

AlliedWare™ OS

## How To | Configure Microsoft Windows 2003 Network Load Balancing Clustering with Allied Telesis Switches

### Introduction

---

This is a guide to Network Load Balancing (NLB) clustering options with Allied Telesis managed layer 3 devices.

NLB is one of the clustering technologies available from Microsoft. It provides high availability for services such as HTTP and FTP, by grouping identical servers into a *cluster* and sharing the network load between all currently-available servers in the cluster.

This How To Note uses scenarios that have been tested in the lab and verified to work. It also deals exclusively with nodes that have dual network cards on the cluster side for bandwidth redundancy and/or aggregation.

This document describes best-practice implementations of NLB clustering on Allied Telesis switches. You should read through this document and decide which setup best suits your requirements. Please pay attention to the details, because some traps in setting these scenarios up can cause unexpected results.

## What information will you find in this document?

This document is divided into the following sections:

- ["Network Load Balancing clustering modes"](#) on page 3
- ["Accepting ARP entries that have a multicast MAC address and a unicast IP address"](#) on page 6
- ["Note about capturing traffic on the clustered servers"](#) on page 7
- ["How to quickly set up Windows 2003 NLB clustering"](#) on page 7
- ["Switch Fault Tolerance, two switches and VRRP"](#) on page 8
- ["Switch Fault Tolerance with a SwitchBlade"](#) on page 17
- ["Adaptive Load Balancing, teamed NICs and a SwitchBlade"](#) on page 19
- ["Adaptive Load Balancing, two switches and VRRP"](#) on page 22
- ["Link Access/Aggregation Control Protocol and a SwitchBlade"](#) on page 29
- ["Static Link Aggregation with a SwitchBlade"](#) on page 34
- ["How Allied Telesis recommends NLB be implemented at layer 3"](#) on page 40
- ["Appendix 1 —Quick Start Guide to setting up a Windows 2003 NLB cluster"](#) on page 43
- ["Appendix 2: IGMP snooping port membership information in switch memory tables"](#) on page 81

## Which product and software version does this information apply to?

The specified solutions only apply to AT-9800 series and SwitchBlade 4000 series, except for the solution in ["How Allied Telesis recommends NLB be implemented at layer 3"](#) on page 40. In that solution, the layer 2 switches can be any Allied Telesis switch and the layer 3 device can be any AR400 or AR700 router.

It applies to AlliedWare versions since 2.7.5. The switch or router implementation varies slightly with the software version—see ["Accepting ARP entries that have a multicast MAC address and a unicast IP address"](#) on page 6.

## Network Load Balancing clustering modes

---

There are two modes for Network Load Balancing with Windows 2003: unicast and multicast. Multicast mode has a further option of IGMP Multicasting. This section discusses how these modes work.

Note that clustering only works, in multicast or unicast mode, if all packets sent to the cluster's IP address are sent to all nodes in the cluster. This means that the cluster-side switch must flood traffic to all ports. To reduce traffic congestion, the cluster should ideally be put in its own VLAN. This recommendation applies regardless of which NLB cluster mode you use.

This document focuses on non-IGMP multicast network load balancing in a layer 3 environment. However, IGMP multicasting appears to be an excellent option for layer 2 networks, because traffic to the servers is not seen on other ports unless they send an IGMP report for that group. Traffic to the servers can be protected from prying eyes by using IGMP filtering and static IGMP entries (see *How To Configure IGMP for Multicasting on Routers and Managed Layer 3 Switches* from [www.alliedtelesis.com/resources/literature/howto.aspx](http://www.alliedtelesis.com/resources/literature/howto.aspx)). In a layer 3 environment you can be less concerned about the security of traffic because your servers should be in their own VLAN, i.e. not in the same segment as clients.

In the section "[How Allied Telesis recommends NLB be implemented at layer 3](#)" on page 40, we use multicast mode with IGMP selected.

### Unicast mode

In **unicast mode**, all hosts in the cluster share a single unicast "cluster" MAC address, to go with the cluster IP address. This mode has the disadvantage that it stops cluster hosts from using their own "burned-in" MAC addresses, so hosts cannot contact each other unless you install a second NIC card on each host and use that for intra-cluster communication.

The cluster operation forces the switch to flood all packets that are destined for the cluster, by stopping the switch from learning the cluster MAC address. The switch cannot learn the cluster MAC address because it never appears in the source field of the Ethernet header. Instead, each NIC uses a special unicast MAC address in the source field of the Ethernet header. The cluster MAC address must never be found in the source field of the Ethernet header, because otherwise the switch will learn the cluster MAC address, stop flooding packets to all ports, and only one node in the cluster will receive traffic for the cluster's IP.

In the two-node cluster that this document deals with, both nodes will use the same cluster MAC address when answering ARP requests for the cluster's address, which is 172.16.0.127.

## Multicast mode

In **multicast mode**, hosts in the cluster use their real “burned-in” MAC address in the source field of the Ethernet header. However, they answer ARP requests for the cluster IP address with a multicast MAC in the ARP packet's payload, while the Ethernet header on the ARP reply still has the real MAC address. This allows cluster hosts to contact each other.

The following tcpdump output shows the multicast MAC:

```
14:02:57.814560 arp who-has 172.16.0.127 (Broadcast) tell 172.16.0.1
14:02:57.814910 arp reply 172.16.0.127 is-at 03:bf:ac:10:00:7f
```

Because the ARP reply is a multicast MAC address, Allied Telesis switches will by default ignore it, so that MAC address is never found in the software IP ARP table. The solution to this depends on the software version—see ["Accepting ARP entries that have a multicast MAC address and a unicast IP address"](#) on page 6.

Because this document deals with dual NICs on each node, multicast is often the only NLB mode that these scenarios will work with. The reason is due to a limitation in the network card drivers, which in many of the setups will simply not recognise the unicast MAC address in the destination field of the Ethernet header. In other words the NICs will not process packets destined to the virtual MAC address despite it being the same source MAC they are sending packets from.

In both multicast and unicast mode the mechanism to cause flooding is as such: ARP requests for the cluster IP are answered with the virtual cluster MAC address in the ARP packet's payload. The node must send all packets, including ARP requests and replies, with a different source MAC in the Ethernet header. This prevents the switch from entering the virtual cluster MAC into the forwarding database (FDB). The end result is that traffic destined for the virtual MAC will be flooded to all ports in the VLAN associated with the cluster IP.

---

**Note:** According to Microsoft if clustering is working properly, one node cannot ping another node in the same cluster. This explains why cluster traffic (like MS NLB heartbeats) generated by each node and teaming packets (like Intel ANS Probes) generated by the NIC driver are broadcasted.

---

## Multicast mode with the IGMP option selected

In **multicast mode with the IGMP option selected**, cluster hosts can also contact each other. This mode uses IGMP to prevent the switch from flooding all ports; instead traffic only goes to NLB ports.

This mode also causes the cluster to use its real “burned in” MAC address in the source field of the ethernet header. The cluster will answer ARP requests for the cluster IP address with a multicast MAC in the ARP packet's payload, while the Ethernet header on the ARP reply still has the real MAC address. Note though, that the MAC address is slightly different to the previously discussed multicast mode—the MAC address starts with the bytes 01:00:5e, which still identifies it as an IP-multicast MAC address:

```
16:07:03.643160 arp who-has 172.16.0.127 tell 172.16.0.66
16:07:03.643330 arp reply 172.16.0.127 is-at 01:00:5e:7f:00:7f
```

The Windows client will happily add this IP / MAC combination to its internal ARP table (on a Windows or Linux PC, use “arp -a” to see all ARP entries). All IP traffic destined for 172.16.0.127 is then sent with an ethernet destination of 01:00:5e:7f:00:7f. The cluster node responds with its burned in MAC in the Ethernet source field.

Microsoft's IGMP multicasting mode cleverly turns the concept of IGMP clients and hosts on its head. As far as an Allied Telesis switch is concerned, the clustered servers are seen as IGMP clients and the IGMP multicast data (stream) is sent by the workstations trying to access the cluster. To ensure that the switches do correctly forward the multicast data, the servers send IGMP reports for a group that corresponds to the MAC address they are putting in their ARP responses. Reports are sent frequently by the NLB servers so there is no concern about IGMP entries timing out on the switch.

The switch's IGMP snooping table (the **show igmpsnooping** command) reveals this information:

```
IGMP Snooping
-----
Status ..... Enabled
Disabled All-groups ports ..... None

Vlan Name (vlan id) ..... default (1)
Fast Leave ..... Off
Group List .....

    Group. 239.255.0.127                Entry timeout 234 secs
    Ports  1,6

    Group. 239.255.255.254            Entry timeout 235 secs
    Ports  1,6

Vlan Name (vlan id) ..... vlan2 (2)
Fast Leave ..... Off
Group List .....

    No group memberships.
```

You can also use the memory tables for information about port membership—see ["Appendix 2: IGMP snooping port membership information in switch memory tables"](#) on page 81.

## Accepting ARP entries that have a multicast MAC address and a unicast IP address

Allied Telesis switches ignore multicast MAC addresses by default. If the Allied Telesis device is layer 3 switching between the clients and the servers, you need to enter at least one extra command to get multicast mode to work, depending on the software version. There are three sets of behaviour:

- with the earliest versions, you need to enter the multicast MAC as a static entry, using the command **add ip arp**.
- with all sb275a versions and with versions 2.7.6 and 276-01, you need to enter the command **enable ip mscluster** and then enter the multicast MAC as a static entry, using the command **add ip arp**
- with later 2.7.6 versions and with version 2.8.1 and later, you need to enter the command **enable ip macdisparity**. You do not need to enter the multicast MAC as a static entry.

### Ports in ARP entries

When you add a static ARP entry for a VLAN, you have to also specify a port number.

SwitchBlade and AT-9800 switches ignore the port number and flood the entry to the VLAN. This means that you can use SwitchBlade and AT-9800 switches in all configurations in this How To Note. You can specify any port in the static ARP entry—the number is irrelevant.

Other switches do not flood the entry to the VLAN. This means that you cannot use other switches with configurations that have multiple ports connected to the server cluster. In this How To Note, the layer 3 switches have multiple ports connected to the cluster in all configurations except the solution in ["How Allied Telesis recommends NLB be implemented at layer 3"](#) on page 40.

### Summary

The following table shows these differences.

	Commands required:		
	add ip arp	enable ip mscluster	enable ip macdisparity
<b>AT-9800 series:</b>			
sb-275 to sb275-04	Yes	No	No
sb-276	Yes	Yes	No
sb276-01	Yes	Yes	No
sb276-02	No	No	Yes
≥ sb-281	No	No	Yes
<b>SwitchBlade:</b>			
≥ sb-275a	Yes	Yes	No
<b>Other layer 3 devices:</b>			
2.7.6	Yes	Yes	No
276-01	Yes	Yes	No
≥ 276-02	No	No	Yes
≥ 2.8.1	No	No	Yes

## Note about capturing traffic on the clustered servers

---

The setups in this How To Note not only use network load balancing, but also teamed network cards. For that reason, when debugging a cluster, you should not do packet sniffing on the servers themselves. **The results cannot be trusted.** For example, inter-cluster node (server-to-server) pings should not work. In our setup, ping would succeed when Ethereal was capturing traffic on a node. So, we conclude that packet sniffers conflict with the network interface driver, or Windows cluster software, and produce erroneous results.

## How to quickly set up Windows 2003 NLB clustering

---

Setting up a Windows 2003 Network Load Balancing cluster requires considerable time for research. The ideas are not overly complex but the mechanisms behind it can take time to understand.

The following sections include the switch configurations for a number of NLB options:

- ["Switch Fault Tolerance, two switches and VRRP" on page 8](#)
- ["Switch Fault Tolerance with a SwitchBlade" on page 17](#)
- ["Adaptive Load Balancing, teamed NICs and a SwitchBlade" on page 19](#)
- ["Adaptive Load Balancing, two switches and VRRP" on page 22](#)
- ["Link Access/Aggregation Control Protocol and a SwitchBlade" on page 29](#)
- ["Static Link Aggregation with a SwitchBlade" on page 34](#)
- ["How Allied Telesis recommends NLB be implemented at layer 3" on page 40](#)

This is followed by ["Appendix 1 —Quick Start Guide to setting up a Windows 2003 NLB cluster" on page 43](#). However, be advised that this implementation is not guaranteed to offer the best security or design for your organisation. The appendix will get you started and after you have had some success you must do more research into cluster policies and design to make your cluster operate efficiently.

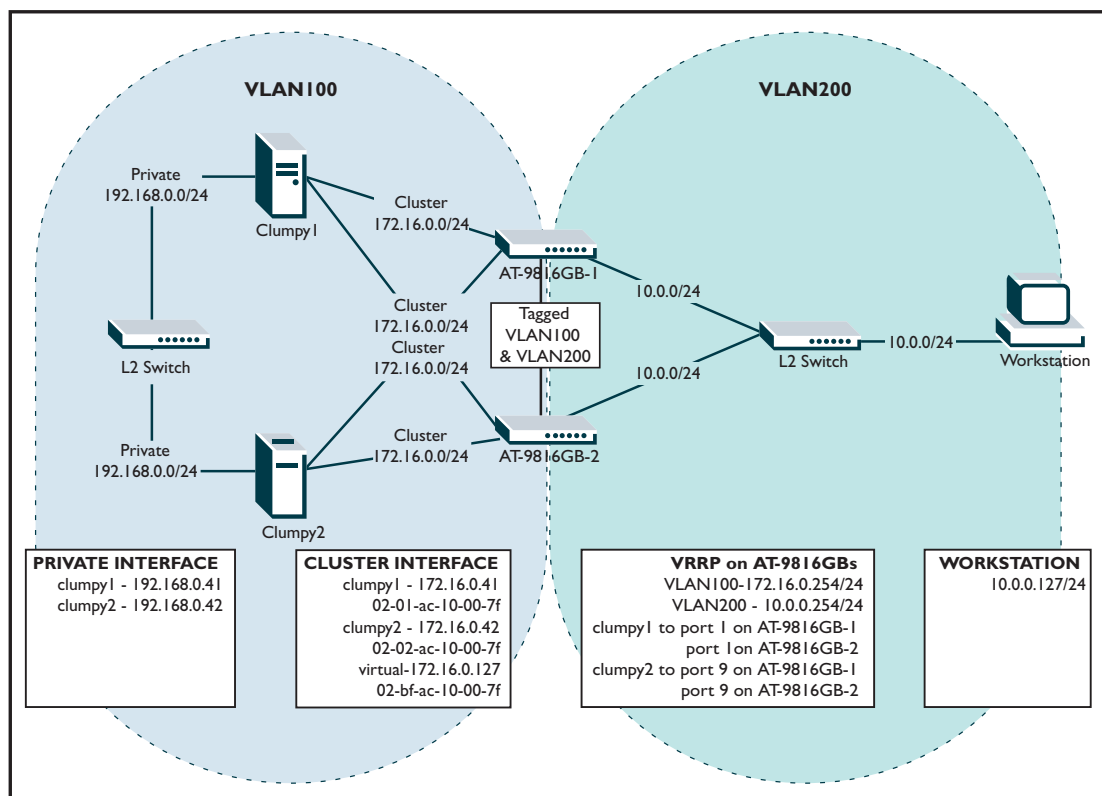
## Switch Fault Tolerance, two switches and VRRP

Switch Fault Tolerance (SFT) provides a failover relationship between two adapters when each adapter is connected to a separate switch. At any given time, one adapter is alive while the other is in standby mode. Using SFT, if one switch fails, then the two switches will maintain node and cluster connectivity without network downtime if each server is attached to both switches.

