NetWare Cluster Services - Internals

Bhogilal Hirani

Technical Instructor
Novell Education (EMEA)
BHirani@Novell.COM

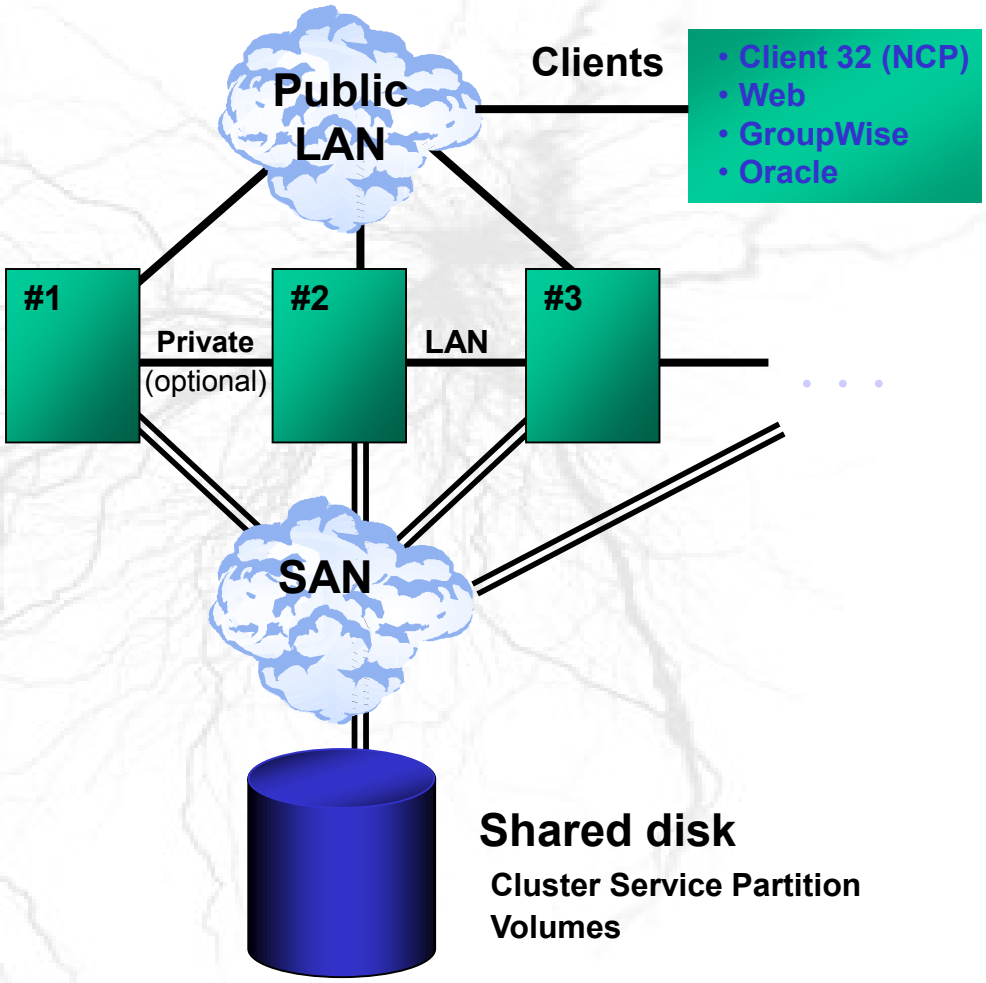
NetWare Cluster Services - Internals

Agenda

- Introduction to NetWare Cluster Services
- Resource definition and failover
  - Seeing is believing
    - DHCP
    - NDPS
    - Enterprise Web Server
    - GroupWise
- Cluster configuration
  - Internal Communications
  - Client Reconnection
- Architecture
  - NLM's / Objects (Ver 1.01 & 1.6)
  - Split Brain
- What's New in NetWare 6 and NWCS 1.6

NetWare Cluster Services - Internals

Introduction



NetWare Cluster Services - Internals

Introduction to NetWare Cluster Services

- High Availability of Resources
- Single System Image
- Single Point of Administration
  - ConsoleOne, Server Console OR Web Browser
- Scalability
  - “1 system” = upto 32 active nodes
- FANOUT failover
- All the features NSS and NDS
- IP Client auto-reconnection

NetWare IS the Platform (Resilience, Reliability, Security etc.)

