IP Cluster Interconnect (IPCI) aka OpenVMS Cluster over IP

Nilakantan Mahadevan

Nilakantan.mahadevan@hp.com

HP, OpenVMS Engineering

GET CONNECTED

People. Training. Technology.

Agenda

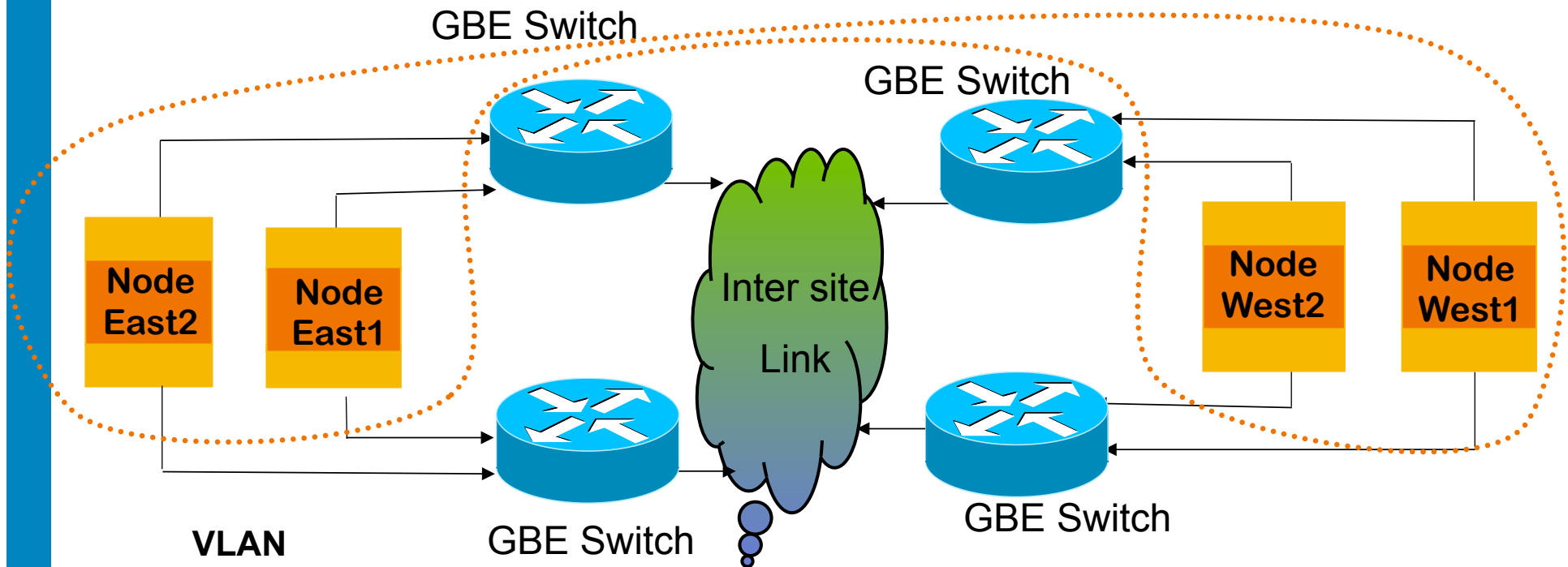- IP Cluster Interconnect (IPCI) aka OpenVMS cluster over IP

IP Cluster Interconnect (OpenVMS cluster over IP)

- Introduction to OpenVMS Clustering Technology

- Cluster communication architecture.

- IPCI technical overview

- IPCI Demonstration.

- IPCI configuration details.

- Salient features of IPCI

- Customer advantage with IPCI

Introduction to OpenVMS Clustering

- Reliability
  - HP OpenVMS is known as "**gold standard**" in disaster tolerance

- Scalability
  - Qualified for 96 nodes and also mixed architecture configuration (IPF-Alpha, Alpha-VAX)
  - OpenVMS supports clusters with up to 500 miles apart.

- Manageability
  - Shared-Everything model with Cluster wide file system
  - Single System Image, Cluster wide management facility

