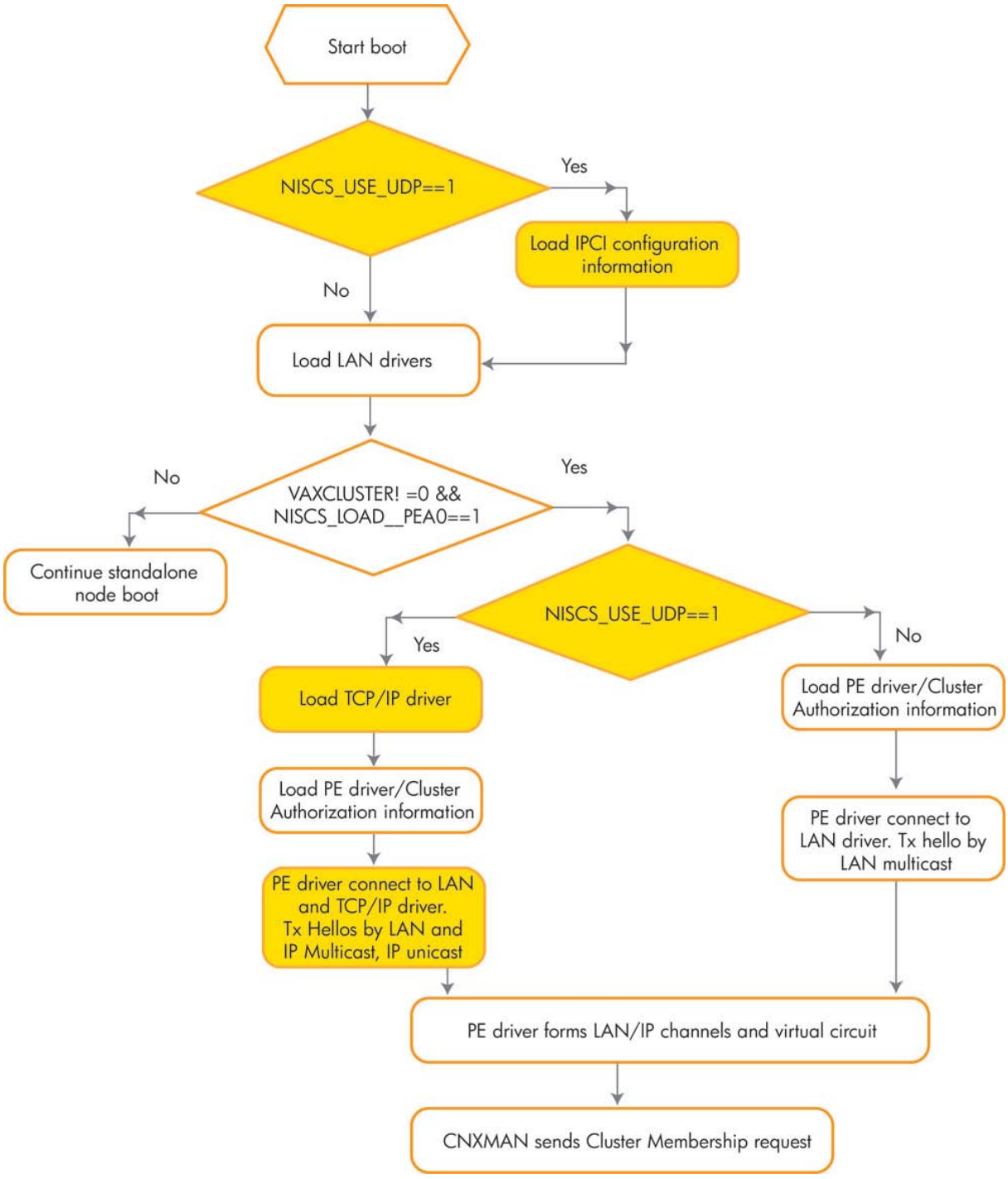
Acknowledgements

OpenVMS cluster over IP was a key project executed and delivered with thanks to sincere work of many people within the OpenVMS Engineering. The article is incomplete without acknowledging the efforts of all the individuals made this project a success. HP OpenVMS management and OpenVMS ambassadors have been supportive during the entire lifecycle of the project.

