



ACCELERATE

outcomes with convergence

HP **TECHFORUM** 2010

CONVERGE

INNOVATE

TRANSFORM

Using an IP Network as an OpenVMS Cluster Interconnect

Keith Parris

Systems/Software Engineer, HP

June 23 2010





Background

- Local Area OpenVMS Clusters today require a OSI Model Layer 2 (Bridged) extended LAN connection between nodes
- OpenVMS Version 8.4 (with TCP/IP Services for OpenVMS Version 5.7) allows the option of using a standard IP network as a native OpenVMS Cluster Interconnect as an alternative to (or in addition to) a LAN



Background

- OpenVMS picks a Multicast LAN address for a cluster based on the Cluster Group Number
- This Multicast LAN address is of the form AB-00-04-01-xx-yy
 - "xx-yy" is based on the cluster group number plus an offset
- Multicast Hello packets are sent out from all LAN adapters enabled for cluster communications, every 1.5 to 3 seconds by default
- Each cluster member enables receipt of packets addressed to the Cluster MAC address, and uses the receipt of Hello packets to discover new nodes, discover new communications paths, track the reachability of a node via a given path, and select the optimal path(s) to be used



Background

- Forming a multi-site OpenVMS cluster implies one of the following:
 - Bridging a VLAN between sites using a bridge (or a router with bridging capability), forming an Extended LAN connection
 - Separate (Private) LAN link between sites, just for OpenVMS Cluster traffic
 - Dark fiber, or lambda over DWDM connection
 - Use of Multi-Protocol Layering System (MPLS), L2TPv3, or DLSW, to encapsulate LAN packets inside IP packets, which are then routed
 - High router CPU overhead
 - Use of Ethernet-over-IP hardware boxes to encapsulate LAN packets inside IP packets



Why Have Customers Asked for this Feature?

- Network groups within some companies refuse to bridge VLANs (IP segments) between different sites
 - IP network purists denigrate bridging because of problems LAN failures / misconfiguration can cause to IP networks (network loops, ARP broadcast storms, etc.)
- In some areas of the world, bridged LAN connections are unavailable between sites
- Bridge/routers tend to favor IP traffic during congestion, and may drop multicast traffic, causing cluster instability
 - Multicast Hello packets are used to track OpenVMS Cluster path presence and latency, for path detection and optimal path selection



DESIGN ALTERNATIVES CONSIDERED

– Create New IP Protocol for OpenVMS Clusters

- Define new protocol and document it through RFC process. More coding work involved. New protocol would be unfamiliar to network managers.

– Use TCP/IP Protocol

- TCP/IP protocol does packet fragmentation & reassembly, retransmission, etc. Packet fragmentation/reassembly not helpful. Some duplication of function with existing PEDRIVER functions (sequencing messages, retransmission). More host CPU overhead. Default timers might be unsuitable for cluster traffic.

– Use UDP/IP Protocol

- Simplest solution. UDP characteristics are closest to LAN packet delivery semantics. This alternative was chosen.