NetWare Cluster Services - Internals

Resource Failover

What is a Resource?

Application (NLM + IP Address + Volume

Cluster Enabled Volume (IP Address + Volume)

Failure Triggers

ONLY when Node communication Broken on LAN or SAN

- Loss of Power
- Server crash / Abend
- LAN failure (NIC, cable, hub/switch)
- SAN Failure (FC controller, GBIC, Cable, laser
- etc

NetWare Cluster Services - Internals

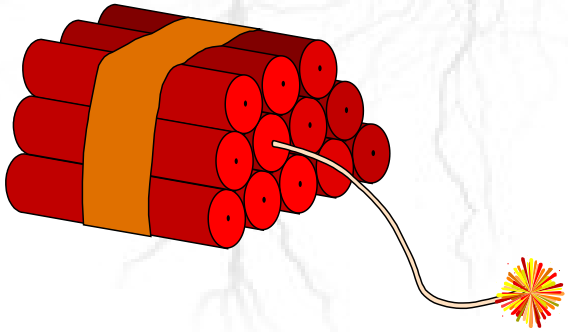
Seeing is believing

Power Off

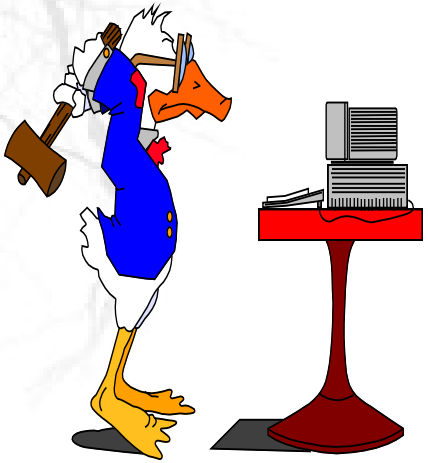
Demonstrations

- DHCP
- NDPS
- Enterprise Web Server
- GroupWise

Abend



Treat it gently



Pull Plug

NetWare Cluster Services - Internals

Cluster Configuration

All nodes maintain the same cluster view/state
All nodes communicate to one another (LAN & SAN)
Master/Slave relationship maintained

- Always 1 Master Node
- Failure of Master triggers election of new Master
- Inter Node communication (protocol setting)
  - Failure timeout (Hearbeat)
    - Master watchdog (Master -> Slave - 1/sec)
    - Hearbeat Slaves -> Master (1/sec)
    - Default tolerance (8/sec)
- SAN communication
  - Each node updates counter on Cluster Partition
  - When updating, checks the counter of other nodes

NetWare Cluster Services - Internals

Cluster Configuration (cont…)

- Quorum Triggers
  - Resources load if membership reached or time out expired
    - Timeout (60 sec)
    - Membership (# nodes)
- Cluster Specific Objects
  - Nodes
    - Link to the NCP Server
    - Node Number
    - IP Address of Node
  - Templates
    - Allows easy resource configuration
  - Resources
    - Application resource
    - Cluster-Enabled Volume