- Equipment**
- Intel Pro/1000 MT Dual Port NIC
  - Linux PC as the workstation
  - 2 x 32 bit PCs with Windows 2003 Enterprise Edition
  - AT-9924SP as L2 switch connected to workstation
  - 2 x AT-9816GB
  - AT-8748SL as L2 switch on private cluster LAN

If only two servers are being used in the cluster then the AT-8748SL may be left out and the nodes connected directly to each other with a crossover cable. The AT-9924SP also serves as a media adapter so that the UTP NIC can connect to the fibre AT-9816GB.

### Network Diagram





## Unicast Mode

First, we will give a configuration with NLB in unicast mode. This configuration uses 2 server clusters with 1 private and 2 teamed NICs.

Note that multicast mode is favoured over unicast mode, because of Windows 2003 and NIC driver variance. We cannot guarantee that unicast will work at all.

### Unicast VRRP switch configurations

#### AT-9800 #1

```
#
# SYSTEM configuration
#
set system name="AT-9816GB-1"
#
# VLAN general configuration
#
create vlan="vlan100" vid=100
create vlan="vlan200" vid=200
#
# VLAN port configuration
#
add vlan="100" port=1,9
add vlan="200" port=3
add vlan="100" port=2 frame=tagged
add vlan="200" port=2 frame=tagged
delete vlan="1" port=2
#
# IP configuration
#
enable ip
add ip int=vlan100 ip=172.16.0.1 mask=255.255.255.0
add ip int=vlan200 ip=10.0.0.1 mask=255.255.255.0
#
# VRRP configuration
#
enable vrrp
create vrrp=100 over=vlan100 ipaddress=172.16.0.254
add vrrp=100 monitoredinterface=vlan100
create vrrp=200 over=vlan200 ipaddress=10.0.0.254
add vrrp=200 monitoredinterface=vlan200
```

**AT-9800 #2**

```
#
# SYSTEM configuration
#
set system name="AT-9816GB-2"
#
# VLAN general configuration
#
create vlan="vlan100" vid=100
create vlan="vlan200" vid=200
#
# VLAN port configuration
#
add vlan="100" port=1,9
add vlan="200" port=3
add vlan="100" port=2 frame=tagged
add vlan="200" port=2 frame=tagged
#
# IP configuration
#
enable ip
add ip int=vlan100 ip=172.16.0.2 mask=255.255.255.0
add ip int=vlan200 ip=10.0.0.2 mask=255.255.255.0
#
# VRRP configuration
#
enable vrrp
create vrrp=100 over=vlan100 ipaddress=172.16.0.254
add vrrp=100 monitoredinterface=vlan100
create vrrp=200 over=vlan200 ipaddress=10.0.0.254
add vrrp=200 monitoredinterface=vlan200
```

**AT-9924SP**

```
#
# SYSTEM configuration
#
set system name="AT-9924SP"
#
# VLAN general configuration
#
create vlan="vlan100" vid=100
add vlan="100" port=1-8
create vlan="vlan200" vid=200
add vlan="200" port=9-12
create vlan="vlan300" vid=300
add vlan="300" port=13-16
create vlan="vlan400" vid=400
add vlan="400" port=17-20
create vlan="vlan500" vid=500
add vlan="500" port=21-24
```

## Table Entries

Here is an example of the entries that should appear in the switch FDB and ARP tables when unicast mode is used with NLB clustering.

```
Manager top> sh swi tab=ip
```

```
CXE IP route table:
```

Index	Seg	Network	Nexthop	ToProc	VRP	VID	Hits	PV	Prio	Perm
0	101	172.16.0.42	<u>02-bf-ac-10-00-7f 0</u>	0	100	0	0	0	0	1
1	101	10.0.0.127	00-50-fc-ee-f5-13 0	1	200	0	0	0	0	1
2	101	172.16.0.127	<u>02-bf-ac-10-00-7f 0</u>	0	100	0	0	0	0	1
3	101	172.16.0.41	<u>02-bf-ac-10-00-7f 0</u>	0	100	0	0	0	0	1
4	102	172.16.0.1	00-00-cd-09-73-50 0	0	100	0	0	0	0	1
5	102	172.16.0.255	00-00-cd-09-73-50 0	0	100	0	0	0	0	1
6	102	10.0.0.1	00-00-cd-09-73-50 0	1	200	0	0	0	0	1
7	102	10.0.0.255	00-00-cd-09-73-50 0	1	200	0	0	0	0	1
8	110	172.16.0.0	00-00-cd-09-73-50 0	0	100	0	0	0	0	1
9	110	10.0.0.0	00-00-cd-09-73-50 0	1	200	0	0	0	0	1

The clustered IP's MAC has been learnt and the AT-9816GB will make routing decisions to 172.16.0.41/42/127 based on the MAC address 02-bf-ac-10-00-7f, however, if we look at the switch's FDB:

```
Manager top> sh swi fdb
```

```
Switch Forwarding Database (hardware)
```

```
-----  
Total Number of Entries = 8  
-----
```

VLAN	MAC Address	Port	Status
1	00-00-cd-09-73-50	CPU	static
100	00-00-cd-09-73-50	CPU	static
200	00-00-cd-09-73-50	CPU	static
100	00-00-5e-00-01-64	CPU	static
200	00-00-5e-00-01-c8	CPU	static
100	<u>02-01-ac-10-00-7f</u>	1	dynamic
100	<u>02-02-ac-10-00-7f</u>	9	dynamic
200	00-50-fc-ee-f5-13	3	dynamic

The table above illustrates that the MAC address 02-bf-ac-10-00-7f does not appear in the switch chip's table, so a packet with a destination MAC address of 02-bf-ac-10-00-7f must be flooded out all ports in the VLAN. Notice that the two highlighted MAC entries are very similar but not the same as the cluster's MAC address. 02-01 tells us that it is a clustering MAC address for the node with priority 1. 02-02 tells us that it is a clustering MAC address for the node with priority 2.

## Multicast Mode

This is the same setup but this time with NLB clustering in multicast mode.

In the following configurations, we have added a static ARP entry. For more recent software versions, you need to also use the command **enable ip mscluster**. For the most recent software versions, you need to use the command **enable ip macdisparity**, but do not need the static ARP entry. For details, see ["Accepting ARP entries that have a multicast MAC address and a unicast IP address" on page 6](#). If you need a static ARP entry, you can see the specific multicast address to use in the NLB Manager (see ["Set up the NLB Cluster" on page 66](#)).

### Multicast VRRP switch configurations

#### AT-9800 #1

```
#
# SYSTEM configuration
#
set system name="AT-9816GB-1"
#
# VLAN general configuration
#
create vlan="vlan100" vid=100
create vlan="vlan200" vid=200
#
# VLAN port configuration
#
add vlan="100" port=1,9
add vlan="200" port=3
add vlan="100" port=2 frame=tagged
add vlan="200" port=2 frame=tagged
delete vlan="1" port=2
#
# IP configuration
#
enable ip
add ip int=vlan100 ip=172.16.0.1 mask=255.255.255.0
add ip int=vlan200 ip=10.0.0.1 mask=255.255.255.0
add ip arp=172.16.0.127 int=vlan100 eth=03-bf-ac-10-00-7f port=1
#
# VRRP configuration
#
enable vrrp
create vrrp=100 over=vlan100 ipaddress=172.16.0.254
add vrrp=100 monitoredinterface=vlan100
create vrrp=200 over=vlan200 ipaddress=10.0.0.254
add vrrp=200 monitoredinterface=vlan200
```

---

**Note:** The MAC address is now slightly different, as shown by the underlined text above.

---

**AT-9800 #2**

```
# SYSTEM configuration
#
set system name="AT-9816GB-2"
#
# VLAN general configuration
#
create vlan="vlan100" vid=100
create vlan="vlan200" vid=200
#
# VLAN port configuration
#
add vlan="100" port=1,9
add vlan="200" port=3
add vlan="100" port=2 frame=tagged
add vlan="200" port=2 frame=tagged
# IP configuration
#
enable ip
add ip int=vlan100 ip=172.16.0.2 mask=255.255.255.0
add ip int=vlan200 ip=10.0.0.2 mask=255.255.255.0
add ip arp=172.16.0.127 int=vlan100 eth=03-bf-ac-10-00-7f port=1#
# VRRP configuration
#
enable vrrp
create vrrp=100 over=vlan100 ipaddress=172.16.0.254
add vrrp=100 monitoredinterface=vlan100
create vrrp=200 over=vlan200 ipaddress=10.0.0.254
add vrrp=200 monitoredinterface=vlan200
```

**AT-9924SP**

```
#
# SYSTEM configuration
#
set system name="AT-9924SP"
#
# VLAN general configuration
#
create vlan="vlan100" vid=100
add vlan="100" port=1-8
create vlan="vlan200" vid=200
add vlan="200" port=9-12
create vlan="vlan300" vid=300
add vlan="300" port=13-16
create vlan="vlan400" vid=400
add vlan="400" port=17-20
create vlan="vlan500" vid=500
add vlan="500" port=21-24
```

## Table Entries

The following tables are the contents of the switch tables for this scenario; there is an entry in the software table for the cluster: IP 172.16.0.127. This has been statically added. There is no entry for the MAC 03-bf-ac-10-00-7f in the forwarding database so traffic will be flooded to all ports in the VLAN that is attached to 172.16.0.0/24.

Note that in this case the individual server IPs (172.16.0.41-42) appear in the ARP table with different MACs and these MACs also appear in the FDB.

```
Manager AT-9816GB-1> sh swi fdb
```

```
Switch Forwarding Database (hardware)
```

```
-----  
Total Number of Entries = 7  
-----
```

VLAN	MAC Address	Port	Status
1	00-00-cd-09-73-50	CPU	static
100	00-00-cd-09-73-50	CPU	static
200	00-00-cd-09-73-50	CPU	static
100	00-00-5e-00-01-64	CPU	static
200	00-00-5e-00-01-c8	CPU	static
100	00-04-23-b5-bb-6f	1	dynamic
100	00-04-23-ab-37-02	2	dynamic

```
Manager AT-9816GB-1> sh ip arp
```

Interface	IP Address	Physical Address	ARP Type	Status
vlan200(3)	10.0.0.127	00-50-fc-ee-f5-13	Dynamic	active
vlan200	10.0.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan100(1)	172.16.0.41	00-04-23-b5-bb-6f	Dynamic	active
vlan100(2)	172.16.0.42	00-04-23-ab-37-02	Dynamic	active
vlan100(1)	172.16.0.127	<u>03-bf-ac-10-00-7f</u>	Static	active
vlan100	172.16.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan100	255.255.255.255	ff-ff-ff-ff-ff-ff	Other	active

## Traffic Flows

Regardless of whether NLB is in unicast or multicast mode, the following traffic flows should apply. Interestingly, both servers will respond to pings to the cluster address 172.16.0.127. This is not the fault of the switch. HTTP balancing works correctly. HTTP load balancing at the packet level follows<sup>1</sup>. Both of these captures are taken from the perspective of clumpy1. The first capture shows the packet flow when clumpy1 answers the HTTP request (in accordance with the defined clustering rules). The second capture is also taken from the perspective of clumpy1 but this time the HTTP request is serviced by clumpy2, so only packets from the workstation can be seen.

### Packet Capture 1

HTTP request from 10.0.0.127 to 172.16.0.127 and served by clumpy1 (172.16.0.41) as seen by clumpy 1:

```
15:53:00.219277 IP 10.0.0.127.40195 > 172.16.0.127.http: SWE 593168138:593168138(0) win
5840 <mss 1460,sackOK,timestamp 148937549 0,nop,wscale 0>
15:53:00.219650 IP 172.16.0.127.http > 10.0.0.127.40195: S 3107871426:3107871426(0) ack
593168139 win 65535 <mss 1460,nop,wscale 0,nop,nop,timestamp 0 0,nop,nop,sackOK>
15:53:00.219723 IP 10.0.0.127.40195 > 172.16.0.127.http: . ack 1 win 5840
<nop,nop,timestamp 148937549 0>
15:53:00.220206 IP 10.0.0.127.40195 > 172.16.0.127.http: P 1:274(273) ack 1 win 5840
<nop,nop,timestamp 148937549 0>
15:53:00.220717 IP 172.16.0.127.http > 10.0.0.127.40195: FP 1:393(392) ack 274 win 65262
<nop,nop,timestamp 229906 148937549>
15:53:00.221659 IP 10.0.0.127.40195 > 172.16.0.127.http: F 274:274(0) ack 394 win 6432
<nop,nop,timestamp 148937549 229906>
15:53:00.221777 IP 172.16.0.127.http > 10.0.0.127.40195:.. ack 275 win 65262
<nop,nop,timestamp 229906 148937549>
```

### Packet Capture 2

An identical HTTP request from 10.0.0.127 to 172.16.0.127 but served by clumpy2 (172.16.0.42) as seen by clumpy1 (172.16.0.41):

```
15:52:56.519534 IP 10.0.0.127.40194 > 172.16.0.127.http: . ack 1481239454 win 5840
<nop,nop,timestamp 148937179 0>
15:52:56.520006 IP 10.0.0.127.40194 > 172.16.0.127.http: P 0:273(273) ack 1 win 5840
<nop,nop,timestamp 148937179 0>
15:52:56.520693 IP 10.0.0.127.40194 > 172.16.0.127.http: . ack 583 win 6972
<nop,nop,timestamp 148937179 230102>
15:52:56.521737 IP 10.0.0.127.40194 > 172.16.0.127.http: F 273:273(0) ack 583 win 6972
<nop,nop,timestamp 148937179 230102>
```

<sup>1</sup>The active port that clumpy1 was connected to on the switch was mirrored and the mirror port was then sniffed by a linux server with tcpdump version 3.8.3 (libpcap version 0.8.3)

## NIC Failover

Should the node's link go down on the NIC driver will rapidly change the state of the standby NIC to active. When operating correctly the node should experience only a few seconds of lost connectivity. If there is a failure further downstream but the link at the NIC end does not go down then the NIC will not change state because it is not aware of any topology change.