hp invent

Disaster Tolerant /Long Distance OpenVMS Clusters

GBE Switch

GBE Switch

GBE Switch

GBE Switch

Inter site

Link

Node East2

Node East1

Node West2

Node West1

VLAN

LAN bridging/Extended LANs using switches

Nodes East1, East2, West1, West2 belong to same LAN/VLAN

Cluster traffic is non-routable

invent

Current OpenVMS Cluster Interconnect Solution

- SCA (aka SCS) – System Communication architecture

  - Cluster communication protocol

- Cluster Interconnect

  - Alpha : LAN, Memory Channel, Shared Memory, CI

  - IPF (Integrity) :LAN

- LAN interconnect for long distance cluster communication

- Bridging and Extended LAN techniques for multi-site long distance clusters

- Nodes belong to same LAN/VLAN for cluster communications

OpenVMS Cluster Communication in LAN
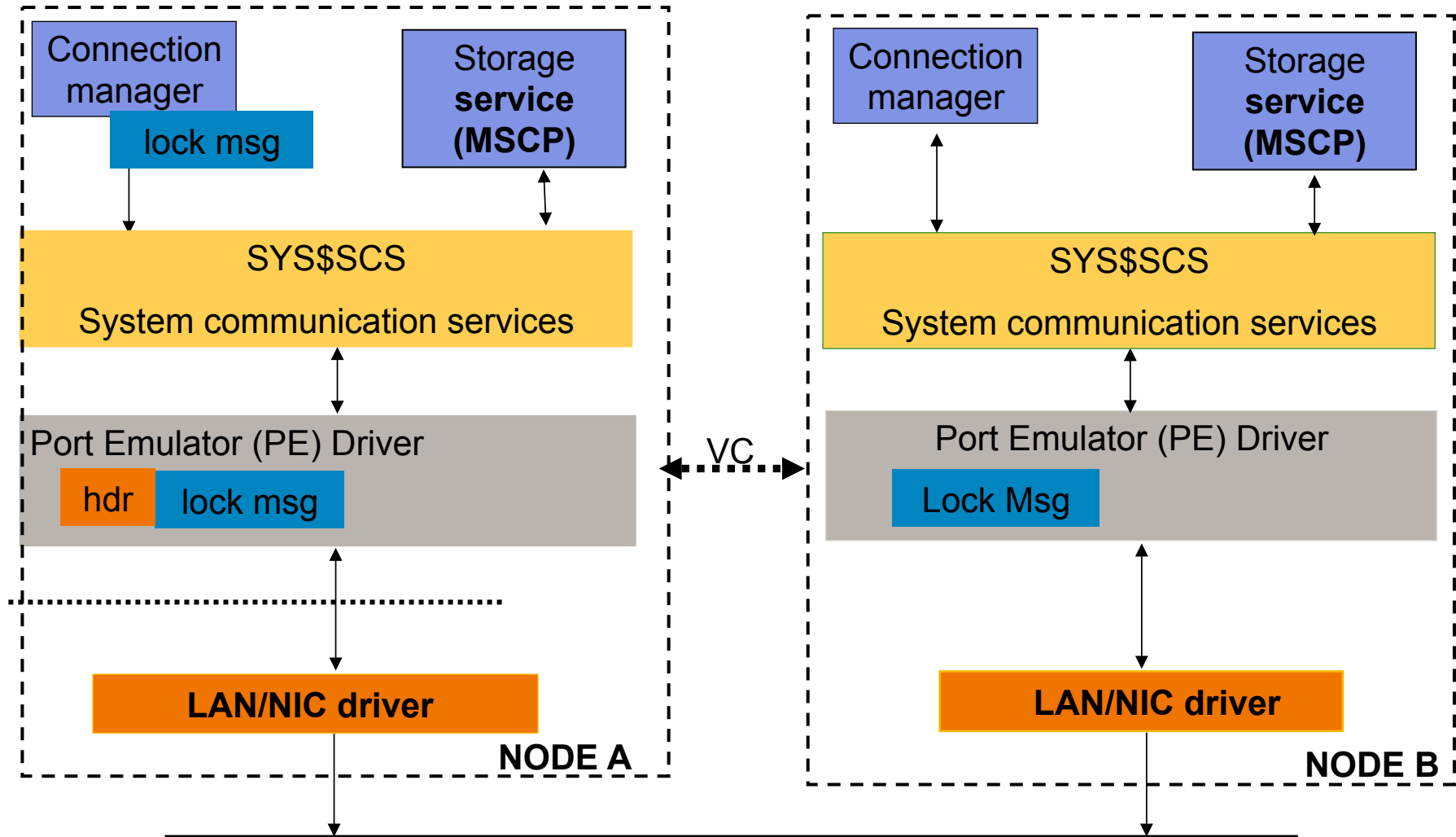
NORSP

MRVP3

Hello pkt

AB-00-04-01-**64-01**

Hello Pkt

Hello pkt
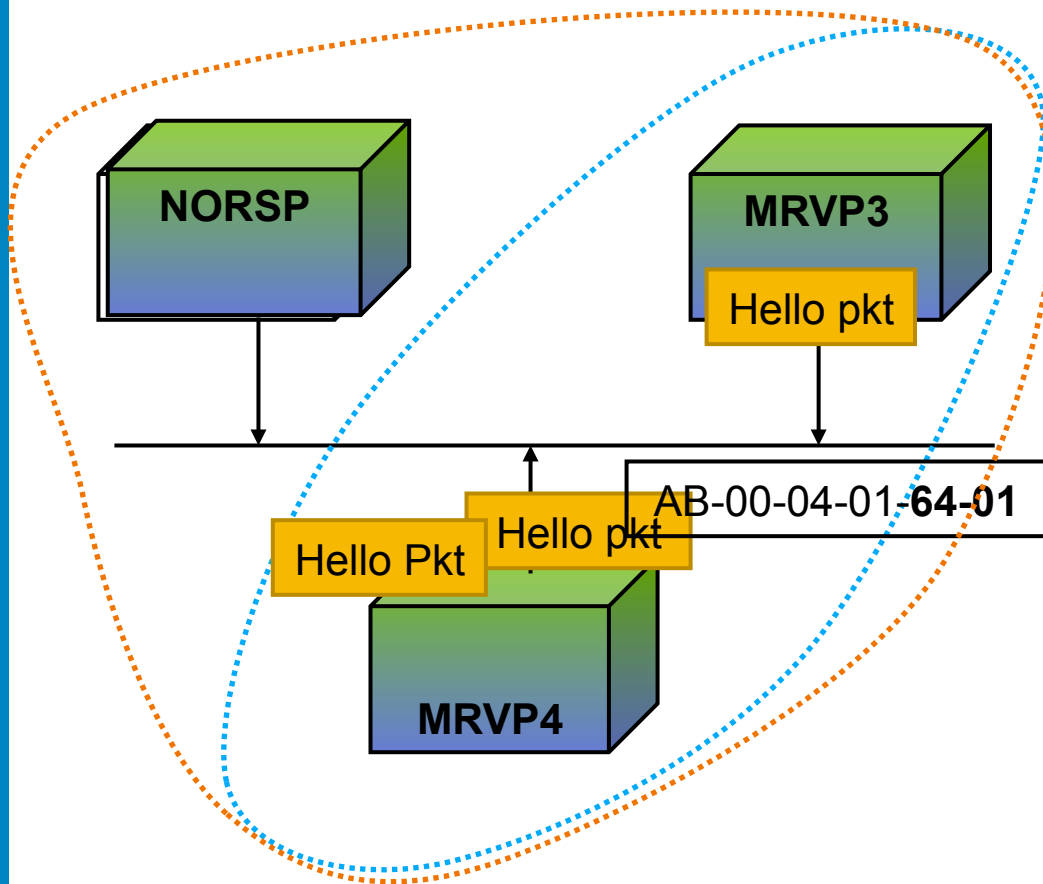
MRVP4

MRVP3 and MRVP4 are in cluster

Cluster group number 100

The multicast address is

AB-00-04-01-00-01 +

the cluster group number

AB-00-04-01-**64-01**

Configure NORSP as cluster and reboot NORSP

hp invent

OpenVMS Cluster Communication Architecture

- VC – Virtual Circuit consists of LAN Channels

Port Emulator (PE) Driver

- Component that implements OpenVMS Cluster communications in LAN (NISCA aka Network Interconnect System Communication Architecture)

- Transmits and receives datagram, sequenced messages and block transfer of data

- Multilayered architecture

- Consists of Transport layer, Channel Control Layer and Data Exchange layer

Current Solution

- OpenVMS customers use LAN interconnect primarily for Cluster communications.

- NISCA protocol which is used for Cluster communication is LAN based.

- OpenVMS Nodes should be in same LAN or VLAN for cluster communications.