NetWare Cluster Services - Internals

Resources

Cluster-Enabled Volume

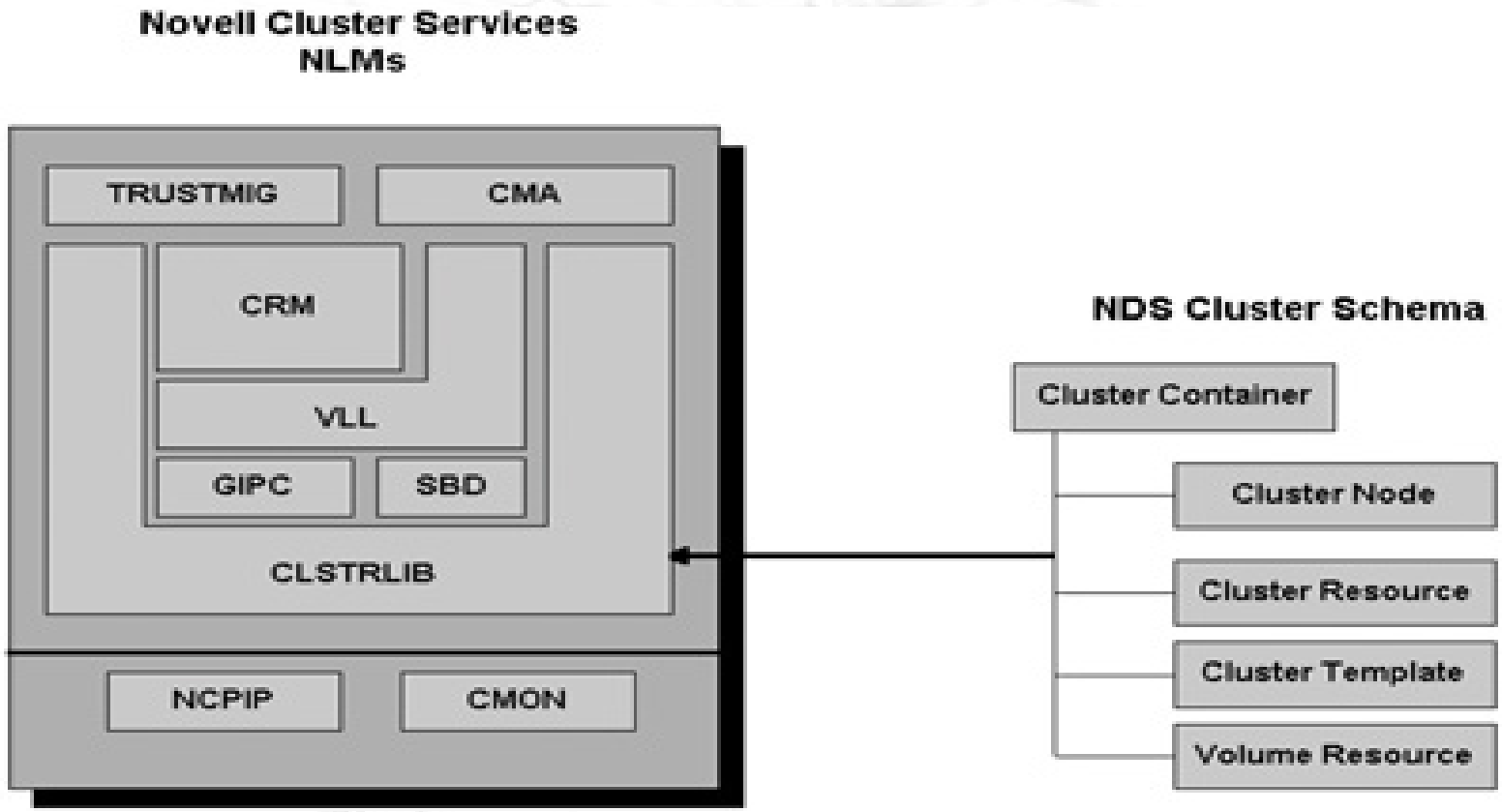
- IP Address Required
- Virtual NCP Server Created and advertised
- Bound on a node as Secondary IP Address
- failover maintains volume ID, server & IP address

Client Reconnection Maintained

- Migration triggers ARP broadcast when IP address is bound
- Client refreshes the cash to reflect the new MAC address

NetWare Cluster Services - Internals

Architecture NWCS 1.x



NetWare Cluster Services - Internals

2 Main Components

- Cluster Object
  - Created during installation (Schema extension)
    - Cluster container
      - Quorum triggers, protocol properties etc.
    - Cluster node
      - Node Number, IP Address
    - Resources
      - Scripts, timeouts, start/failover/failback policies, nodes list

NetWare Cluster Services - Internals

2 Main components cont…

- NLM's - The power of 7 NLM's (1.02 Mb)
  - **CLSTRLIB** - **CL**u**S**te**R** Configuration **L**IBrary
  - **GIPC** - **G**roup **I**nter**P**rocess **C**ommunication
  - **SBD** - **S**plit **B**rain **D**etector
  - **VLL** - **V**irtual Interface Architecture **L**ink **L**ayer
  - **CRM** - **C**luster **R**esource **M**anager
  - **TRUSTMIG** - **TRUST**ee **MIG**ration
  - **CMA** - **C**luster **M**anagement **A**gent

- The NLM's run on each node in the cluster (LDNCS.NCF)

NetWare Cluster Services - Internals

Architecture cont...