Probes can be used to determine link failures however; in the configuration here they will not help. In fact, in our lab setup with Intel ANS Probes, the NIC driver disabled the wrong NIC when a downstream link was broken. So in this setup, there appears to be no advantage to enabling probes.

## Summary

This setup is an excellent redundancy option. While the combined bandwidth of the NICs is never utilised, failover is fast if the link to the active NIC goes down. This is not a solution for adapting to failures that occur further down the network; however, there is no solution on offer in this document for such a scenario.



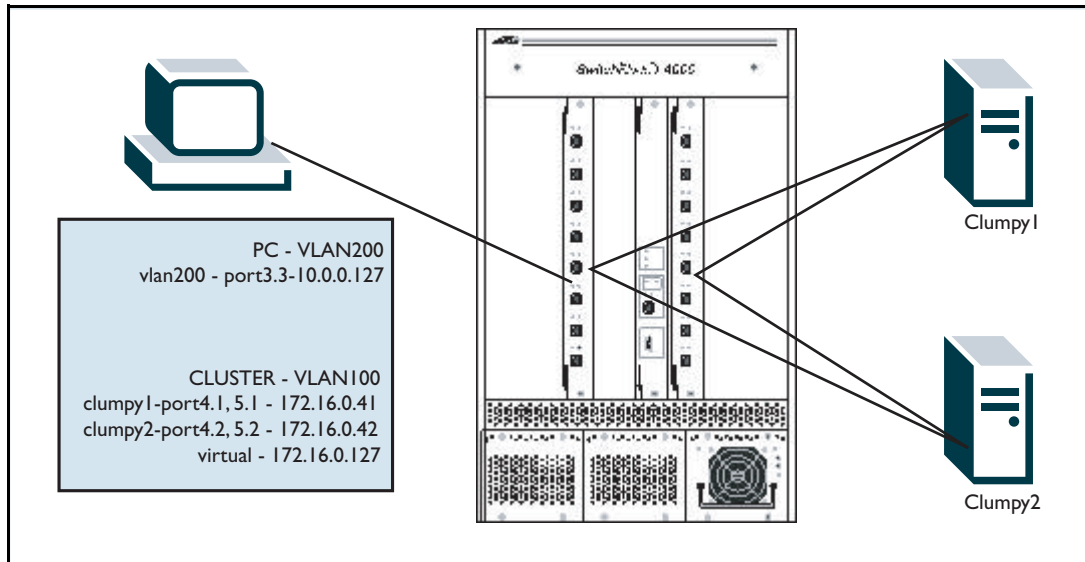
# Switch Fault Tolerance with a SwitchBlade

Switch Fault Tolerance (SFT) provides a failover relationship between two adaptors when each adaptor is connected to a separate switch. At any given time, one adaptor is alive while the other is in standby mode. Using SFT and one SwitchBlade will permit you to hot swap line cards without network downtime if each server is attached to two different line cards.

**Equipment** This example uses:

- Intel Pro/1000 MT Dual Port NIC
- SwitchBlade chassis (AT-SB4108-00)
  - Switch Control Card (AT-SB4211)
  - 2 x 8 port UTP trispeed line card (AT-SB44118-100 (RJ45))
- 2 x 32-bit PCs running Windows 2003 Enterprise Edition

**Network Diagram**



## Configuration

Set the Windows 2003 NLB cluster servers to use multicast mode. Remember to add a static ARP for the multicast MAC, and to use the command **enable ip mscluster**.

```
set system name="Switchblade"

create vlan="vlan100" vid=100
create vlan="vlan200" vid=200

add vlan="100" port=4.1-4.2,5.1-5.2
add vlan="200" port=4.3

enable ip
add ip int=vlan100 ip=172.16.0.254 mask=255.255.255.0
add ip int=vlan200 ip=10.0.0.254 mask=255.255.255.0
enable ip mscluster
add ip arp=172.16.0.127 int=vlan100 eth=03-bf-ac-10-00-7f port=4.1
```

## Description

As in all the other cases where NLB multicast mode is used, there should never be an entry in the forwarding database for the multicast MAC. The static ARP entry will, of course, appear in the IP ARP table.

## Summary

This setup is an excellent redundancy option. While the combined bandwidth of the NICs is never utilised, failover is fast if the link to the active NIC goes down. This is not a solution for adapting to failures that occur further down the network; however, there is no solution on offer in this document for such a scenario. This setup will give you full redundancy (both at the cluster and node level) when you hot swap a SwitchBlade line card.

## Adaptive Load Balancing, teamed NICs and a SwitchBlade

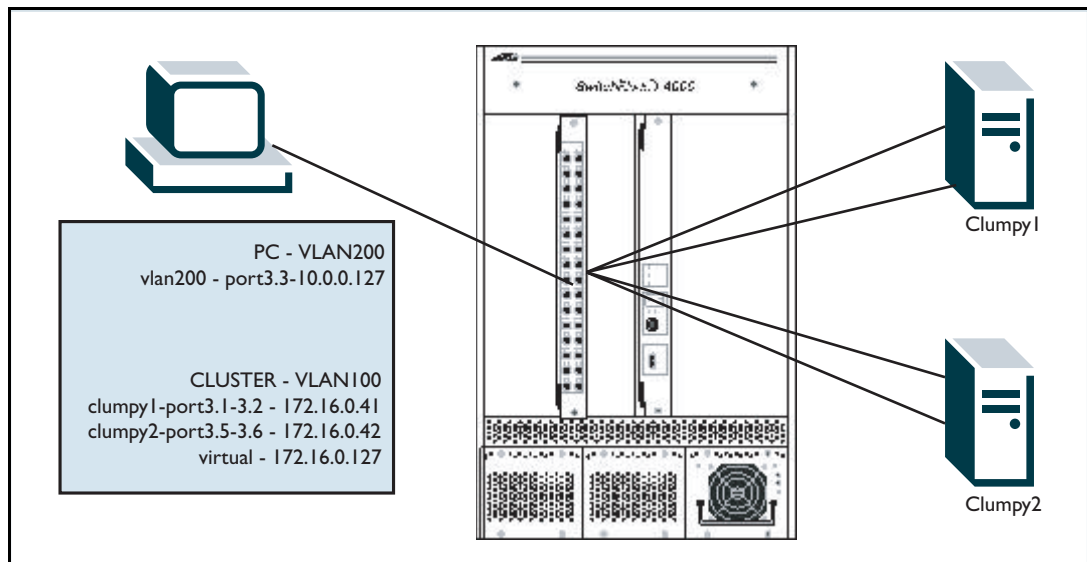
This section describes Adaptive Load Balancing (ALB). ALB provides load balancing and adapter fault tolerance. The NIC driver should analyse the send and receive load on each adapter and balance the rate across them based on the NIC vendor's algorithm/rule set (for example Intel base it on the destination address).

If you want to provide load balancing with dual NICs in your cluster, then you may instead use LACP, which is discussed in "[Link Access/Aggregation Control Protocol and a SwitchBlade](#)" on page 29. Another alternative is Static Link Aggregation, which load balances at the server level and is discussed in "[Static Link Aggregation with a SwitchBlade](#)" on page 34. ALB is a less efficient but easy-to-implement solution.

**Equipment** This example uses:

- Intel Pro/1000 MT Dual Port NIC
- SwitchBlade chassis (AT-SB4108-00)
- Switch Control Card (AT-SB4211)
- 8 port UTP trispeed line card (AT-SB4411 8-1000T(RJ45))
- 2 x 32-bit PCs running Windows 2003 Enterprise Edition

### Network diagram



## Configuration

Set the Windows 2003 NLB cluster servers to use multicast mode. Remember to add a static ARP for the multicast MAC, and to use the command **enable ip mscluster**.

```
set system name="Switchblade"

create vlan="vlan100" vid=100
create vlan="vlan200" vid=200

add vlan="100" port=3.1-3.2,3.5-3.6
add vlan="200" port=3.3

enable ip
add ip int=vlan100 ip=172.16.0.254 mask=255.255.255.0
add ip int=vlan200 ip=10.0.0.254 mask=255.255.255.0
enable ip mscluster
add ip arp=172.16.0.127 int=vlan100 eth=03-bf-ac-10-00-7f port=3.1
```

For better performance and a less chatty network, turn off probes and Receive Load Balancing (RLB) on your NIC team.

## Description

The SwitchBlade will send any traffic destined for 172.16.0.127 to all ports in the VLAN that belongs to the 172.16.0.0/24 subnet, including both teamed NICs on both servers. This setup is therefore less efficient than LACP or Static Link Aggregation, which would only have the SwitchBlade send one packet to one port in each team/trunk.

## Packet Captures

The following packet capture illustrates the duplicate packets that are sent to the node. Notice that only traffic from 10.0.0.127 is duplicated (marked in red). Traffic from 172.16.0.127 (in this case clumpy1) is seen only once (marked in blue).<sup>1</sup>

```
15:59:00.609752 IP (tos 0x0, ttl 63, id 23053, offset 0, flags [DF], length: 60)
10.0.0.127.42721 > 172.16.0.127.http: SWE [tcp sum ok] 3360891801:3360891801(0) win 5840
<mss 1460,sackOK,timestamp 244021652 0,nop,wscale 0>
15:59:00.609757 IP (tos 0x0, ttl 63, id 23053, offset 0, flags [DF], length: 60)
10.0.0.127.42721 > 172.16.0.127.http: SWE [tcp sum ok] 3360891801:3360891801(0) win 5840
<mss 1460,sackOK,timestamp 244021652 0,nop,wscale 0>

15:59:00.609936 IP (tos 0x0, ttl 128, id 17555, offset 0, flags [none], length:64)
172.16.0.127.http > 10.0.0.127.42721: S [tcp sum ok] 1420104149:1420104149(0) ack
3360891802 win 16384 <mss 1460,nop,wscale 0,nop,nop,timestamp 0 0,nop,nop, sackOK>

15:59:00.610002 IP (tos 0x0, ttl 63, id 23054, offset 0, flags [DF], length: 52)
10.0.0.127.42721 > 172.16.0.127.http: . [tcp sum ok] 1:1(0) ack 1 win 5840
<nop,nop,timestamp 244021652 0>
15:59:00.610006 IP (tos 0x0, ttl 63, id 23054, offset 0, flags [DF], length: 52)
10.0.0.127.42721 > 172.16.0.127.http: . [tcp sum ok] 1:1(0) ack 1 win 5840
<nop,nop,timestamp 244021652 0>

<continued on next page>
```

<continued from previous page>

```

15:59:00.612316 IP (tos 0x0, ttl 63, id 23056, offset 0, flags [DF], length: 52)
10.0.0.127.42721 > 172.16.0.127.http: F [tcp sum ok] 274:274(0) ack 397 win 6432
<nop,nop,timestamp 244021652 843939>
15:59:00.612321 IP (tos 0x0, ttl 63, id 23056, offset 0, flags [DF], length: 52)
10.0.0.127.42721 > 172.16.0.127.http: F [tcp sum ok] 274:274(0) ack 397 win 6432
<nop,nop,timestamp 244021652 843939>

15:59:00.612535 IP (tos 0x0, ttl 128, id 17557, offset 0, flags [DF], length: 52)
172.16.0.127.http > 10.0.0.127.42721: . [tcp sum ok] 397:397(0) ack 275 win 65262
<nop,nop,timestamp 843939 244021652>

```

## Unicast Mode

Setting your cluster to use unicast with ALB will lead to connection timeouts. The switch chip learns the cluster MAC thanks to an MS-Load balancing packet that has the Ethernet source MAC field set to the cluster MAC (this should not happen). Unfortunately, this ruins all the effort that the clustering process had previously put into hiding the cluster MAC from the switch and will stop the switch from flooding packets that are destined for the cluster MAC.

## Summary

This solution provides NIC redundancy and transmit-aggregation. It will not be tolerant to failures downstream but the cluster should recover quickly if one or all ports fail on a node. This is easy to implement but not as efficient as link aggregation.

---

I. The two ALB ports that clumpy1 was connected to on the switch were mirrored and the mirror port was then sniffed by a linux server with tcpdump version 3.8.3 (libpcap version 0.8.3).

## Adaptive Load Balancing, two switches and VRRP

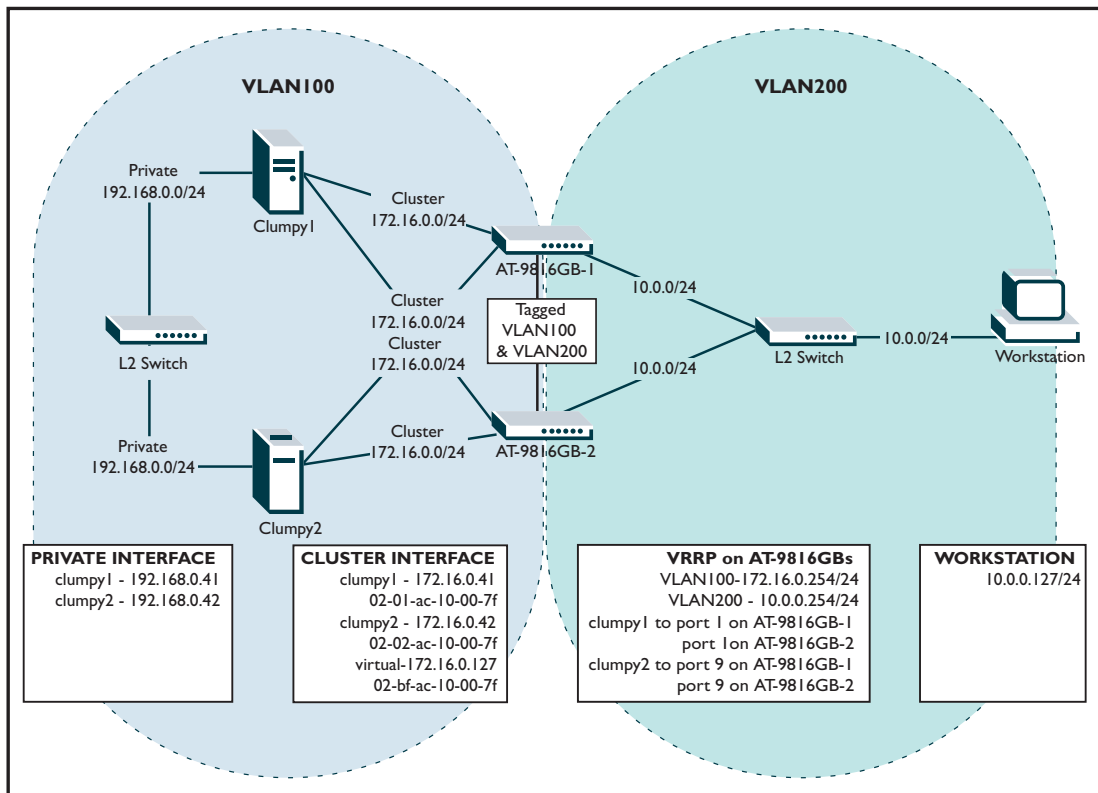
This section describes Adaptive Load Balancing (ALB). ALB provides load balancing and adapter fault tolerance. The NIC driver should analyse the send and receive load on each adapter and balance the rate across them based on the NIC vendor's algorithm/rule set (for example Intel base it on the destination address).