- To make OpenVMS cluster work between sites/ Data center customers use bridging and Extended LAN techniques

Technical Motivation for Cluster over IP

- Network switch during higher loads give priority to IP traffic than cluster (SCS) traffic

- Cluster instability during periods of heavy IP usage

- High router utilization for transporting cluster packets

- IP is de-facto industry standard

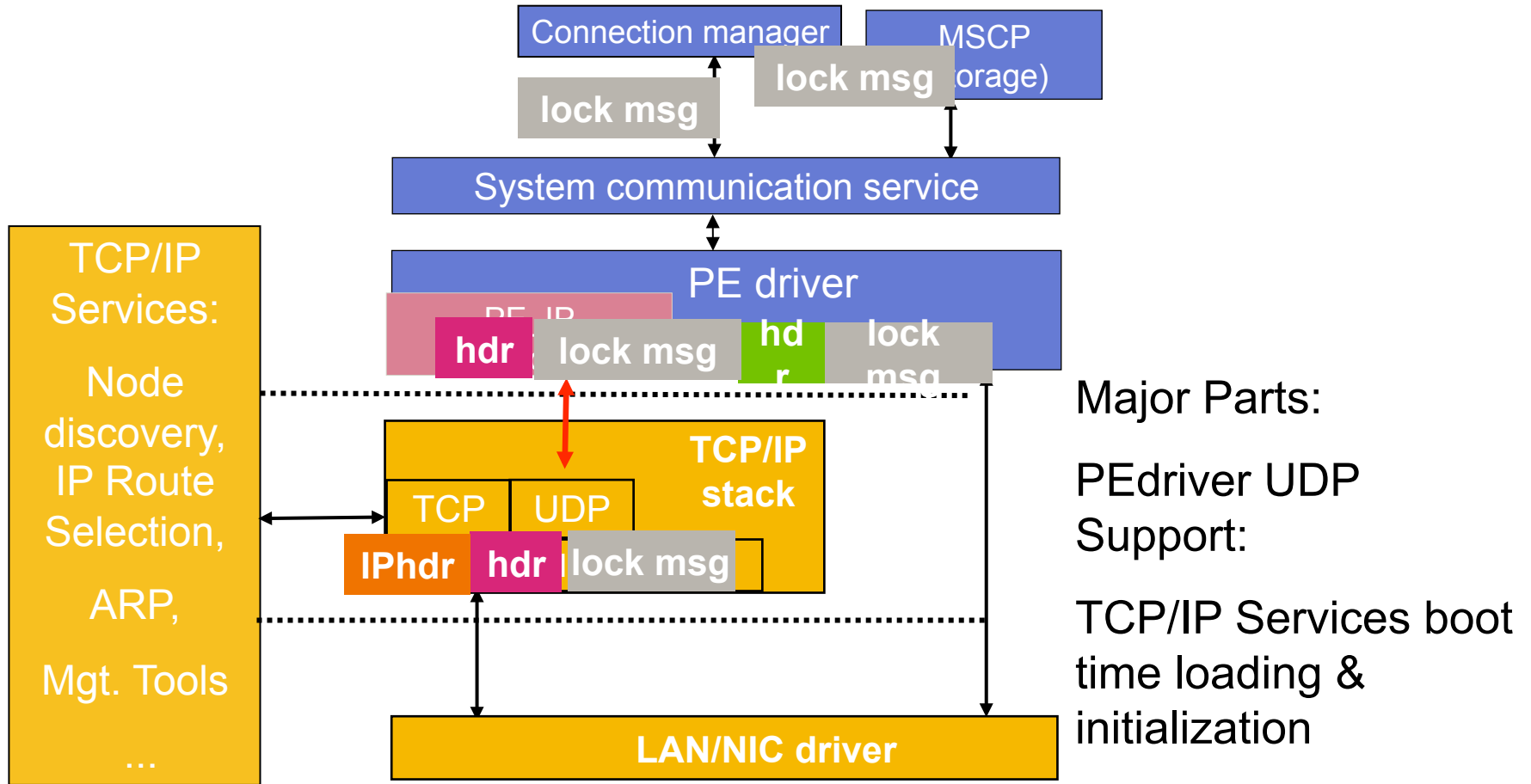- Leveraging benefits of improvements in IP technology

Business/Market motivation for Cluster over IP

- Major Switch vendor dropping support for bridging

- Extra Cost/license for LAN bridging /layer 2 service.

- Corporate policies restricting scope of non-ip protocols

- Mandate of IP only network

- Specialized hardware/human resource costs for setting up multi site DT cluster with LAN bridging

hp invent

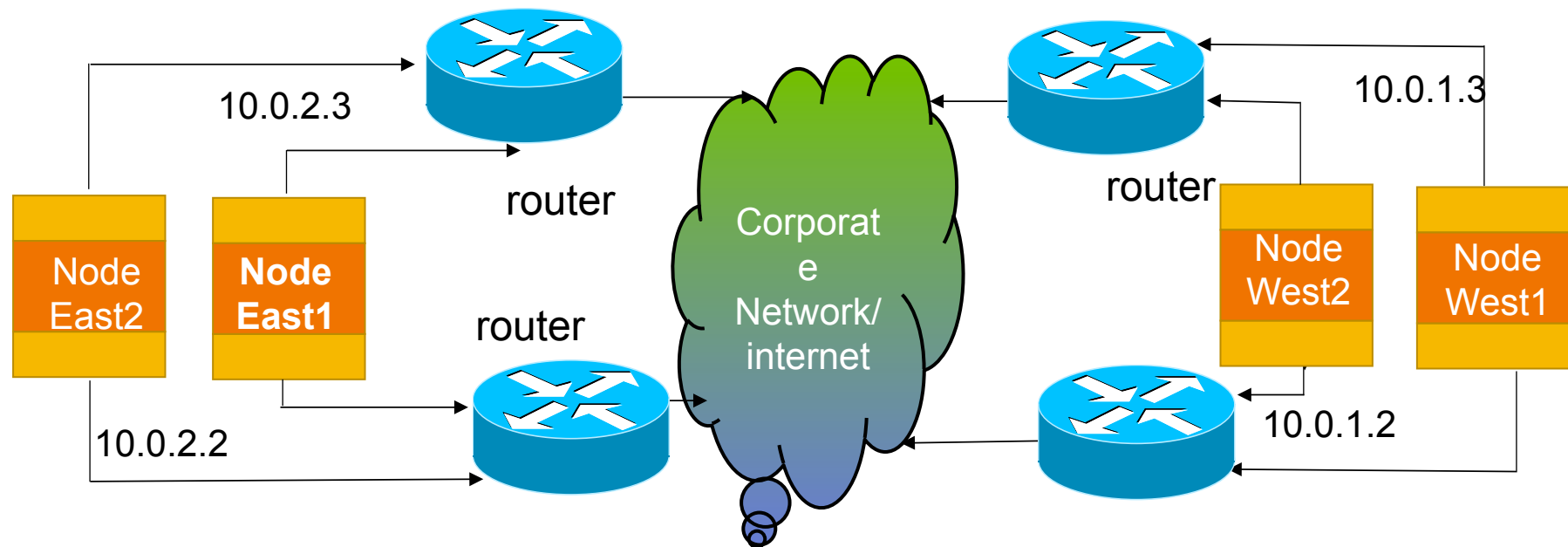# Solution: IP Cluster Interconnect (IPCI) aka OpenVMS Cluster over IP