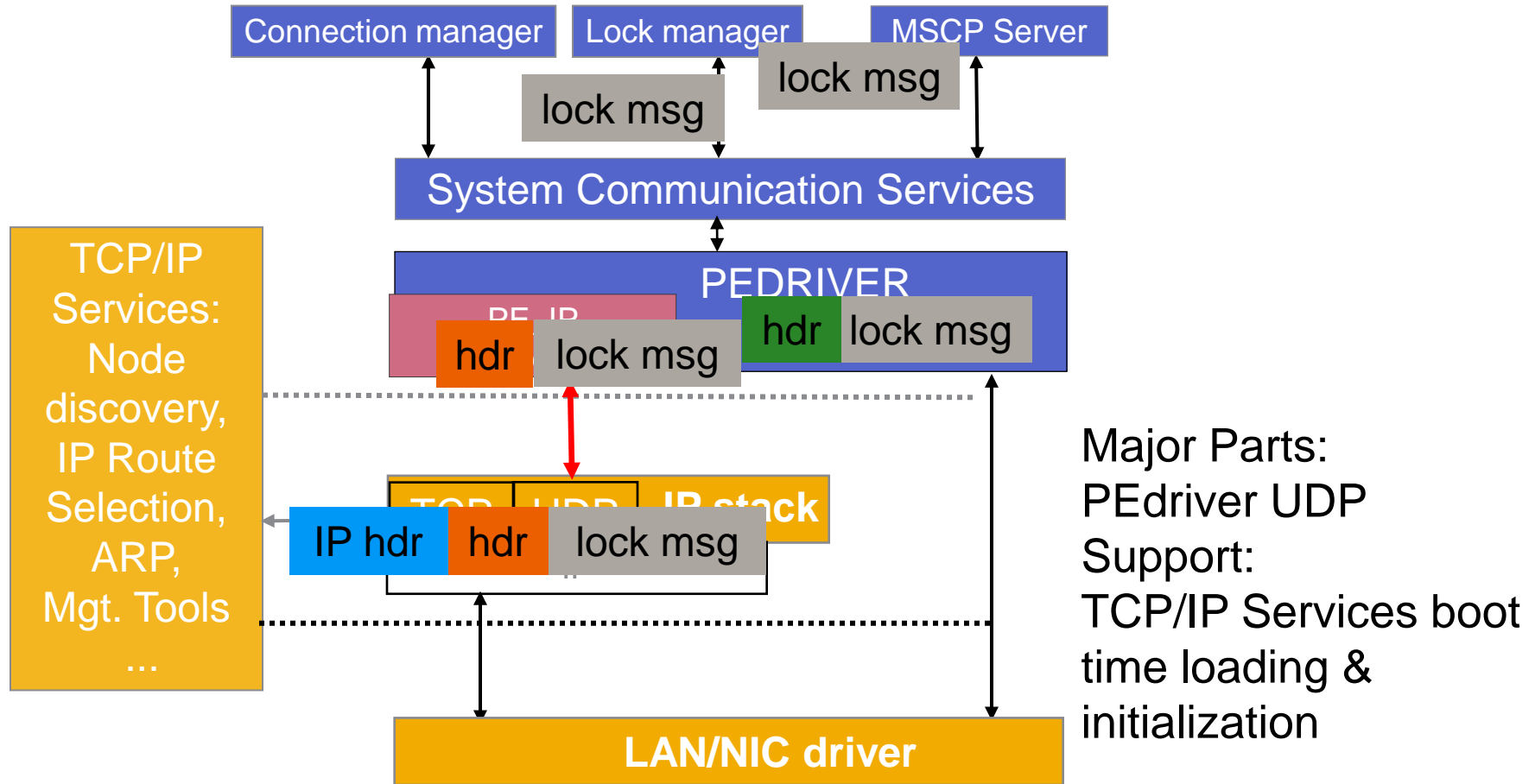


CHANGES REQUIRED TO SUPPORT IP AS A CLUSTER INTERCONNECT

- Load TCP/IP Software early in boot
- Driver-level interface between PEDRIVER and TCP/IP Software
 - Read IP configuration file SYS\$SYSTEM: PE\$IP_CONFIG.DAT early in boot
 - PEDRIVER can send/receive UDP packets
- Node discovery using either IP Multicast or IP Unicast packets
- CLUSTER_CONFIG[_LAN].COM
- TCPIP\$CONFIG.COM
- SCACP
- SDA
- Availability Manager



IPCI SOLUTION – PE DRIVER OVER UDP



Existing Cluster Components
 New PEDRIVER component

New Component-Component interaction
 Existing VMS/TCP/IP Components



SETTING UP CLUSTERS OVER IP

- OpenVMS Version 8.4 Installation / Upgrade Procedure
- CLUSTER_CONFIG_LAN.COM



CONFIGURATION FILES

- Files used for configuring Cluster with IP interconnect are:
 - SYS\$SYSTEM:PE\$IP_CONFIG.DAT and
 - SYS\$SYSTEM:TCPIP\$CLUSTER.DAT
- Loaded during the boot process
- Set up by CLUSTER_CONFIG[_LAN].COM



CONFIGURATION FILES: SYS\$SYSTEM:PE\$IP_CONFIG.DAT

- SYS\$SYSTEM:PE\$IP_CONFIG.DAT contains:
 - UDP port number to use for IPCI
 - IP unicast addresses of the nodes of the cluster
 - Optional IP multicast address for the cluster
- IP multicast messages are used for discovering a node within the same IP multicast domain. If IP Multicast is not available, or some nodes are in a different IP multicast domain, nodes can use IP unicast messages to join the cluster.
 - Administratively scoped multicast address space for IPv4 is the range 239.0.0.0 to 239.255.255.255.
- When IP unicast addresses are used for node discovery, a node is allowed to join the cluster only if its IP address is present in PE\$IP_CONFIG.DAT
- If changes are made to this text file manually, put them into effect using
 - \$MCR SCACP RELOAD
- PE\$IP_CONFIG.DAT can be common for all the nodes of a cluster.



CONFIGURATION FILES: SYS\$SYSTEM:PE\$IP_CONFIG.DAT

Sample PE\$IP_CONFIG.DAT file:

```
! CLUSTER_CONFIG_LAN creating for CHANGE operation on  
15-JUL-2008 15:23:56.05
```

```
multicast_address=239.242.7.193
```

```
ttl=32
```

```
udp_port=49152
```

```
unicast=10.0.2.3
```

```
unicast=10.0.2.2
```

```
unicast=10.0.1.2
```



CONFIGURATION FILES: SYS\$SYSTEM:TCPIP\$CLUSTER.DAT

- SYS\$SYSTEM:TCPIP\$CLUSTER.DAT contains:
 - IP interface to be used for IPCI
 - IP addresses on which cluster communication is enabled
 - TCP/IP route information
- SYS\$SYSTEM:TCPIP\$CLUSTER.DAT is unique for each node in a cluster