CLSTRLIB.NLM

- 1st node to join the cluster = MASTER
- CLSTRLIB on the master caches cluster specific DS information
- Slave nodes cache information from master when joining
- Master node responsible for committing the changes to NDS

GIPC.NLM

- Monitors Group membership
- Membership (sub-protocol) maintains group membership
- Heartbeat (sub-protocol) maintains heartbeat packets
- Contacts the VLL.NLM if the heartbeat is not received

NetWare Cluster Services - Internals

Architecture cont...

VLL.NLM

- Contacts the SBD.NLM to check SAN connectivity
- Interface between GIPC.NLM and SBD.NLM
- Contacts CRM.NLM (resource failover)

SBD.NLM

- Has access to the SBD (Cluster Services) partition
- Checks the SBD partition if node is communicating with the SAN
- Confirms SAN connectivity to VLL.NLM
  - If no SAN connectivity, VLL.NLM contacts CRM.NLM

NetWare Cluster Services - Internals

Architecture cont...

CRM.NLM

- On the master node
  - CRM Maintains an accurate view of resources
    - Nodes list
    - Resource information
      - scripts, start/failover/failback
    - Maintains real time status of all resources
    - Distributes real time status to all slaves
  - Resource states
    - Unassigned, Offline, Loading, Unloading
      Comatose, Running, Alert , NDS sync, Quorum wait

NetWare Cluster Services - Internals

Architecture cont...

CMA.NLM

- Proxy to the ConsoleOne
- Interacts with CRM to provide the view to ConsoleOne

TRUSTMIG.NLM

- Monitors NDS Trustee assignments on a cluster-enabled volume
- Migrates the trustee information to the new node
- Trustee information stored in _NetWare directory (TRUSTMIG.FIL)
- Relies on the tree being healthy

CMON.NLM

- Monitors and displays cluster status on each server

NetWare Cluster Services - Internals

Architecture cont...