- IPCI is the ability to make use of IP for OpenVMS clusters communications

- IPCI involves making OpenVMS cluster communication module (PE driver) to use IP stack

- IPCI will coexist with LAN interconnect for Cluster communication

- IP unicast and optionally IP multicast for node discovery

- File based mechanism for unicast node discovery

IPCI solution – PE driver over UDP

Connection manager

lock msg

MSCP (storage)

lock msg

System communication service

TCP/IP Services:

Node discovery, IP Route Selection, ARP, Mgt. Tools ...

PE driver

PE IP

hdr **lock msg** **hdr** **lock msg**

TCP/IP stack

TCP UDP

IPhdr **hdr** **lock msg**

LAN/NIC driver

Major Parts:

PEdriver UDP Support:

TCP/IP Services boot time loading & initialization

Existing Cluster Component | New Component-Component interaction

NEW PEdriver component | Existing VMS/TCP/IP component

hp invent

Cluster using IPCI

Network diagram showing:

- Node East2, Node East1 (left side) connecting to routers labeled 10.0.2.3 and 10.0.2.2
- Corporate Network/internet (center cloud)
- Node West2, Node West1 (right side) connecting to routers labeled 10.0.1.3 and 10.0.1.2
- router labels above and below

• Node East1, East2, West1,West2 can be part of the same or **different LAN** for cluster communications using IPCI. *Cluster traffic is routable*

•**East1 and West 2 has a Virtual Circuit (VC) VC consists of IP channels for SCS traffic**
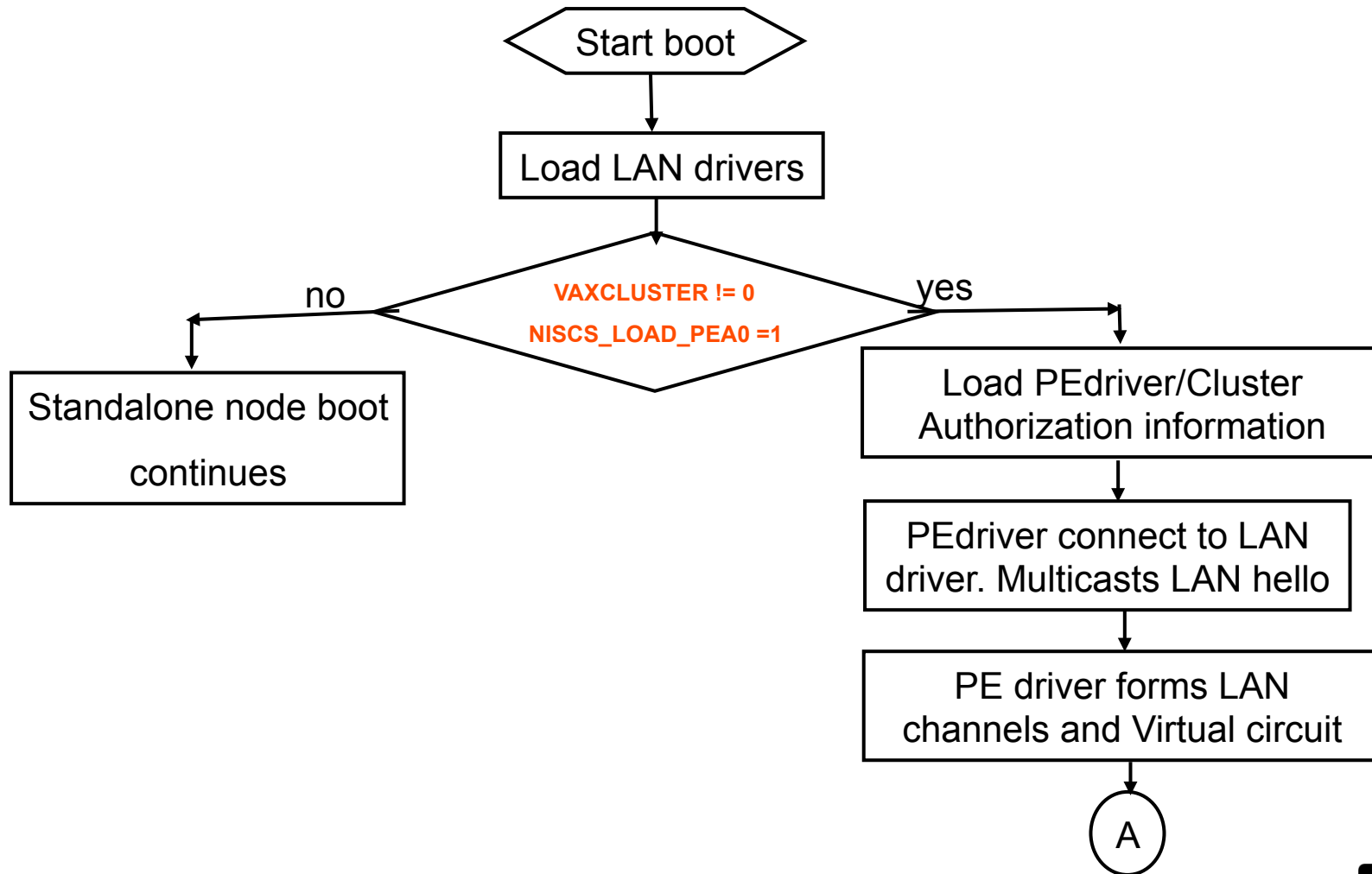
PE driver over UDP

- The IP UDP service has the same packet delivery characteristics as 802 LANs

- PEdriver uses the IP based UDP datagram service as another LAN device

- Only the lowest layer of PEdriver has extension to locate and connect to the TCP/IP

- Uses Kernel mode Interface to talk to IP stack.