CONFIGURATION FILES: SYS\$SYSTEM:TCPIP\$CLUSTER.DAT

Sample TCPIP\$CLUSTER.DAT file:

```
interface=LE1,LLB0,10.0.1.2,255.0.0.0
```

```
default_route=10.0.1.1
```



NEW SYSGEN PARAMETERS

– NISCS_USE_UDP

- 1 = Enable use of IP as a cluster interconnect
- 0 = Disable use of IP As a cluster interconnect (Default)

– NISCS_USE_LAN

- 1 = Enable use of LAN as a cluster interconnect (Default)
- 0 = Disable use of LAN as a cluster interconnect

– NISCS_UDP_PORT

- Port number to use for cluster communications over UDP

– NISCS_UDP_PKTSZ

- Similar to NISCS_MAX_PKTSZ for LANs, but this applies to UDP traffic



NEW SCACP COMMANDS

\$ MCR SCACP

SCACP> RELOAD

SCACP> SHOW CHANNEL /IP

SCACP> SHOW CHANNEL /LANCHANNEL

SCACP> SHOW IP_INTERFACE

SCACP> SET IP_INTERFACE

SCACP> START IP_INTERFACE

SCACP> STOP IP_INTERFACE



NEW SDA COMMANDS

\$ ANALYZE/SYSTEM

SDA> PE IP_INTERFACE

SDA> PE LAN_INTERFACE



MYTHS

- Myth: If I'm using a bridged connection between sites today, I should plan on converting to IP as a cluster interconnect
- Fact: Unless the bridged LAN connection is being taken away, it is typically best to continue using it
 - LAN paths are preferred over IP paths by PEDRIVER by default
 - Adding IP as an alternate path could help hold the cluster together if the network gets congested.



MYTHS

- Myth: Longer-distance clusters can be built using IP as a cluster interconnect than with a bridged LAN connection
- Fact: Speed of light over the distance between sites causes packet latency, and the impact of this latency on application performance is the primary limitation on inter-site distance in OpenVMS clusters.



MYTHS

- Myth: Using IP as a cluster interconnect will provide better performance than a bridged LAN connection
- Facts:
 - Host CPU overhead is likely to be slightly higher using IP
 - Unless unusual bridging techniques or hardware (DLSW, L2TPv3, MPLS, Ethernet-over-IP) were in place, latency is likely to be about the same or perhaps even slightly higher over IP



PERFORMANCE TUNING RECOMMENDATIONS FOR IPCI

- Use Fast_Path to put TCP/IP Software, PEDRIVER, and LAN adapters used for cluster traffic on the same non-Primary CPU
 - If CPU saturation in interrupt state occurs on that one CPU, then interrupts will need to be spread across multiple CPUs



LIMITATIONS

- Only HP's TCP/IP Services for OpenVMS is supported
 - IPCI not available with other TCP/IP stacks at this time
- Only supported in OpenVMS Version 8.4 or above
 - No plans to back-port to 8.3 or 8.3-1H1
- No VAX support planned
 - Only Alpha and Integrity
- Initial release supports IPv4 only
 - No IPv6 support at this time



PROMISING NEW AREAS OF APPLICATION

- Hobbyist clusters over the Internet
- Quorum node at 3rd site in disaster-tolerant clusters



RESOURCES

- OpenVMS Documentation for Version 8.4 on HP OpenVMS website:
 - <http://www.hp.com/go/openvms/>
- White paper: "Guidelines to Configure Cluster over IP"
 - http://www.connect-community.org/resource/resmgr/library_whitepapers/cluster_over_ip_whitepaper.pdf
- Presentation "IP Cluster Interconnect (IPCI) aka OpenVMS Cluster over IP" by Nilakantan Mahadevan:
 - 2008: http://www.woertman.com/VMSTUD2008/OpenVMS_clusterimprovements_IPCI.ppt
 - 2009: http://www.connect-community.de/Events/OpenVMS2009/folien/07-OpenVMS_clusteroverIP.pdf



Q&A



SPEAKER CONTACT INFO

E-mail:

Keith.Parris@hp.com

Website:

<http://www2.openvms.org/kparris/>



NEXT STEPS

Visit the EXPO to learn more about OpenVMS 8.4 features
EXPO G3, South Convention Center Level 2

Session: Case Studies of DT & DR with OpenVMS, Thu. 10:30 AM,
Breakers G, South Convention Center Level 2

Discussion: OpenVMS Clusters, Thu. 12:00 Noon
Connections Cafe, Table 3, Mandalay Bay G, South Convention Center Level 2

Session: Using Shadowsets with More Than 3 Members, Thu. 1:30 PM,
Breakers C, South Convention Center Level 2

Discussion: Disaster Tolerance & DR with OpenVMS, Thu. 3:00 PM
Connections Cafe, Table 2, Mandalay Bay G, South Convention Center Level 2



OUTCOMES THAT MATTER.