If you want to provide load balancing with dual NICs in your cluster, then you may instead use LACP, which is discussed in "[Link Access/Aggregation Control Protocol and a SwitchBlade](#)" on [page 29](#). Another alternative is Static Link Aggregation, which load balances at the server level and is discussed in "[Static Link Aggregation with a SwitchBlade](#)" on [page 34](#). ALB is a less efficient but easy to implement solution.

- Equipment**
- Intel Pro/1000 MT Dual Port NIC
  - Linux PC as the workstation
  - 2 x 32 bit PCs with Windows 2003 Enterprise Edition
  - AT-9924SP as L2 switch connected to workstation
  - 2 x AT-9816GB
  - AT-8748SL as L2 switch on private cluster LAN

If only two servers are being used in the cluster, then the AT-8748SL may be left out and the nodes connected directly to each other with a crossover cable. The AT-9924SP also serves as a media adapter so that the UTP NIC can connect to the fibre AT-9816GB.

### Network diagram



## Configuration

Again, we will only use multicast clustering here.

In the following configurations, we have added a static ARP entry to both switches. For more recent software versions, you need to also use the command **enable ip mscluster**. For the most recent software versions, you need to use the command **enable ip macdisparity**, but do not need the static ARP entry. For details, see ["Accepting ARP entries that have a multicast MAC address and a unicast IP address" on page 6](#). If you need a static ARP entry, you can see the specific multicast address to use in the NLB Manager (see ["Set up the NLB Cluster" on page 66](#)).

### AT-9800 #1

```
#
# SYSTEM configuration
#
set system name="AT-9816GB-1"
#
# VLAN general configuration
#
create vlan="vlan100" vid=100
create vlan="vlan200" vid=200
#
# VLAN port configuration
#
add vlan="100" port=1,9
add vlan="200" port=3
add vlan="100" port=2 frame=tagged
add vlan="200" port=2 frame=tagged
delete vlan="1" port=2
#
# IP configuration
#
enable ip
add ip int=vlan100 ip=172.16.0.1 mask=255.255.255.0
add ip int=vlan200 ip=10.0.0.1 mask=255.255.255.0
add ip arp=172.16.0.127 int=vlan100 eth=03-bf-ac-10-00-7f port=1
#
# VRRP configuration
#
enable vrrp
create vrrp=100 over=vlan100 ipaddress=172.16.0.254
add vrrp=100 monitoredinterface=vlan100
create vrrp=200 over=vlan200 ipaddress=10.0.0.254
add vrrp=200 monitoredinterface=vlan200
```

**AT-9800 #2**

```

#
# SYSTEM configuration
#
set system name="AT-9816GB-2"
#
# VLAN general configuration
#
create vlan="vlan100" vid=100
create vlan="vlan200" vid=200
#
# VLAN port configuration
#
add vlan="100" port=1,9
add vlan="200" port=3
add vlan="100" port=2 frame=tagged
add vlan="200" port=2 frame=tagged
#
# IP configuration
#
enable ip
add ip int=vlan100 ip=172.16.0.2 mask=255.255.255.0
add ip int=vlan200 ip=10.0.0.2 mask=255.255.255.0
add ip arp=172.16.0.127 int=vlan100 eth=03-bf-ac-10-00-7f port=1
#
# VRRP configuration
#
enable vrrp
create vrrp=100 over=vlan100 ipaddress=172.16.0.254
add vrrp=100 monitoredinterface=vlan100
create vrrp=200 over=vlan200 ipaddress=10.0.0.254
add vrrp=200 monitoredinterface=vlan200

```

**AT-9924SP**

```

#
# SYSTEM configuration
#
set system name="AT-9924SP"
#
# VLAN general configuration
#
create vlan="vlan100" vid=100
add vlan="100" port=1-8
create vlan="vlan200" vid=200
add vlan="200" port=9-12
create vlan="vlan300" vid=300
add vlan="300" port=13-16
create vlan="vlan400" vid=400
add vlan="400" port=17-20
create vlan="vlan500" vid=500
add vlan="500" port=21-24

```

**Tip:** For better performance and a less chatty network, turn off probes and Receive Load Balancing (RLB) on your NIC team.



## Description

In ALB mode, one of the NICs will answer an ARP request for the virtual IP address with its burned-in MAC. This behaviour will once again defeat the way clustering should work because the cluster IP ARP entry has a MAC address found in the forwarding database. This is avoided by entering a static IP ARP entry.

### This is wrong:

```
Manager AT-9816GB-1> sh ip arp
```

Interface	IP Address	Physical Address	ARP Type	Status
vlan200(3)	10.0.0.127	00-50-fc-ee-f5-13	Dynamic	active
vlan200	10.0.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan100(2)	172.16.0.41	00-04-23-b5-bb-6f	Dynamic	active
vlan100(9)	172.16.0.42	00-04-23-ab-37-02	Dynamic	active
vlan100(2)	172.16.0.127	00-04-23-b5-bb-6f	Dynamic	active
vlan100	172.16.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan200	255.255.255.255	ff-ff-ff-ff-ff-ff	Other	active

```
Manager AT-9816GB-1> sh swi fdb
```

```
Switch Forwarding Database (hardware)
```

```
-----  
Total Number of Entries = 8  
-----
```

VLAN	MAC Address	Port	Status
1	00-00-cd-09-73-50	CPU	static
100	00-00-cd-09-73-50	CPU	static
200	00-00-cd-09-73-50	CPU	static
100	00-04-23-b5-bb-6f	2	dynamic
100	00-04-23-ab-37-02	9	dynamic
100	00-04-23-ab-37-03	2	dynamic
100	00-04-23-b5-bb-6e	1	dynamic
200	00-50-fc-ee-f5-13	3	dynamic

### This is right:

```
Manager AT-9816GB-1> sh ip arp
```

Interface	IP Address	Physical Address	ARP Type	Status
vlan200	10.0.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan100(9)	172.16.0.42	00-04-23-ab-37-02	Dynamic	active
vlan100(1)	172.16.0.127	<u>03-bf-ac-10-00-7f</u>	Static	active
vlan100	172.16.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan200	255.255.255.255	ff-ff-ff-ff-ff-ff	Other	active

Example output continues on the next page.

```
Manager AT-9816GB-1> sh swi fdb
```

```
Switch Forwarding Database (hardware)
```

```
-----  
Total Number of Entries = 10  
-----
```

VLAN	MAC Address	Port	Status
1	00-00-cd-09-73-50	CPU	static
100	00-00-cd-09-73-50	CPU	static
200	00-00-cd-09-73-50	CPU	static
100	00-00-5e-00-01-64	CPU	static
200	00-00-5e-00-01-c8	CPU	static
100	00-04-23-b5-bb-6f	2	dynamic
100	00-04-23-ab-37-02	9	dynamic
100	00-04-23-ab-37-03	2	dynamic
100	00-04-23-b5-bb-6e	1	dynamic
200	00-50-fc-ee-f5-13	3	dynamic

```
Manager AT-9816GB-2> sh ip arp
```

Interface	IP Address	Physical Address	ARP Type	Status
vlan200(2)	10.0.0.127	00-50-fc-ee-f5-13	Dynamic	active
vlan200(2)	10.0.0.254	00-00-5e-00-01-c8	Dynamic	active
vlan200	10.0.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan100(1)	172.16.0.41	00-04-23-b5-bb-6f	Dynamic	active
vlan100(2)	172.16.0.42	00-04-23-ab-37-02	Dynamic	active
vlan100(1)	172.16.0.127	<u>03-bf-ac-10-00-7f</u>	Static	active
vlan100(2)	172.16.0.254	00-00-5e-00-01-64	Dynamic	active
vlan100	172.16.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan200	255.255.255.255	ff-ff-ff-ff-ff-ff	Other	active

```
Manager AT-9816GB-2> sh swi fdb
```

```
Switch Forwarding Database (hardware)
```

```
-----  
Total Number of Entries = 11  
-----
```

VLAN	MAC Address	Port	Status
1	00-00-cd-03-00-e8	CPU	static
100	00-00-cd-03-00-e8	CPU	static
200	00-00-cd-03-00-e8	CPU	static
100	00-04-23-b5-bb-6f	1	dynamic
100	00-04-23-ab-37-02	2	dynamic
100	00-04-23-ab-37-03	9	dynamic
100	00-04-23-b5-bb-6e	2	dynamic
100	00-00-cd-09-73-50	2	dynamic
100	00-00-5e-00-01-64	2	dynamic
200	00-00-5e-00-01-c8	2	dynamic
200	00-50-fc-ee-f5-13	2	dynamic

At this point we can observe an interesting phenomenon. Remember that we added the static IP ARP entry to port 1 on both switches. One port on each node's NIC is connected to port 1 of a switch. The second port on each node's NIC is attached to port 9 on the other

switch. In the following example, one of the switches now cannot ping the virtual server address:

```
Manager AT-9816GB-1> ping 172.16.0.127

Echo reply 1 from 172.16.0.127 time delay 0 ms
Echo reply 2 from 172.16.0.127 time delay 0 ms
Echo reply 3 from 172.16.0.127 time delay 0 ms
Echo reply 4 from 172.16.0.127 time delay 0 ms
Echo reply 5 from 172.16.0.127 time delay 0 ms

Manager AT-9816GB-2> ping 172.16.0.127

Request 1 timed-out: No reply from 172.16.0.127
Request 2 timed-out: No reply from 172.16.0.127
Request 3 timed-out: No reply from 172.16.0.127
Request 4 timed-out: No reply from 172.16.0.127
Request 5 timed-out: No reply from 172.16.0.127
```

To understand the reason for this, examine the following fix:

```
Manager AT-9816GB-2> del ip arp=172.16.0.127

Info (1005272): ARP entry successfully deleted.

Manager AT-9816GB-2> add ip arp=172.16.0.127 ether=03-bf-ac-10-00-7f
int=vlan100
port=9

Info (1005275): ARP successfully added.

Manager AT-9816GB-2> ping 172.16.0.127

Echo reply 1 from 172.16.0.127 time delay 0 ms
Echo reply 2 from 172.16.0.127 time delay 0 ms
Echo reply 3 from 172.16.0.127 time delay 0 ms
Echo reply 4 from 172.16.0.127 time delay 0 ms
Echo reply 5 from 172.16.0.127 time delay 0 ms
```

Example output continues on the next page:

```
Manager AT-9816GB-2> sh ip arp
```

Interface	IP Address	Physical Address	ARP Type	Status
vlan200	10.0.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan100(2)	172.16.0.41	00-04-23-b5-bb-6f	Dynamic	active
vlan100(9)	172.16.0.42	00-04-23-ab-37-02	Dynamic	active
vlan100(9)	172.16.0.127	<u>03-bf-ac-10-00-7f</u>	Static	active
vlan100	172.16.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan200	255.255.255.255	ff-ff-ff-ff-ff-ff	Other	active

```
Manager AT-9816GB-2> sh swi fdb
```

```
Switch Forwarding Database (hardware)
```

```
-----  
Total Number of Entries = 7  
-----
```

VLAN	MAC Address	Port	Status
1	00-00-cd-03-00-e8	CPU	static
100	00-00-cd-03-00-e8	CPU	static
200	00-00-cd-03-00-e8	CPU	static
100	00-04-23-b5-bb-6f	2	dynamic
100	00-04-23-ab-37-02	9	dynamic
100	00-00-5e-00-01-64	2	dynamic
200	00-00-5e-00-01-c8	2	dynamic

The static IP ARP entry has now been set to port 9 instead of port 1. The next thing to understand is that an IP ARP entry will be used by the switch when it does software lookups. The switch consults its IP ARP table when it pings an address and in the first instance sends pings to port 1 only. From this information and our observation in the example above, we conclude that changing the static IP ARP port worked because under ALB one of the NIC ports is not listened to by the NIC driver for some arbitrary rule at the NIC driver level.

ALB should be able to assume that traffic is flooded to all ports and hence the behaviour is not necessarily incorrect.

Data that is switched through the switch (as opposed to being sent from the switch CPU) will be flooded. The port lookup for packets going through the switch is done by consulting the FDB, which does not have an entry for the cluster MAC address.

## Summary

This solution provides NIC redundancy and transmit-aggregation. It will not be tolerant to failures downstream but the cluster should recover quickly if one or all ports fail on a node.

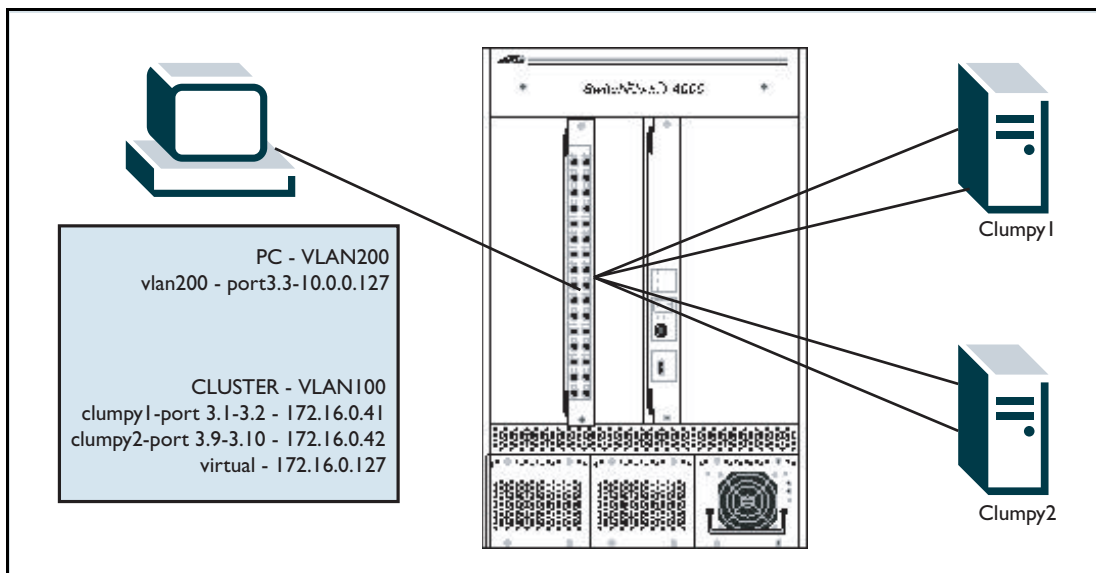
# Link Access/Aggregation Control Protocol and a SwitchBlade

The following configuration demonstrates how to increase your NLB bandwidth by teaming two NICs. We will use Link Aggregation Control Protocol (LACP) to maximise this bandwidth.