- Node discovery using IP Unicast through configuration file

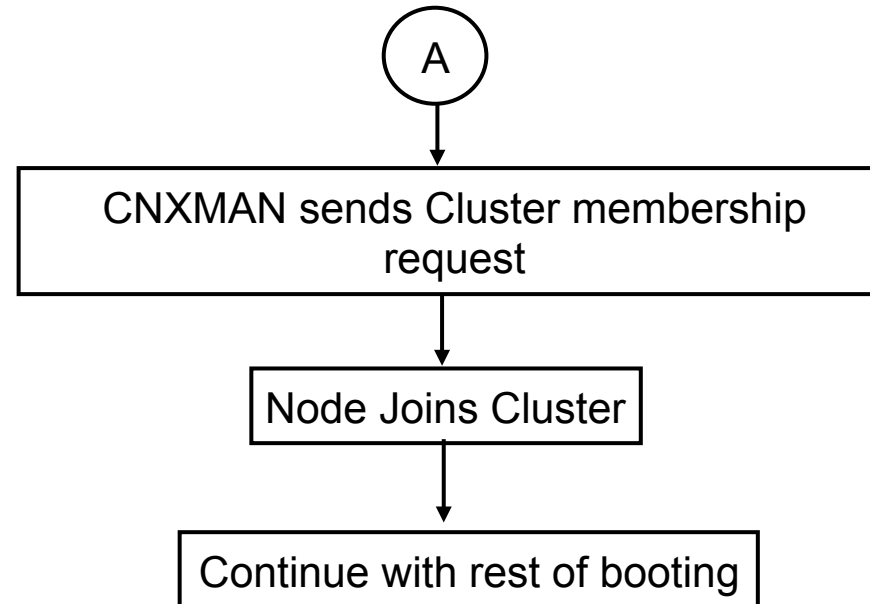- Alternate and Additional Mechanism: IP Multicast

TCP/IP Services boot time loading and Initialization

- Cluster communications are available in an IP only network environment

- Existing boot sequence – LAN,PE driver, TCP/IP

- Boot Sequence with IPCI – LAN,TCP/IP, PE driver

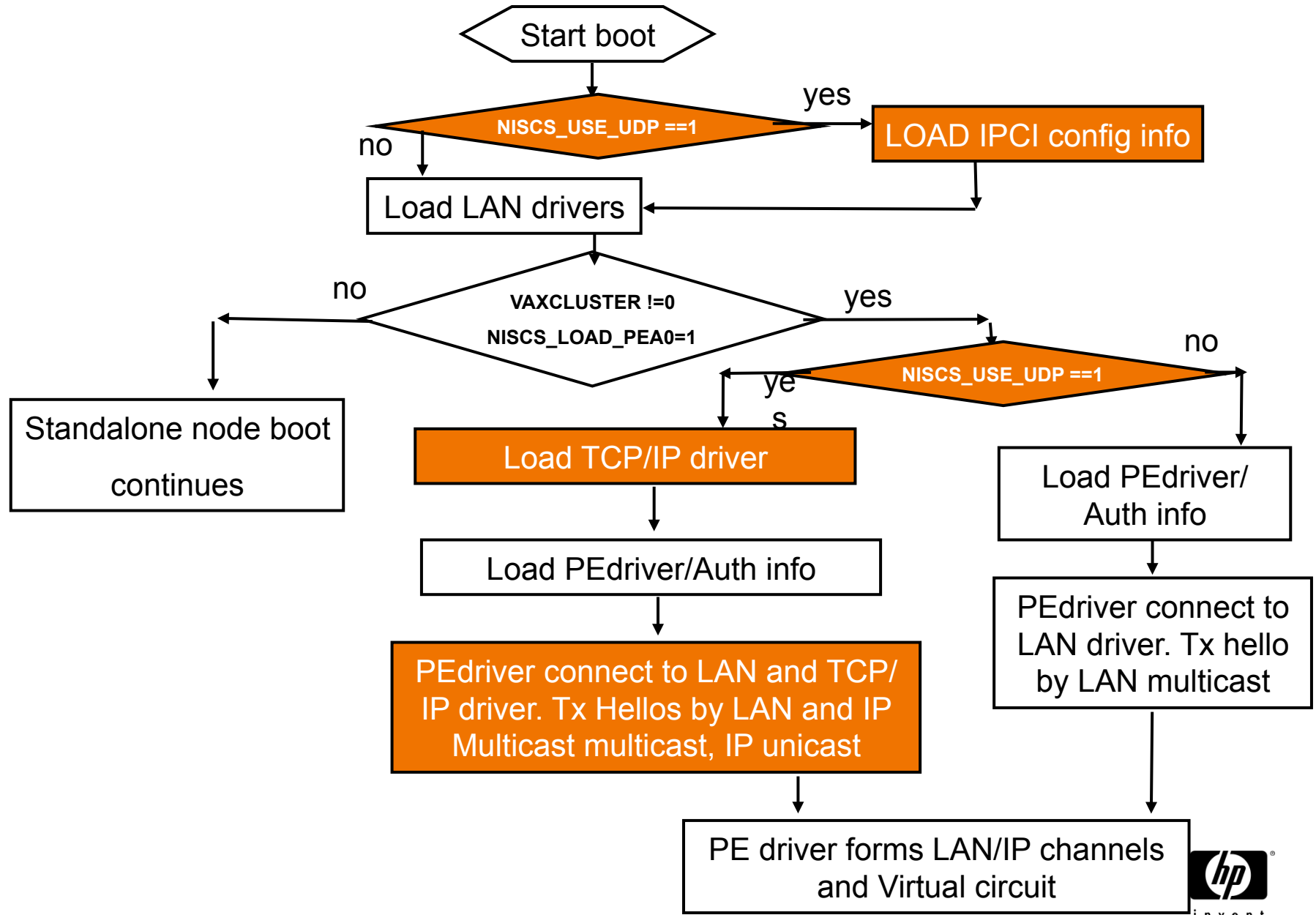- Ability to make use of boot time configuration information to initialize TCP/IP services
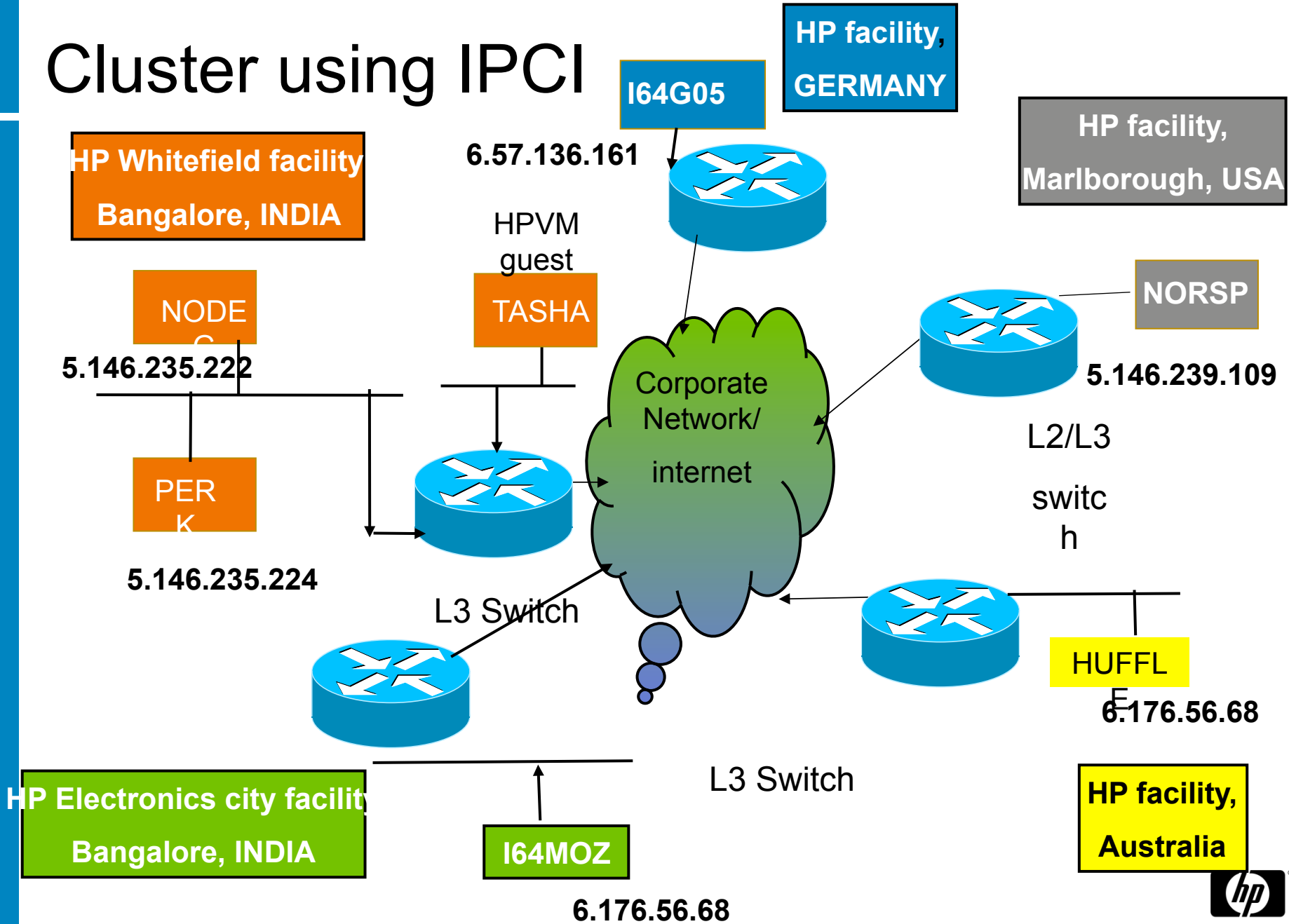
OpenVMS Cluster formation (LAN)

```
        ┌─────────────────┐
        │   Start boot    │
        └────────┬────────┘
                 │
                 ▼
        ┌─────────────────┐
        │ Load LAN drivers│
        └────────┬────────┘
                 │
                 ▼
no  ◄───────◇ VAXCLUSTER != 0 ◇───────► yes
    │         NISCS_LOAD_PEA0 =1          │
    ▼                                     ▼
┌──────────────────┐        ┌──────────────────────────┐
│ Standalone node  │        │ Load PEdriver/Cluster    │
│ boot             │        │ Authorization information│
│                  │        └────────────┬─────────────┘
│ continues        │                     │
└──────────────────┘                     ▼
                            ┌──────────────────────────┐
                            │ PEdriver connect to LAN  │
                            │ driver. Multicasts LAN   │
                            │ hello                    │
                            └────────────┬─────────────┘
                                         │
                                         ▼
                            ┌──────────────────────────┐
                            │ PE driver forms LAN       │
                            │ channels and Virtual      │
                            │ circuit                  │
                            └────────────┬─────────────┘
                                         │
                                         ▼
                                        (A)
```