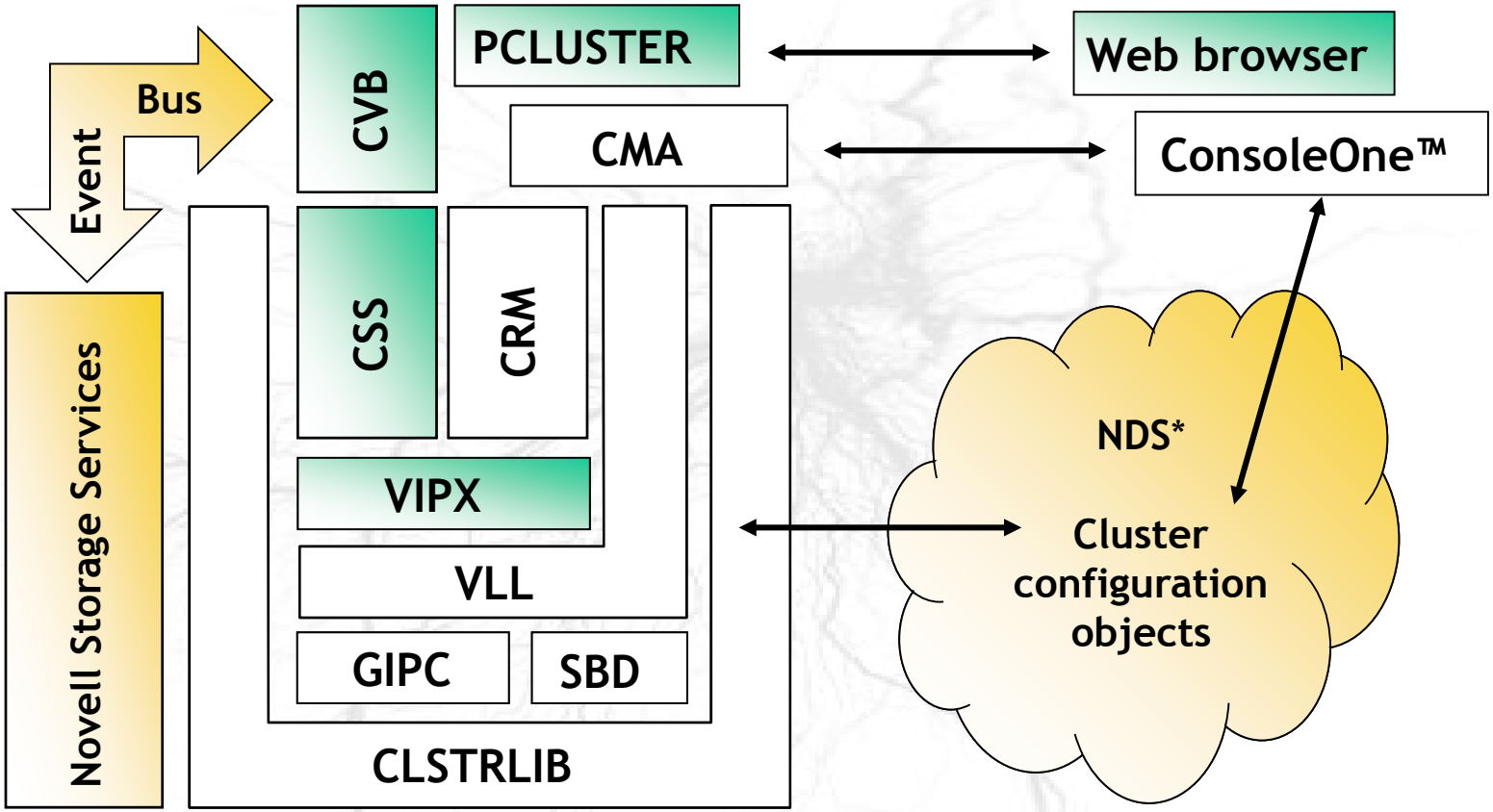
ALL IN ACTION

Groupwise running on Nodes A,B and C and node A fails

- Node A fails (no Hearbeat from Node A's GIPC.NLM)
- GIPC.NLM on Nodes B and C notify their VLL modules
- VLL modules contacts SBD modules to check SAN connectivity
- SBD confirms with VLL that Node A is unavailable
- The VLL modules contact the CRM modules
- CRM checks with CLSTRLIB's info on each node (Node list)
- Node B executes and confirms execution of load script with node C
  (In case node B fails, Node C can take over)
- TRUSTMIG.NLM ensures migration of trustees of volume
- CMON updates Cluster monitor Screen on each server

Architecure cont..