- Equipment**
- Intel Pro/1000 MT Dual Port NIC
  - SwitchBlade (AT-SB4108-00)
  - Switch Control Card (AT-SB4211)
  - 24 port trispeed line card (AT-SB4412)
  - 2 x 32-bit PCs running Windows 2003 Enterprise Edition

**Note:** LACP requires a “K revision” line card in the SwitchBlade. Any control card should be okay.

## Network diagram



**Note:** The card in the SwitchBlade in the diagram above shows 32 ports rather than the 24 that our setup actually uses.

## Configuration

First, set up LACP on your SwitchBlade. At the hardware level you will need a “K revision” line card, such as a 24 port trisped card.

Set the Windows 2003 NLB cluster servers to use multicast mode.

Remember to add a static ARP for the multicast MAC, and to use the command **enable ip mscluster**.

```
set system name="Switchblade"

create vlan="vlan100" vid=100
create vlan="vlan200" vid=200

add vlan="100" port=3.1-3.2,3.9-3.10
add vlan="200" port=3.3

enable ip
add ip int=vlan100 ip=172.16.0.254 mask=255.255.255.0
add ip int=vlan200 ip=10.0.0.254 mask=255.255.255.0
enable ip mscluster
add ip arp=172.16.0.127 int=vlan100 eth=03-bf-ac-10-00-7f port=3.1

ena lacp
delete lacp port=3.3-3.8,3.11-3.24
set lacp port=3.1-3.2 adminkey=100
set lacp port=3.9-3.10 adminkey=200
```

Your clustered servers' NICs must support Link Aggregation (802.3ad) mode. This will either be a part of the driver set already or you will need to download the latest drivers from whoever your NIC vendor is. Set up link aggregation on the NICs last.

---

**Note:** It may take a few moments after Windows starts up for NLB to begin.

---

## Description

When your setup is working you should see results similar to the following:

```
Manager Switchblade> sh ip arp
```

Interface	IP Address	Physical Address	ARP Type	Status
vlan200(3.3)	10.0.0.127	00-50-fc-ee-f5-13	Dynamic	active
vlan200	10.0.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan100(3.2)	172.16.0.41	00-04-23-b5-bb-6e	Dynamic	active
vlan100(3.10)	172.16.0.42	00-04-23-ab-37-02	Dynamic	active
vlan100(3.1)	172.16.0.127	<u>03-bf-ac-10-00-7f</u>	Static	active
vlan100	172.16.0.255	ff-ff-ff-ff-ff-ff	Other	active
vlan200	255.255.255.255	ff-ff-ff-ff-ff-ff	Other	active

Manager Switchblade> sh swi fdb inst=3.0

Switch Forwarding Database (hardware)

Blade Switch Instance: 3.0, (AT-SB4412 24-10/100/1000T(RJ45))

-----  
 Total Number of Entries = 8  
 -----

VLAN	MAC Address	Port	Status
1	00-00-cd-05-01-14	CPU	static
100	00-00-cd-05-01-14	CPU	static
200	00-00-cd-05-01-14	CPU	static
100	00-04-23-ab-37-02	3.10	dynamic
100	00-04-23-b5-bb-6e	3.2	dynamic
100	00-04-23-b5-bb-6f	3.2	dynamic
100	00-04-23-ab-37-03	3.10	dynamic
200	00-50-fc-ee-f5-13	3.3	dynamic

Manager Switchblade> sh swi tab=ip inst=3.0

Blade Switch Instance: 3.0, (AT-SB4412 24-10/100/1000T(RJ45))

CXE IP route table summary information:

Total number of entries = 18  
 -----

CXE IP route table:

Index	Seg	Network	NextHop	ToProc	VRP	VID	Hits	PV	Prio	Perm
0	101	192.168.255.1	00-00-cd-05-00-00	0	0	4094	0	0	0	1
1	101	172.16.0.127	03-bf-ac-10-00-7f	0	6	100	0	0	0	1
2	101	172.16.0.42	00-04-23-ab-37-02	0	6	100	0	0	0	1
3	101	10.0.0.127	00-50-fc-ee-f5-13	0	5	200	0	0	0	1
4	101	172.16.0.41	00-04-23-b5-bb-6e	0	6	100	0	0	0	1
5	102	192.168.255.12	00-00-cd-05-00-58	0	0	4094	0	0	0	1
6	102	192.168.255.127	00-00-cd-05-00-58	0	0	4094	0	0	0	1
7	102	192.168.255.140	00-00-cd-05-00-58	0	1	4093	0	0	0	1
8	102	192.168.255.255	00-00-cd-05-00-58	0	1	4093	0	0	0	1
9	102	10.0.0.254	00-00-cd-05-01-14	0	5	200	0	0	0	1
10	102	10.0.0.255	00-00-cd-05-01-14	0	5	200	0	0	0	1
11	102	172.16.0.254	00-00-cd-05-01-14	0	6	100	0	0	0	1
12	102	172.16.0.255	00-00-cd-05-01-14	0	6	100	0	0	0	1
13	109	192.168.255.0	00-00-cd-05-00-58	0	0	4094	0	0	0	1
14	109	192.168.255.128	00-00-cd-05-00-58	0	1	4093	0	0	0	1
15	110	10.0.0.0	00-00-cd-05-01-14	0	5	200	0	0	0	1
16	110	172.16.0.0	00-00-cd-05-01-14	0	6	100	0	0	0	1
17	134	0.0.0.0	00-00-cd-05-00-00	0	0	4094	0	0	0	1

```

Manager Switchblade> sh lacp trunk

LACP Dynamic Trunk Group Information
-----

Trunk group name ..... lacp30:
  Speed ..... 1000 Mbps
  Ports in Trunk ..... 3.1-3.2
  LAG ID:
  [(8000,00-00-cd-05-01-14,0004,00,0000), (ffff,00-04-23-b5-bb-6e,0011,00,0000)]

Trunk group name ..... lacp29:
  Speed ..... 1000 Mbps
  Ports in Trunk ..... 3.9-3.10
  LAG ID:
  [(8000,00-00-cd-05-01-14,0005,00,0000), (ffff,00-04-23-ab-37 02,0011,00,0000)]

```

Notice that the MAC address 03-bf-ac-10-00-7f only appears in the IP ARP table, it never appears in the FDB. Two of the MAC addresses in the FDB appear in the IP ARP table. Another two (00-04-23-b5-bb-6f and 00-04-23-ab-37-03) appear only in the FDB, they are the second ports of each node. Both ports of the node are associated with the same port. This is a designed feature of trunking, where all the MAC addresses in a trunk are associated in FDB with one port in that trunk. This does not break trunking because traffic will still be balanced across all the ports in the trunk.

In this setup, with minimal load, traffic may be seen to only be received on one port in the trunk and reply sent out the other port. To check that the LACP trunk is working, ping the cluster address.

**On port 3.1 you should see only echo requests:**

```

16:45:42.208807 IP (tos 0x0, ttl 63, id 0, offset 0, flags [DF], length: 84) 10
.0.0.127 > 172.16.0.127: icmp 64: echo request seq 57088
16:45:43.208795 IP (tos 0x0, ttl 63, id 0, offset 0, flags [DF], length: 84) 10
.0.0.127 > 172.16.0.127: icmp 64: echo request seq 57344
16:45:44.208777 IP (tos 0x0, ttl 63, id 0, offset 0, flags [DF], length: 84) 10
.0.0.127 > 172.16.0.127: icmp 64: echo request seq 57600

```

**and only echo replies on port 3.2:**

```

16:47:26.208996 IP (tos 0x0, ttl 128, id 6872, offset 0, flags [DF], length:
84) 172.16.0.127 > 10.0.0.127: icmp 64: echo reply seq 57088
16:47:27.208993 IP (tos 0x0, ttl 128, id 6921, offset 0, flags [DF], length:
84) 172.16.0.127 > 10.0.0.127: icmp 64: echo reply seq 57344
16:47:28.208992 IP (tos 0x0, ttl 128, id 6922, offset 0, flags [DF], length:
84) 172.16.0.127 > 10.0.0.127: icmp 64: echo reply seq 57600

```

The important point is that the ICMP echo request is sent to one port only in the trunk. The NLB server will decide which port to send the reply over (either 3.1 or 3.2); in this case it was over port 3.2. Note that both servers may respond to the ICMP packet depending on the NLB rules so your client could report duplicate replies.

You may see the same behaviour with HTTP requests (one port egress, one port ingress) but as traffic increases it is more likely that the SwitchBlade and Windows will balance traffic



across both ports. The key is that you should never see duplicate unicast packets over two ports in the same trunk.

Port 3.1 would appear as such:

```
16:20:13.856129 IP 10.0.0.127.42235 > 172.16.0.127.http: SWE
3617082630:3617082630(0) win 5840 <mss 1460,sackOK,timestamp 235508243
0,nop,wscale 0>
16:20:13.857435 IP 10.0.0.127.42235 > 172.16.0.127.http: . ack 4292649670 win
5840 <nop,nop,timestamp 235508244 0>
16:20:13.858012 IP 10.0.0.127.42235 > 172.16.0.127.http: P 0:273(273) ack 1
win 5840 <nop,nop,timestamp 235508244 0>
16:20:13.860174 IP 10.0.0.127.42235 > 172.16.0.127.http: F 273:273(0) ack 397
win 6432 <nop,nop,timestamp 235508245 23725>
```

Port 3.2 may appear as such:

```
17:10:27.509249 IP (tos 0x0, ttl 128, id 11982, offset 0, flags [none],
length:64) 172.16.0.127.http > 10.0.0.127.42628: S [tcp sum ok]
1493187854:1493187854(0) ack 2433262544 win 16384 <mss 1460,nop,wscale
0,nop,nop,timestamp 0 0,nop,nop,sackOK>
17:10:27.510746 IP (tos 0x0, ttl 128, id 11983, offset 0, flags [DF], length:
447) 172.16.0.127.http > 10.0.0.127.42628: FP 1:396(395) ack 274 win 65262
<nop,nop,timestamp 22780 235809610>
17:10:27.511880 IP (tos 0x0, ttl 128, id 11984, offset 0, flags [DF], length:
52) 172.16.0.127.http > 10.0.0.127.42628: . [tcp sum ok] 397:397(0) ack 275
win 65262 <nop,nop,timestamp 22780 235809610>
```

Please see the section "[Special Note About LACP and 802.3ad](#)" on page 36.

## Summary

This method will provide efficient fault tolerance and load balancing. This setup will not give an individual node redundancy if a line card is hot swapped. However, better use of bandwidth is gained. Cluster redundancy is preserved if the two nodes each connect to two 24-port trispeed cards, because hot swapping one line card will leave the other node(s) up.

LACP will give you management capabilities over link aggregation. The benefits of that may only be slight. For more information about LACP on Allied Telesis switches, see your switch's Software Reference. For more information on just LACP, consult the IEEE standard 802.3-2002.

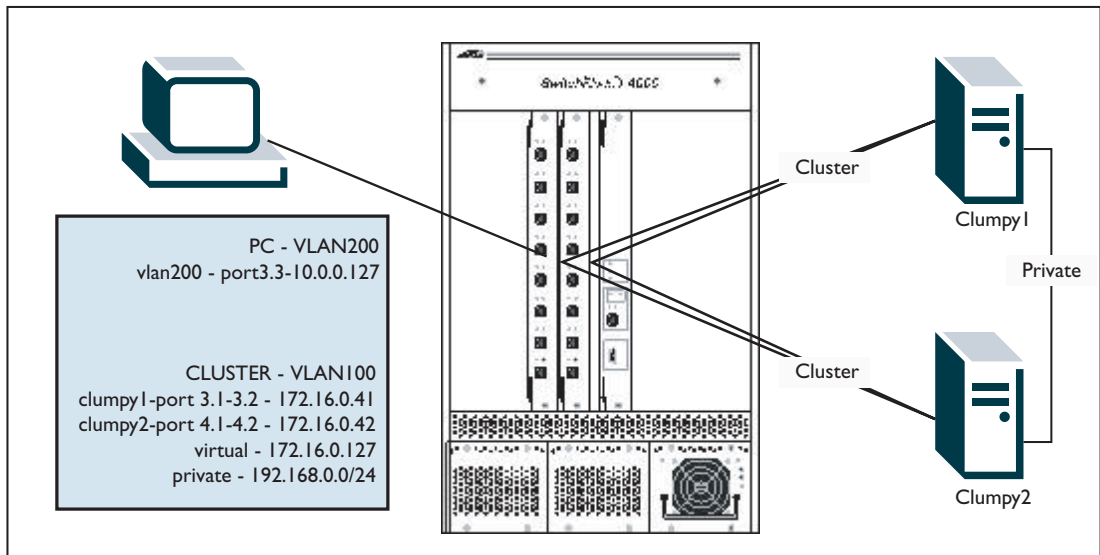
The alternative is Static Link Aggregation, which does not rely on special hardware and software to run.

# Static Link Aggregation with a SwitchBlade

Static link aggregation was developed to increase throughput between switches. Because Allied Telesis switches support 802.3ad, this is a viable option.

- Equipment**
- Intel Pro/1000 MT Dual Port NIC
  - Linux PC as the workstation
  - 2 x 32 bit PCs with Windows 2003 Enterprise Edition
  - SwitchBlade chassis (AT-SB4108-00)
  - Switch Control Card (AT-SB4211)
  - 2 x 8 port UTP trispeed line card (AT-SB4411 8-1000T(RJ45))

**Network diagram**



## Configuration

```
#
# SYSTEM configuration
#
set system name="Switchblade"
#
# VLAN general configuration
#
create vlan="vlan100" vid=100
add vlan="100" port=3.1-3.2,4.1-4.2
create vlan="vlan200" vid=200
add vlan="200" port=3.3
#
# IP configuration
#
enable ip
add ip int=vlan100 ip=172.16.0.254 mask=255.255.255.0
add ip int=vlan200 ip=10.0.0.254 mask=255.255.255.0
enable ip mscluster
add ip arp=172.16.0.127 int=vlan100 eth=03-bf-ac-10-00-7f port=3.1
#
# SWITCH configuration
#
create switch trunk=clumpy2 port=4.1-4.2 speed=1000m
create switch trunk=clumpy1 port=3.1-3.2 speed=1000m
```

## Description

802.3ad trunking demands that both ports in a trunk be on the same switch instance. Therefore, you cannot connect two ports on the same node to two different cards and expect to be able to trunk them. By putting each node on its own line card, only cluster redundancy is gained. NLB multicast mode must be used in the cluster.

Please see the section ["Special Note About LACP and 802.3ad"](#) on page 36.

## Summary

This method will provide efficient fault tolerance and load balancing with the added bonus of working on any Allied Telesis SwitchBlade line card. This setup will not give an individual node redundancy if a line card is hot swapped. However, better use of bandwidth is gained. Cluster redundancy is preserved because hot swapping one line card will leave the other node(s) up.