OpenVMS Cluster formation (LAN)

A

CNXMAN sends Cluster membership request

Node Joins Cluster

Continue with rest of booting

OpenVMS Cluster formation (IPCI)

Start boot

NISCS_USE_UDP ==1
— yes → LOAD IPCI config info
— no →

Load LAN drivers

VAXCLUSTER !=0
NISCS_LOAD_PEA0=1

— no → Standalone node boot continues

— yes →

NISCS_USE_UDP ==1
— no → Load PEdriver/ Auth info → PEdriver connect to LAN driver. Tx hello by LAN multicast

— yes →

Load TCP/IP driver

Load PEdriver/Auth info

PEdriver connect to LAN and TCP/IP driver. Tx Hellos by LAN and IP Multicast multicast, IP unicast

PE driver forms LAN/IP channels and Virtual circuit

hp invent

Node Discovery – Unicast and Multicast

- IP unicast used for node discovery and hello packets.

- IP multicast can also be used (Administratively scoped IP multicast address)

- Remote nodes not in IP multicast domain use IP unicast technique to join Cluster and send hello packets.

- Basic Cluster principle: All nodes must see all.

Cluster using IPCI

HP facility, GERMANY

I64G05

6.57.136.161

HP facility, Marlborough, USA

HP Whitefield facility Bangalore, INDIA

HPVM guest

NODE C

TASHA

NORSP

5.146.235.222

Corporate Network/ internet

5.146.239.109

L2/L3

PER K

switc h

5.146.235.224

L3 Switch

HUFFL E

6.176.56.68

HP Electronics city facility Bangalore, INDIA

L3 Switch

I64MOZ

HP facility, Australia

6.176.56.68

hp invent

SHOW CLUSTER

SYSTEMS		MEMBERS
NODE	SOFTWARE	STATUS
I64MOZ	VMS XBXH-J2I	MEMBER
NODEG	VMS XBXH-J2I	MEMBER
HUFFLE	VMS XBYE-BL1	MEMBER
STKTSC	VMS XBYE-BL1	MEMBER
NORSP	VMS XBYD-J2I	MEMBER
PERK	VMS XBXI-BL1	MEMBER

```
$
$
```

invent

Cluster using IPCI

- 2 nodes in Bangalore Whitefield facility
- 1 node in Bangalore (India) Electronics city facility
- 1 node in HP, Marlborough, USA
- 1 node in HP Germany , 1 node in HP Australia
- 1 HPVM guest node
- Distance between Bangalore facilities <50 miles. PING latency 4 ms. Cluster Latency approx 4ms.
- Distance approx 8000 miles. Ping Latency between Bangalore and Marlborough 350ms.Cluster Latency approx 350ms.