Appendix A: Cluster formation in LAN

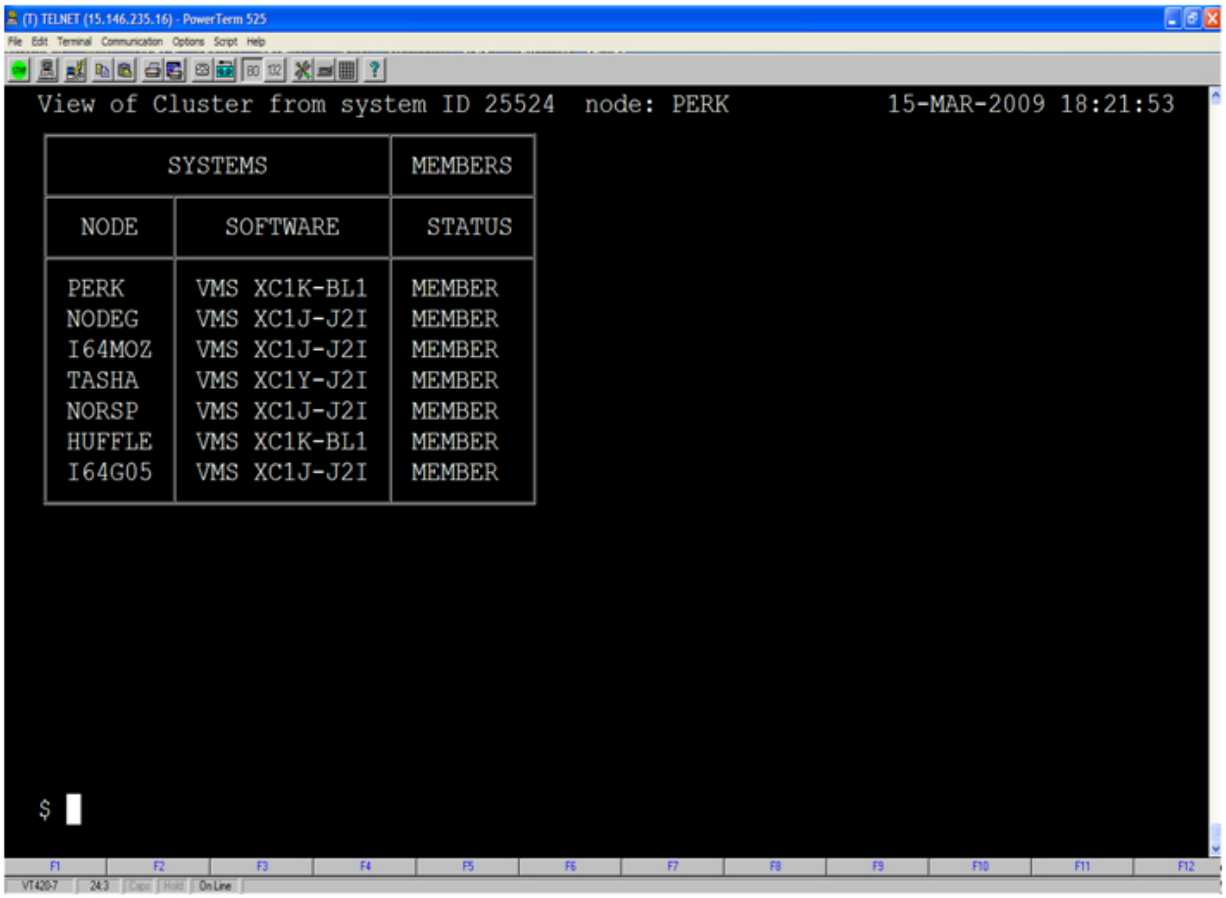


Appendix B: Cluster formation in IP only environment (IPCI)



Appendix C: Cluster formed with nodes in different Continents (Americas, Asia, Europe, Australia)

SHOW CLUSTER



4AA4-0525ENW, Created April 2012