## Special Note About LACP and 802.3ad

The SwitchBlade will send IP packets destined for the virtual address to one port in each trunk. The node's NIC driver will decide to send return traffic via one NIC, which is a decision the NIC driver makes based on its own algorithm. An IP packet may be sent by the SwitchBlade to port 1 of the node and the reply sent to the SwitchBlade from port 2 on the node. An IP packet may be sent by the SwitchBlade to port 2 of the node and the reply sent to the SwitchBlade again from port 2 on the node.

Therefore, all traffic cannot be seen on one port alone but sniffing both ports should reveal the entire traffic flow with no duplicate packets.

The port that the SwitchBlade sends traffic to is dependant on the layer(s) that you choose to involve in the hashing calculation that determines the egress port for a given packet. Layer 2 and layer 3 hashing is normally on. Layer 4 hashing may or may not be on by default, depending on the SwitchBlade software release and control cards. Under layer 4 hashing, the SwitchBlade will use TCP port numbers to determine which physical port in a trunk to send traffic over. Put simply, layer 4 hashing will give a randomly balanced traffic shape for packets with the same source and destination IP.

The use of layer 2 and/or layer 3 hashing without layer 4 will cause the SwitchBlade to send all IP traffic to the same port on the node every time for packets with the same source and destination IP. The node will continue to send return traffic based on its own algorithm.

In the following example output from the command **show switch**, you can see that the SwitchBlade control card is using layer 2 and layer 3 hashing.

```

Switch Configuration
-----
Silicon Revision ..... CXE K1
Switch Type ..... Control Card
Blade Position ..... 9
Blade Switch Instance ..... 9.0
Blade Instance Health ..... Alive
IP Address ..... 192.168.255.0
IPX Address ..... fffffff0
PCI errors ..... 0
CAM size ( entries ) ..... 106496
Switch Address ..... 00-00-cd-05-01-14
Learning ..... ON
Ageing Timer ..... ON
Number of Fixed Ports ..... 0
Number of Uplink Ports ..... 0
Mirroring ..... DISABLED
Mirror port ..... None
Ports mirroring on Rx ..... None
Ports mirroring on Tx ..... None
Ports mirroring on Both .... None
Pause Flow Control Mode .... Basic
Basic Flow Control Length .. 10
Number of WAN Interfaces ... 0
Name of Interface(s) ..... -
Ageingtime ..... 300
VLAN classification ..... To be defined
STP Forwarding ..... Disabled
UpTime ..... 04:09:55
Hashingfield ..... L2 L3
L3 Test mode ..... MODE1
-----

Switch Configuration
-----
Silicon Revision ..... CXE K1
Switch Type ..... Control Card
Blade Position ..... 9
Blade Switch Instance ..... 9.1
Blade Instance Health ..... Alive
IP Address ..... 192.168.255.0
IPX Address ..... fffffff0
PCI errors ..... 0
CAM size ( entries ) ..... 106496
Switch Address ..... 00-00-cd-05-01-14
Learning ..... ON
Ageing Timer ..... ON
Number of Fixed Ports ..... 0
Number of Uplink Ports ..... 0
Mirroring ..... DISABLED
Mirror port ..... None
Ports mirroring on Rx ..... None
Ports mirroring on Tx ..... None
Ports mirroring on Both .... None
Pause Flow Control Mode .... Basic
Basic Flow Control Length .. 10
Number of WAN Interfaces ... 0
Name of Interface(s) ..... -
Ageingtime ..... 300
VLAN classification ..... To be defined
STP Forwarding ..... Disabled
UpTime ..... 04:09:55
Hashingfield ..... L2 L3
L3 Test mode ..... MODE1

```

To force the SwitchBlade to mix traffic between ports on the same node, turn on layer 4 hashing. Also, to avoid problems with cards that have pre K-revision silicon, consider turning off layer 3 hashing on all control and line cards:

```
enable switch mirror
create switch trunk=clumpy2 port=4.1-4.2 speed=1000m
create switch trunk=clumpy1 port=3.1-3.2 speed=1000m
set switch port=3.1 mirror=both

disable switch hash=13 instance=9.0
disable switch hash=13 instance=9.1
disable switch hash=13 inst=3.0
disable switch hash=13 inst=4.0

ena swi hash=14 inst=3.0
ena swi hash=14 inst=4.0
ena swi hash=14 inst=9.0
ena swi hash=14 inst=9.1
```

This results in the example output from the command **show switch** as shown on the next page.

```

Switch Configuration
-----
Silicon Revision ..... CXE K1
Switch Type ..... Control Card
Blade Position ..... 9
Blade Switch Instance ..... 9.0
Blade Instance Health ..... Alive
IP Address ..... 192.168.255.0
IPX Address ..... fffffff0
PCI errors ..... 0
CAM size ( entries ) ..... 106496
Switch Address ..... 00-00-cd-05-01-14
Learning ..... ON
Ageing Timer ..... ON
Number of Fixed Ports ..... 0
Number of Uplink Ports ..... 0
Mirroring ..... DISABLED
Mirror port ..... None
Ports mirroring on Rx ..... None
Ports mirroring on Tx ..... None
Ports mirroring on Both .... None
Pause Flow Control Mode .... Basic
Basic Flow Control Length .. 10
Number of WAN Interfaces ... 0
Name of Interface(s) ..... -
Ageingtime ..... 300
VLAN classification ..... To be defined
STP Forwarding ..... Disabled
UpTime ..... 04:33:37
Hashingfield ..... L2 L4
L3 Test mode ..... MODE1
-----

Switch Configuration
-----
Silicon Revision ..... CXE K1
Switch Type ..... Control Card
Blade Position ..... 9
Blade Switch Instance ..... 9.1
Blade Instance Health ..... Alive
IP Address ..... 192.168.255.0
IPX Address ..... fffffff0
PCI errors ..... 0
CAM size ( entries ) ..... 106496
Switch Address ..... 00-00-cd-05-01-14
Learning ..... ON
Ageing Timer ..... ON
Number of Fixed Ports ..... 0
Number of Uplink Ports ..... 0
Mirroring ..... DISABLED
Mirror port ..... None
Ports mirroring on Rx ..... None
Ports mirroring on Tx ..... None
Ports mirroring on Both .... None
Pause Flow Control Mode .... Basic
Basic Flow Control Length .. 10
Number of WAN Interfaces ... 0
Name of Interface(s) ..... -
Ageingtime ..... 300
VLAN classification ..... To be defined
STP Forwarding ..... Disabled
UpTime ..... 04:33:37
Hashingfield ..... L2 L4
L3 Test mode ..... MODE1

```

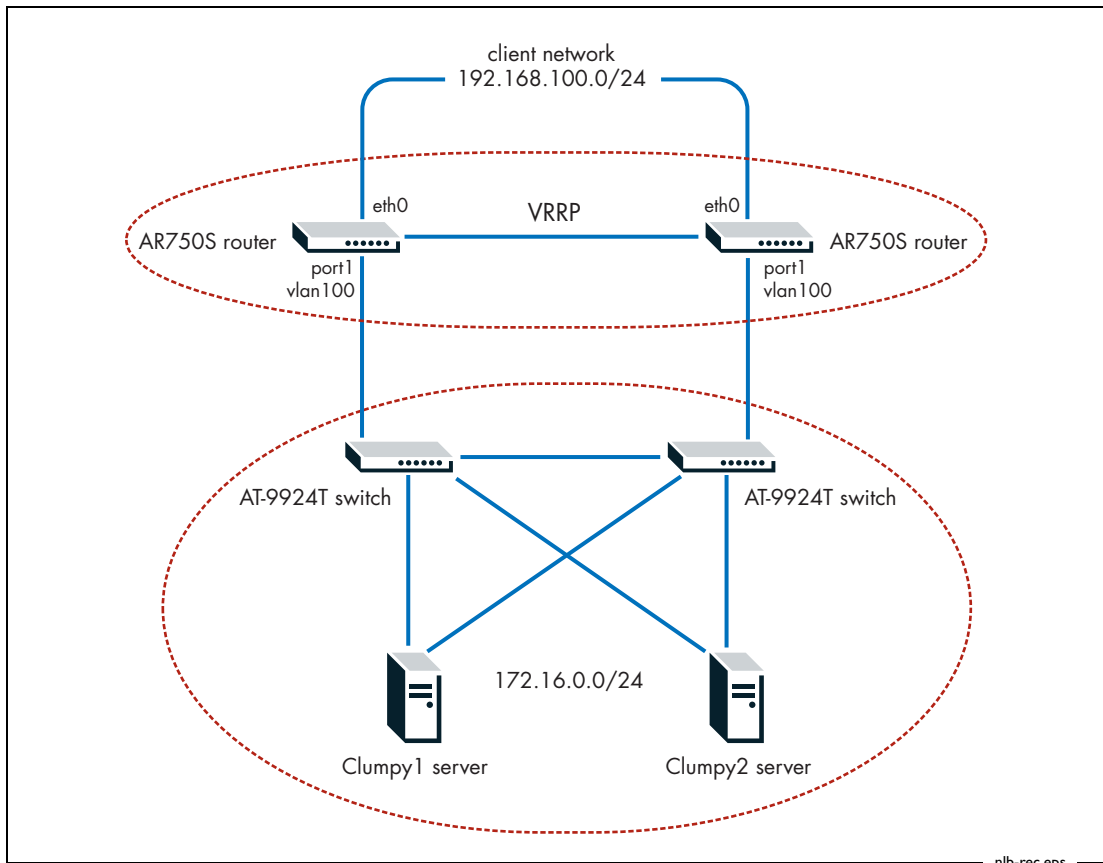
## How Allied Telesis recommends NLB be implemented at layer 3

If you are interested in teamed NICs for redundancy, aggregation or otherwise then the mode you select is entirely your decision. This example will use the most commonly implemented teaming - Switch Fault Tolerance (SFT).

The following method provides the best level of redundancy and failover at layer 2 and layer 3.

At layer 3, it uses AR750S routers, but other routers will also work. Layer 3 switches will also work, as long as each layer 3 switch only has one port connected to one layer 2 switch. The layer 2 switches can be any Allied Telesis switch.

### Network diagram



nlb-rec.eps



## Configuration

The AR750S is in charge of routing traffic between the rest of the network and the L2 network that the servers are on. This network can be easily scaled up with more switches and servers, or scaled down to one switch and two servers each with a single or dual NIC.

Add an IP interface and route information to both of the AR750S routers.

```
enable ip
add ip interface=eth0 ip=192.168.100.x mask=255.255.255.0
add ip interface=vlan100 ip=172.16.0.x mask=255.255.255.0
add ip arp=172.16.0.127 int=vlan100 eth=01-00-5e-7f-00-7f port=1
```

In the above configuration, we have added a static ARP entry to the VLAN interface to which the server's layer 2 network connects. For more recent software versions, you need to also use the command **enable ip mcluster**. For the most recent software versions, you need to use the command **enable ip macdisparity**, but do not need the static ARP entry. For details, see ["Accepting ARP entries that have a multicast MAC address and a unicast IP address" on page 6](#). If you need a static ARP entry, you can see the specific multicast address to use in the NLB Manager (see ["Set up the NLB Cluster" on page 66](#)).

It is important to note that only one switch connects to the AR750S. This will not work properly if a switch is connected to more than one router.

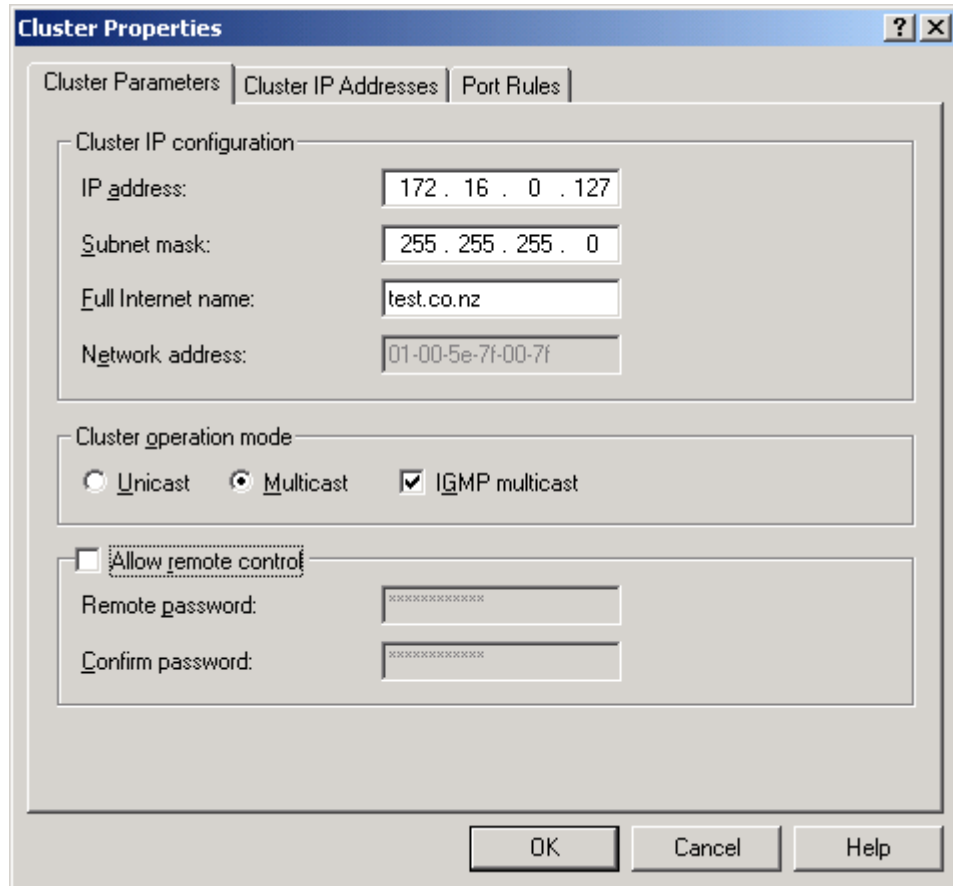
Then create a VRRP interface to be the next hop for the 192.168.100.0/24 network to reach the NLB server network. You should also add a VRRP interface for the 172.16.0.0/24 subnet. The following configuration is applied to both AR routers and includes a monitored interface so that the VRRP master will change if the status of vlan100 or eth0 changes.

```
enable vrrp
create vrrp=1 over=eth0 ipaddress=192.168.100.254
add vrrp=1 monitoredinterface=vlan100
create vrrp=2 over=vlan100 ipaddress=172.16.0.254
add vrrp=2 monitoredinterface=eth0
```

The switches can be configured with or without an IP address but they are not configured with a static IP ARP. The switches should not be configured with multiple VLANs. At the most basic level, the switches do not need any configuration at all. A switch that supports IGMP snooping is all that is actually required. IGMP snooping is enabled by default on Allied Telesis switches. An IGMP querier configuration is required on at least one switch, ideally on both for redundancy:

```
enable ip igmp
ena ip igmp interface=vlan100
```

Configure the NLB servers as normal, but select “IGMP multicast”, as shown in the following figure.