Add a node into a IPCI cluster

$ @SYS$MANAGER:CLUSTER_CONFIG_LAN

        Cluster Configuration Procedure
        CLUSTER_CONFIG_LAN Version V2.80
         Executing on an IA64 System

DECnet Phase IV is installed on this node.
IA64 satellites will use TCP/IP BOOTP and TFTP services for downline loading

TCP/IP is installed and running on this node.

    Enter a "?" for help at any prompt. If you are familiar with
    the execution of this procedure, you may want to mute extra notes
    and explanations by invoking it with "@CLUSTER_CONFIG_LAN BRIEF".

   This IA64 node is not currently a cluster member.
MAIN Menu

    1. ADD I64MOZ to existing cluster, or form a new cluster.
    2. MAKE a directory structure for a new root on a system disk.
    3. DELETE a root from a system disk.
    4. EXIT from this procedure.

hp invent

1. ADD I64MOZ to existing cluster, or form a new cluster.
2. MAKE a directory structure for a new root on a system disk.
3. DELETE a root from a system disk.
4. EXIT from this procedure.

Enter choice [4]: 1
Is the node to be a clustered node with a shared SCSI/FIBRE-CHANNEL bus (Y/N)? n

  IA64 node, using LAN for cluster communications. PEDRIVER will be loaded.
  No other cluster interconnects are supported for IA64 nodes.
Enter this cluster's group number: 1985
Enter this cluster's password:
Re-enter this cluster's password for verification:
ENABLE IP for cluster communications (Y/N)? Y
UDP port number to be used for Cluster Communication over IP[49153]?
Enable IP multicast for cluster communication(Y/N)[Y]?
What is IP the multicast address[239.242.7.193]?
What is the TTL (time to live) value for IP multicast packets [32] ?
Do you want to enter unicast address(es)(Y/N)[Y]?
What is the unicast address[Press [RETURN] to end the list]? 6.118.162.109
What is the unicast address[Press [RETURN] to end the list]? 6.50.12.169
What is the unicast address[Press [RETURN] to end the list]? 6.176.56.68
What is the unicast address[Press [RETURN] to end the list]? 5.146.235.222
What is the unicast address[Press [RETURN] to end the list]? 6.138.182.6
What is the unicast address[Press [RETURN] to end the list]?

What is the unicast address[Press [RETURN] to end the list]?

```
******************************************************************

    Cluster Communications  over IP  has  been  enabled.  Now
    CLUSTER_CONFIG_LAN will run the  SYS$MANAGER:TCPIP$CONFIG
    procedure. Please select the IP interfaces to be used for
    Cluster Communications  over IP  (IPCI). This can be done
    selecting "Core Environment"  option from the main menu
    followed by the "Interfaces"  option.  You may also use
    this opportunity to configure other aspects.
******************************************************************
```

 TCP/IP Network Configuration Procedure

    This procedure helps you define the parameters required

    to run HP TCP/IP Services for OpenVMS on this system.

- %TCPIP-I-IPCI, TCP/IP Configuration is limited to IPCI.
- -TCPIP-I-IPCI, Rerun TCPIP$CONFIG after joining the cluster.

Press Return to continue ...

- HP TCP/IP Services for OpenVMS Interface & Address Configuration Menu

Hostname Details: Configured=Not Configured, Active=Not Configured

Configuration options:

  0 - Set The Target Node (Current Node: I64MOZ)

  1 - IE0 Menu (EIA0: TwistedPair 100mbps)

  2 - IE1 Menu (EIB0: TwistedPair 1000mbps)

 [E] - Exit menu

Enter configuration option: 1

* IPCI Address Configuration *

Only IPCI addresses can be configured in the current environment.
After configuring your IPCI address(es) it will be necessary to
run TCPIP$CONFIG once your node has joined the cluster.

IPv4 Address may be entered with CIDR bits suffix.
E.g. For a 16-bit netmask enter 10.0.1.1/16

Enter IPv4 Address []: 6.138.182.6/24

Requested configuration:

    Node   : I64MOZ

    Interface: IE0

    IPCI   : Yes

    Address : 16.138.182.6/24

    Netmask : 255.255.255.0 (CIDR bits: 24)

Is this correct [YES]:

Updated Interface in IPCI configuration file: SYS $SYSROOT:[SYSEXE]TCPIP$CLUSTER.DAT;

HP TCP/IP Services for OpenVMS Interface & Address Configuration Menu

Hostname Details: Configured=Not Configured, Active=Not Configured
Configuration options:

  0 - Set The Target Node (Current Node: I64MOZ)
  1 - IE0 Menu (EIA0: TwistedPair 100mbps)
  2 - 16.138.182.6/24   *noname*   IPCI
  3 - IE1 Menu (EIB0: TwistedPair 1000mbps)
 [E] - Exit menu

Enter configuration option: E
Enter your Default Gateway address []: 6.138.182.1
The default gateway will be: 16.138.182.1. Correct [NO]: YES
Updated Default Route in IPCI configuration file: SYS$SYSROOT:[SYSEXE]TCPIP$CLUSTER.DAT;
TCPIP-I-IPCIDONE, Finished configuring IPCI address information.
Will I64MOZ be a boot server [Y]?

SYS$SYSTEM:PE$IP_CONFIG.DAT

Configuration File for IPCI.

! CLUSTER_CONFIG_LAN creating for CHANGE operation on 8-NOV-2008 10:46:19.26

multicast_address=239.242.7.193

ttl=32

udp_port=49153

unicast=6.118.162.109

unicast=6.50.12.169

unicast=6.176.56.68

unicast=5.146.235.222

unicast=6.138.182.6

SYS$SYSTEM:PE$IP_CONFIG.DAT

- Generated by CLUSTER_CONFIG_LAN.COM

- Read early in the boot sequence.

- Provides information to PEdriver.

- Can be common through out cluster.

- Remote node IP address should be present in local node PE$IP_CONFIG.DAT in order to allow remote node join the cluster.

- Best practice for IP unicast: Include all IP address and have one copy throughout the cluster.

- "$MC SCACP reload" to be used to refresh IP unicast list on a live system.