NWCS 1.6 Architecture Changes



NetWare Cluster Services - Internals

Split Brain

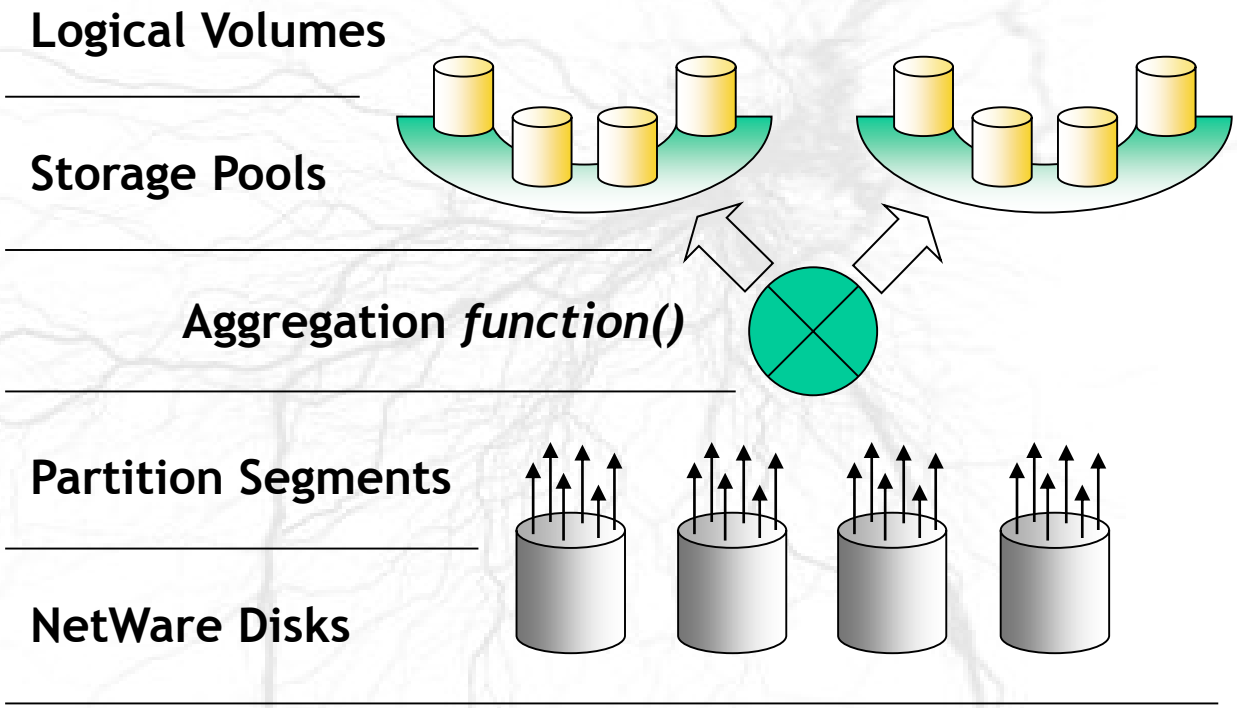
- Avoids a cluster being split into 2 Sub-Cluster
  - Caused by a LAN failure
- Data corruption is avoided by forced abends on one side
  - Auto restart after abend=0
    - set by Cluster Services Installation
- Which side wins?
  - Majority number of nodes
  - If Equal side with Master node wins
  - In a 2 node cluster, Patch required
    - Avoids the master with the network cable unplugged surviving

NetWare Cluster Services

What's new in Ver 1.6

- Installation Changes
  - Install using Deployment Manager
  - Multiple node select during Install
  - Master_IP_Address_Resource
    - Allows cluster management using the Browser
    - Only runs on the master
- Only local devices are activated by default
- Closely tied with NSS version 3.0
  - Device/volume Management done through ConsoleOne
- Device on the SAN must be flagged "Shareable for Clustering"

What's New in Ver 1.6 cont..



What's New in Ver 1.6 cont..

- Failover is done at shared POOL level
  - Cluster Enabling a volume in a pool activates the Pool
    - All volumes in a Pool are tied to the pool
    - 1 volume per pool emulates nw5.x/NWCS 1.01
  - Virtual Server Name derived from the pool name
    - Virtual server name is customisable
    - Clustered volume name is customisable
    - "_" can be eliminated
  - Dynamic storage expansion
    - Cluster Volume Broker consumes events from NSS
    - Distributes events to all nodes in the cluster
    - Provides a consistent view from all nodes

What's New in Ver 1.6 cont..

Grey Data Protection

- Flush on close
  - Cache flushed on closing a file
  - Performance Penalty
- Cluster console commands
  - manage resources from any nodes' console screen
    - migrate/offline/start/stop/status/alert etc.
  - Manage resources with a Web Browser
    - uses the Master_IP_Address resource
    - http://<master ip address>:8009
    - Full alternative to ConsoleOne

What's New in Ver 1.6 cont..

- SNMP
  - Node Traps
    - node Join, Leave or Fail
  - Cluster discovery
    - Cluster Name, Node name, IP addresses
- SMTP
  - Event based E-Mail notification
    - Plain text or XML format
  - Node events
    - Join, Leave or Fail
  - Resource events
    - Cluster resource state changes

What's New in Ver 1.6 cont..

- Diagnostic tool
  - monitor LAN and SAN heartbeat
- Persistent Event Log
- Resource Prioritisation
  - Deterministic load order
- Resource Management
  - Online on creation
  - If offline, online to any node (not tied to preferred node)
  - Resource UPTIME information
- 2-Node Split Brain Fix

More Information

Product Information

http://www.Novell.COM/products

Documentation

http://www.Novell.COM/documentation

Profession Education Programmes (PEP)

Advanced Technical Training (3 days)

http://www.Novell.COM/offices/emea/education/

Customized Training

Virginie Alfonsi (VAlfonsi@Novell.COM)