## Appendix 1 —Quick Start Guide to setting up a Windows 2003 NLB cluster

---

### ► Install Windows 2003

Defaults are okay during the install but you must ensure you use NTFS. If you already have Windows 2003 installed, but with the FAT file system, then open up a command prompt (otherwise known as a MSDOS window) and type “convert c: /FS:NTFS /X” and then reboot. When the PC reboots it will analyse the hard drive and then convert the file system to NTFS without destroying data (hopefully).

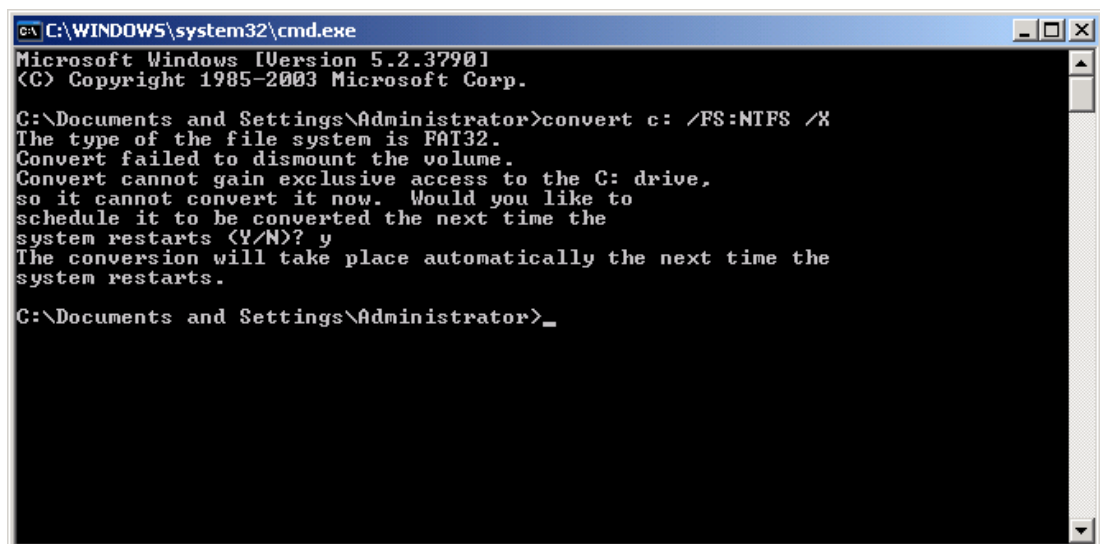
Before you begin. Set up the private node-to-node ethernet connection and assign each node an IP. This example uses the 192.168.0.0/24 network for the private LAN.

Copy the i386 directory to your hard drive if you have space.

## Installation of first NLB Server - clumpy1

### ► Check that your drive partition uses an NTFS file system

If your Windows 2003 installation partition file system is not NTFS, then convert the drive now. You cannot proceed without NTFS.



```
c:\WINDOWS\system32\cmd.exe
Microsoft Windows [Version 5.2.3790]
(C) Copyright 1985-2003 Microsoft Corp.

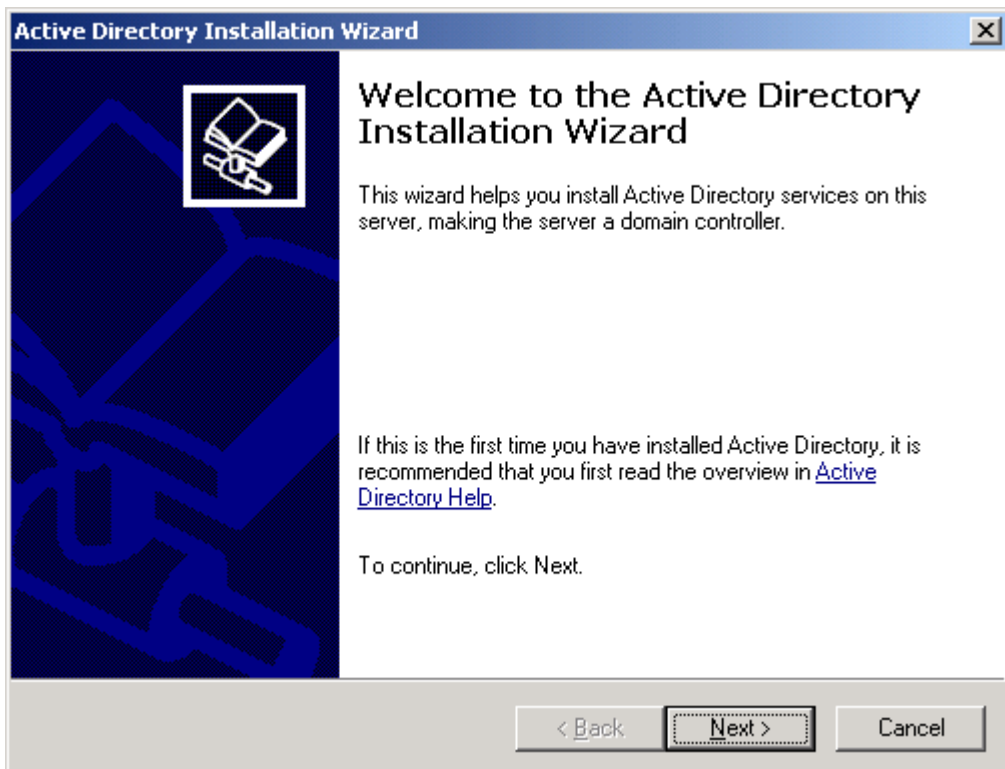
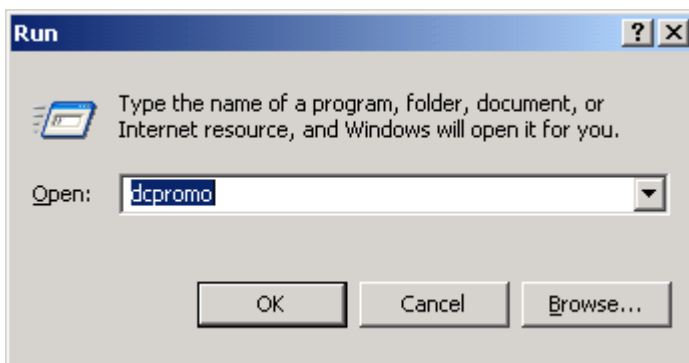
C:\Documents and Settings\Administrator>convert c: /FS:NTFS /X
The type of the file system is FAT32.
Convert failed to dismount the volume.
Convert cannot gain exclusive access to the C: drive,
so it cannot convert it now. Would you like to
schedule it to be converted the next time the
system restarts (Y/N)? y
The conversion will take place automatically the next time the
system restarts.

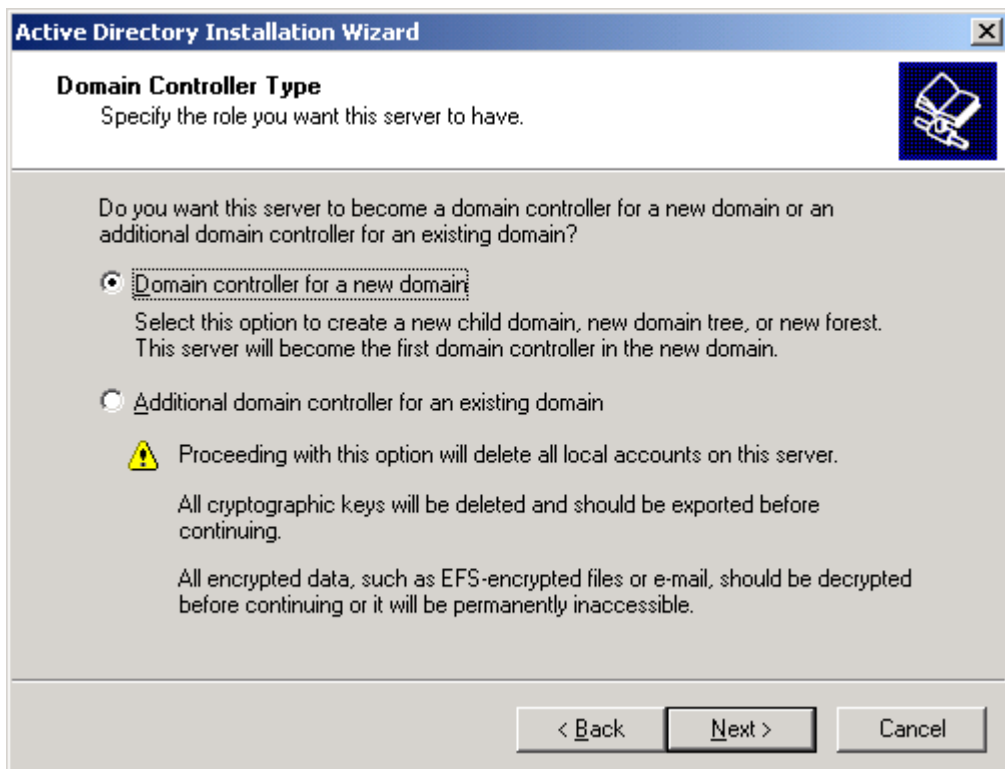
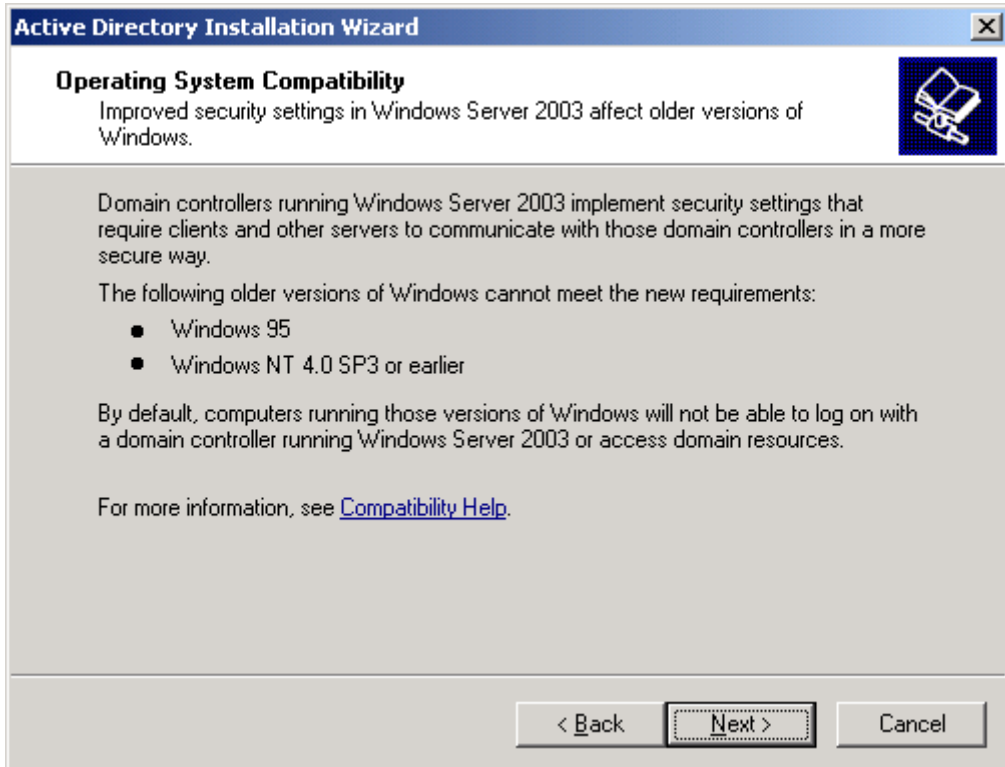
C:\Documents and Settings\Administrator>_
```

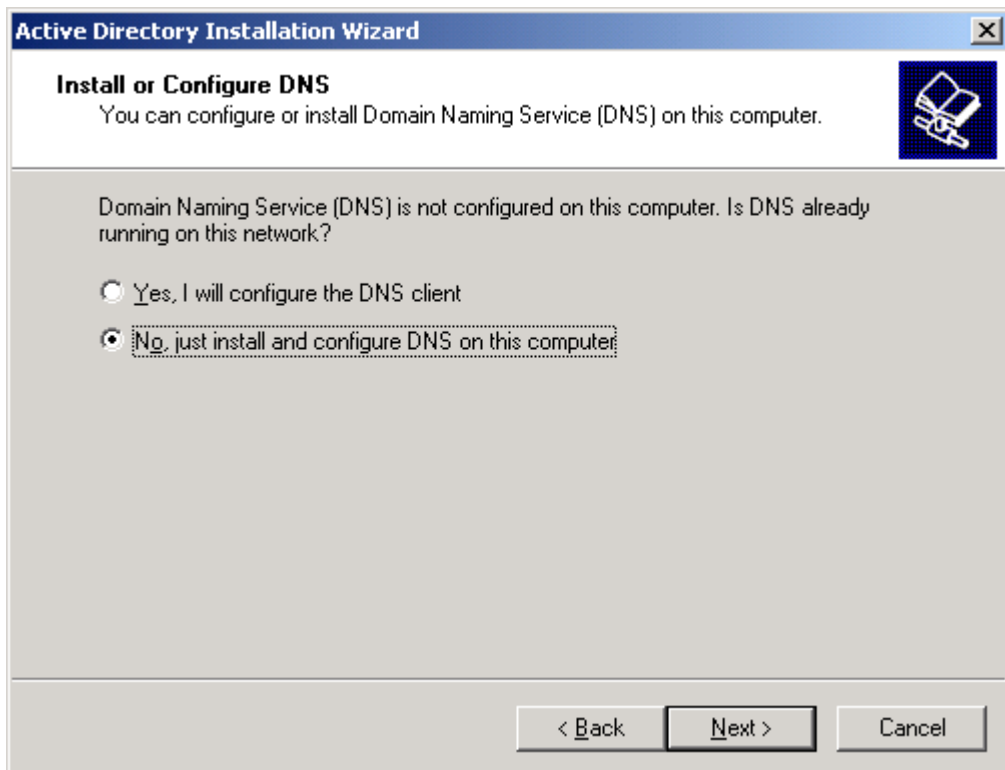
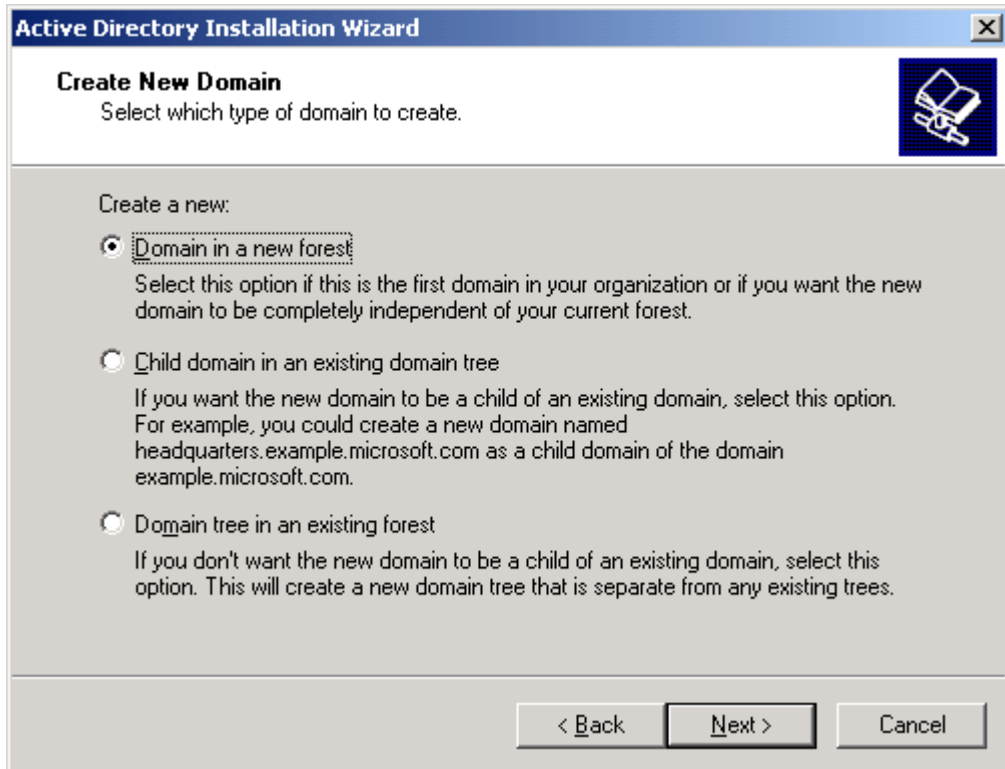
► Install a domain controller (DC) if you do not have one