SYS$SYSTEM:TCPIP$CLUSTER.DAT

Configuration File for IPCI.

- default_route=6.138.182.1
- interface=IE0,EIA0,6.138.182.6,255.255.255.0

SYS$SYSTEM:TCPIP$CLUSTER.DAT

- Generated by TCPIP$CONFIG which is invoked by CLUSTER_CONFIG_LAN.COM

- Read early in the boot sequence.

- Provides information to PEdriver 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.

Console Messages

HP OpenVMS Industry Standard 64 Operating System, Version XBXH-J2I

© Copyright 1976-2008 Hewlett-Packard Development Company, L.P.

%DECnet-I-LOADED, network base image loaded, version = 05.16.00

%VMScluster-I-LOADIPCICFG, loading the IP cluster configuration files

%VMScluster-S-LOADEDIPCICFG, Successfully loaded IP cluster configuration files

%SMP-I-CPUTRN, CPU #1 has joined the active set.

%SYSINIT-I- waiting to form or join an OpenVMS Cluster

%VMScluster-I-LOADSECDB, loading the cluster security database

%EIA0, Auto-negotiation mode assumed set by console

Console Messages

%EWE0, Link up: 1000 mbit, full duplex, flow control (txrx)

%EWD0, Link up: 1000 mbit, full duplex, flow control (txrx)

%PEA0, Configuration data for IP clusters found

%PEA0, IP Multicast enabled for cluster communication, Multicast address, 239.242.7.193

%PEA0, Cluster communication enabled on IP interface, WE0

%PEA0, Successfully initialized with TCP/IP services

%PEA0, Remote node Address, 6.138.185.6,!INDIA added to unicast list of IP bus, IE0

%PEA0, Remote node Address, 5.146.235.222, !INDIA added to unicast list of IP bus, IE0

%PEA0, Remote node Address, 5.146.239.109,!USA added to unicast list of IP bus, IE0

%PEA0, Remote node Address, 5.146.235.224,INDIA added to unicast list of IP bus, IE0

%PEA0, Remote node Address, 6.176.56.68,AUSTRALIA added to unicast list of IP bus, IE0

%PEA0, Remote node Address, 6.50.12.169 ! SWEDEN, added to unicast list of IP bus, IE0

%PEA0, Remote node Address, 5.146.238.251 ! HPVMGUEST, added to unicast list of IP bus, IE0

%PEA0, Hello sent on IP bus IE0

%PEA0, Cluster communication successfully initialized on IP interface , IE0

%CNXMAN, Sending VMScluster membership request to system NORSP

%CNXMAN, Now a VMScluster member -- system I64MOZ

%STDRV-I-STARTUP, OpenVMS startup begun at 29-OCT-2008 15:20:41.10

hp
invent

SCACP commands

$ MC SCACP SHOW CHANNEL PERK

```
NODEG PEA0 Channel Summary 10-MAY-2008 05:09:51.38:


Remote LAN/IP Dev Channel ECS Buffer Delay  Packets
Node     Loc Rmt    State   state      Size  uSec  (S+R)
-------------------------------------------------------------
PERK     WE0 IE0   Open   N(T,I,F) 1394   708.9 30410576
PERK     EWA EIA   Open   Y(T,P,F) 1426   551.9 26586732
PERK     WE0 IE1   Open   N(T,I,F) 1394   784.2 38475399
PERK     EWA EIB   Open   Y(T,P,F) 1426   572.4 23780101
PERK     EIA EIB   Open   Y(T,P,F) 1426   694.1 15288091
```

IP channels between nodes   **LAN channels between nodes**

NEW SCACP commands

$ MC SCACP SHOW CHANNEL <nodename>/IP and /LAN

$ MC SCACP SHOW IP_INTERFACE <ip_interface>
   Example. MC SCACP SHOW IP_INTERFACE we0

$ MC SCACP START IP_INTERFACE <ip_interface>
   Example. MC SCACP START IP WE0

$MC SCACP START IP_INTERFACE <ip_interface>

$ MC SCACP SET IP_INTERFACE <ip_interface>
   Example MC SCACP SET IP_INTERFACE WE0/PRIORITY=4

$MC SCACP STOP IP_INTERFACE <ip_interface>
   Example MC SCACP STOP IP_INTERFACE WE0
$ MC SCACP RELOAD

Virtual Circuit (VC)

- VC is logical connection between two nodes

- VC consists of Channels (LAN or IP)

- VC check summing and compression applicable to Cluster over IP also.

Equivalent Channel Set (ECS)

- Set of Eligible Channels used by PEdriver for communication

- ECS technique applicable to IP channels also

- Tight (T), Peer (P), Fast(F) characteristics are required for a channel to be part of ECS

- TPF charactertistics applicable to IP channels also.

- LAN channels preferred over IP channels

Distance and Latency

- Current Distance limitation will still be applicable.

- Speed of Light causes approx 1 millisecond for 50 mile roundtrip.

- Distance more than 500 miles require site specific configuration and we suggest contact HP DTCS. (HP Disaster Tolerant Cluster services) or Product management.

Security

- Normal intranet and Internet Security principles
- VPN (virtual Private Network)
- TTL (Time to live)
- Firewalls.

Performance

- Engineering will conduct some performance test to recommend configurations for optimal performance.

- PING – PONG tests will be conducted and recommendations will be based on the results.

- Observation reveal tcp/ip ping latency close to latency reported by PEdriver

Feature details

- Available with OpenVMS V8.4 only (Alpha and Integrity)
  - Will be available with V8.4 Field Test
  - No Prior version support
- Requires HP TCP/IP services for OpenVMS V5.7
  - Not available with other TCP/IP stacks at this time
  - Initial release supports IPv4 only; no IPv6
  - Requires static IP addresses and IP Unicast; optionally uses IP Multicast
- Coexists with LAN interconnect for Cluster communication
- Support for Satellite nodes included
- Existing intra-node distance/latency limitation (500 miles) applies

Salient Features

- Ability to Discover nodes and form cluster in an IP only network

- Perform rolling upgrades to the new version without a cluster reboot

- Dynamically load balance using all the available healthy interconnect interfaces

- Interoperability with prior versions

- Delay measurement technique for discovering path with minimal latency

Customer Advantage

- Only IP services (No LAN bridging) are provided by some Telco Vendors. Customers can now use OpenVMS clusters with IPCI

- Lower infrastructural and Operational costs

- No extra license/cost for LAN bridging (Layer 2 service)

- Leverage the benefits from the improvements in IP and LAN interconnect technology

- One network infrastructure in Data center for all purposes