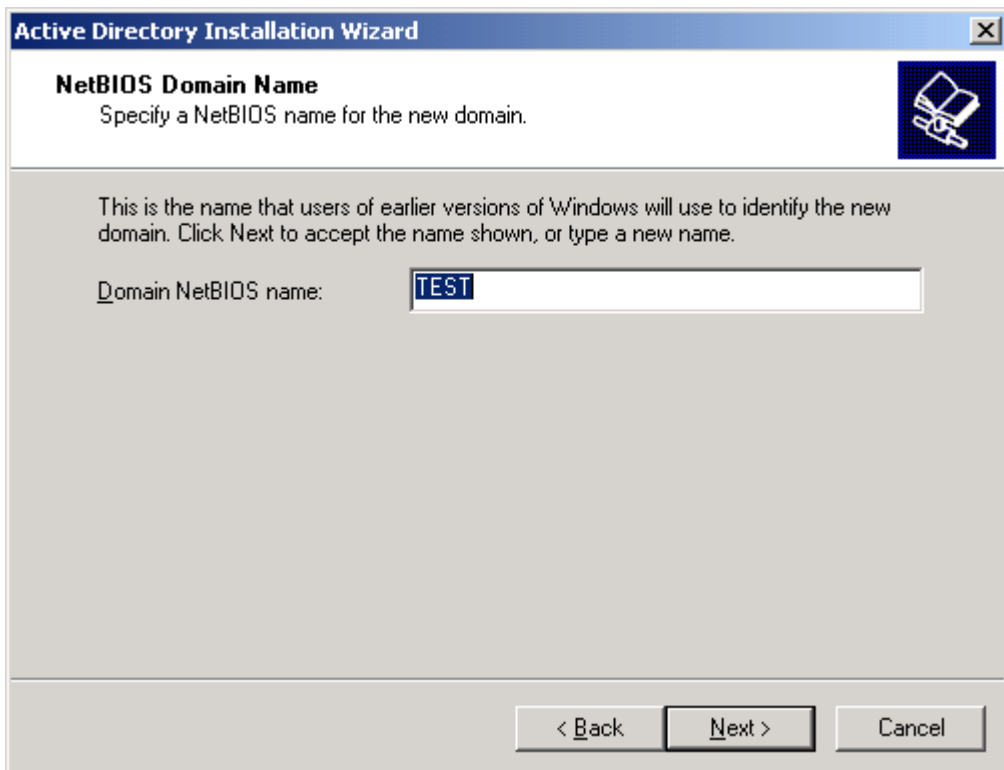
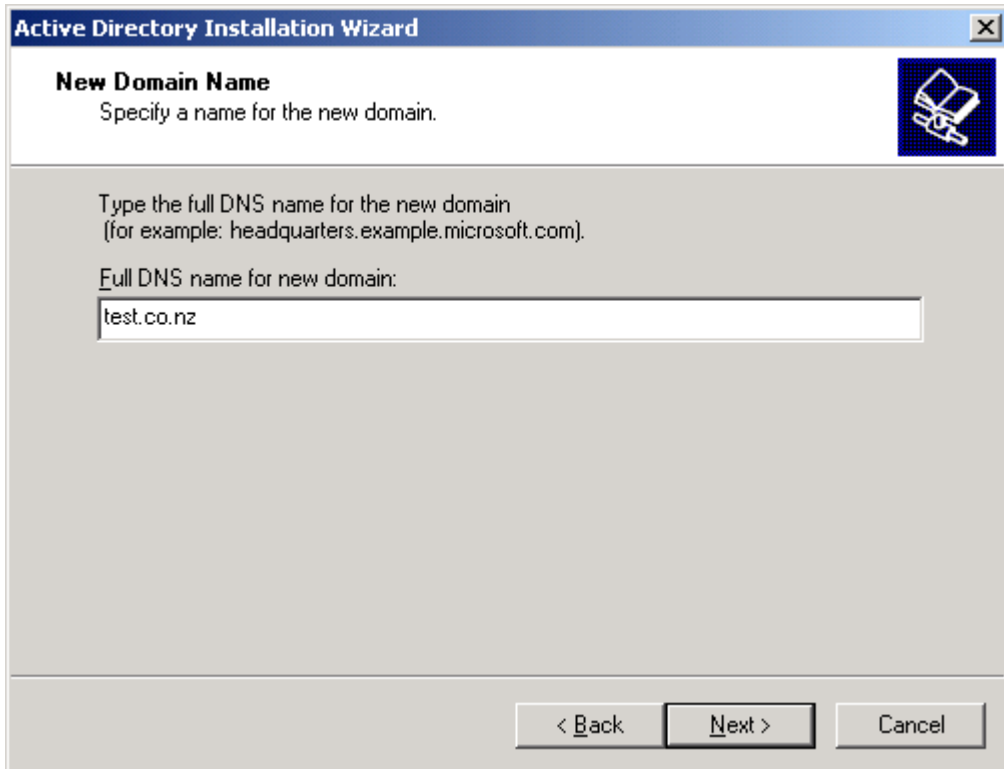
Windows 2003 clustering requires DNS lookups to work and in the case of Clustering requires an account within the domain to be set up. It may be possible to avoid setting up a DC (you could manually set up DNS perhaps) but it is easier to have a domain controller in a Windows environment. Run dcpromo, follow the prompts as per the screenshots and then reboot afterwards.

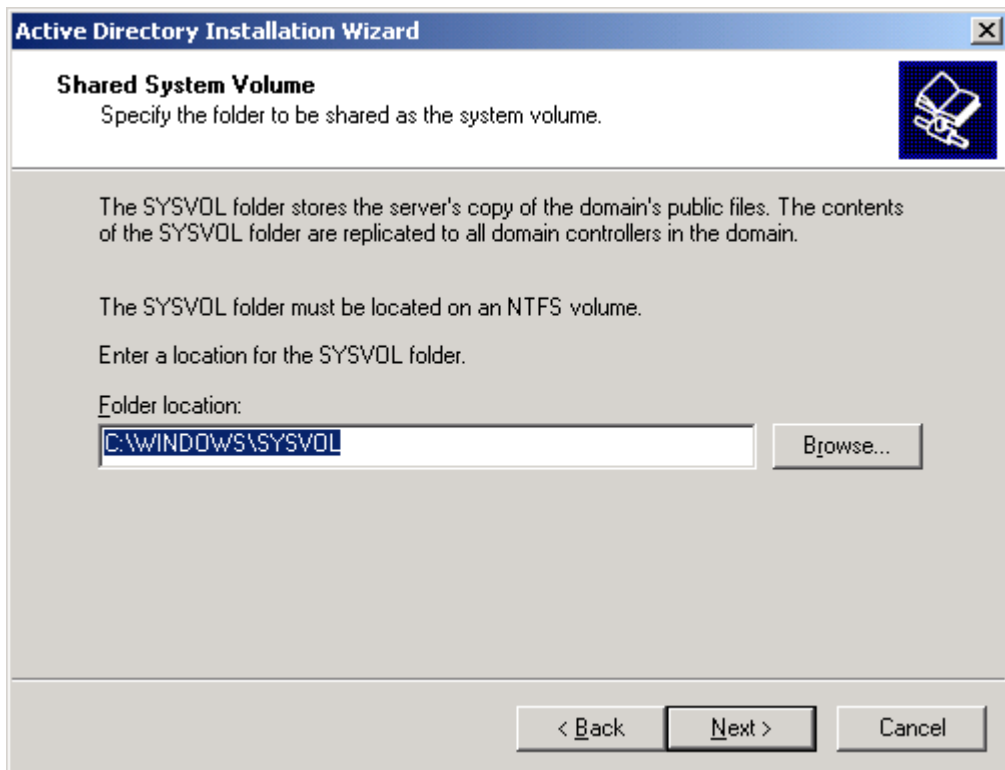
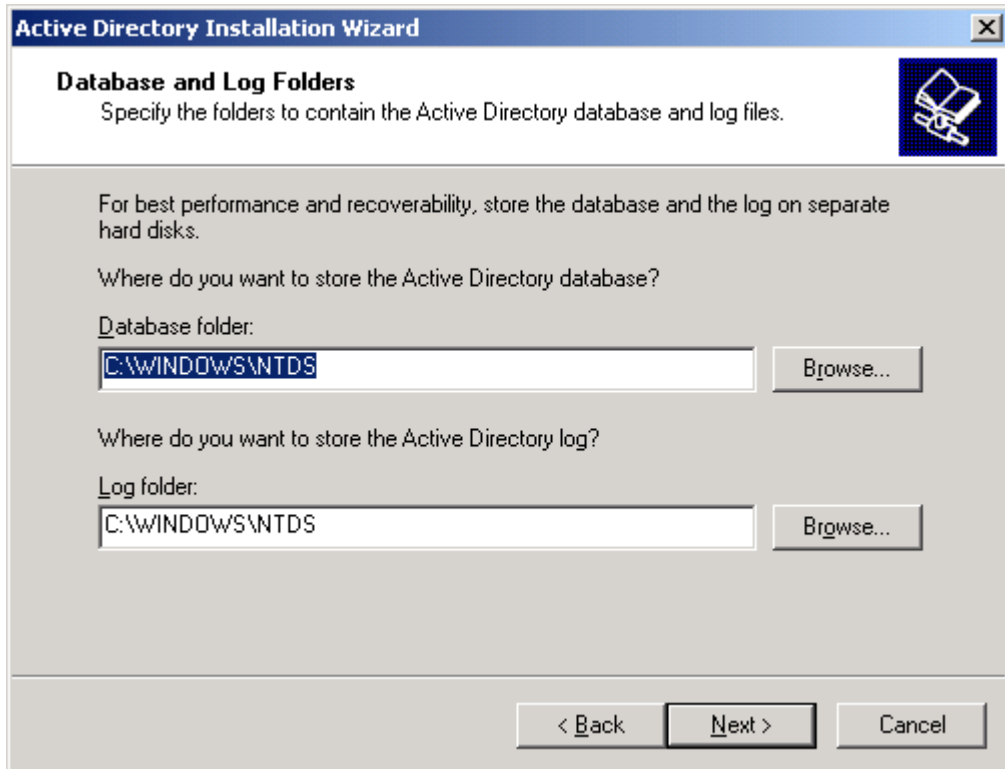
► Set up a Domain Controller



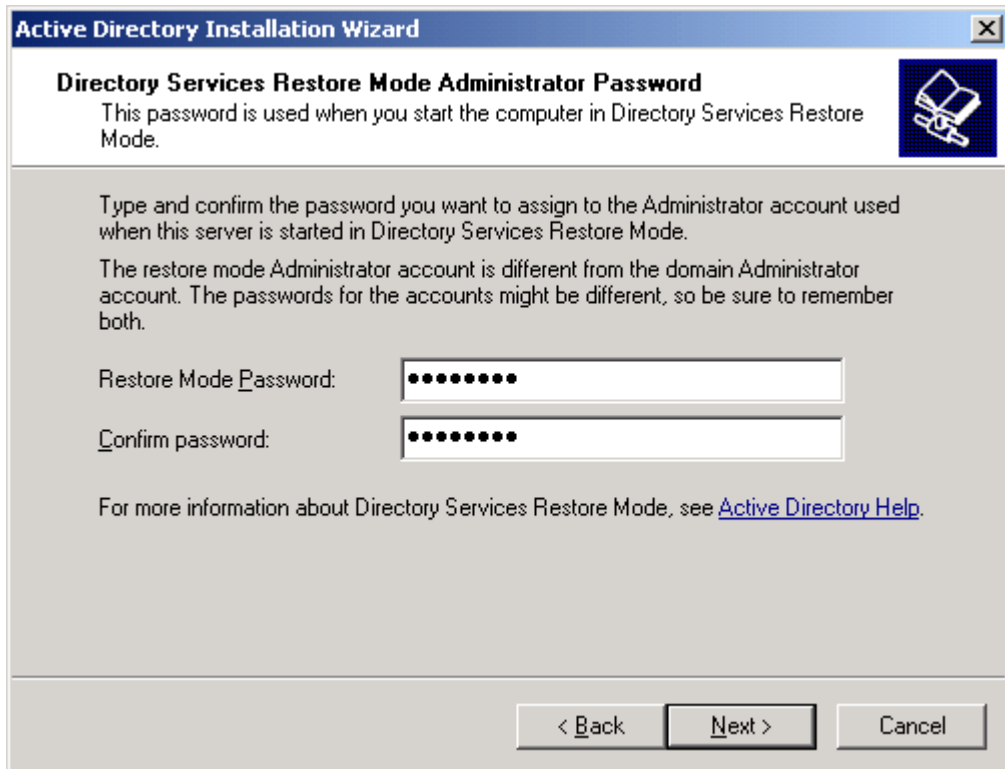
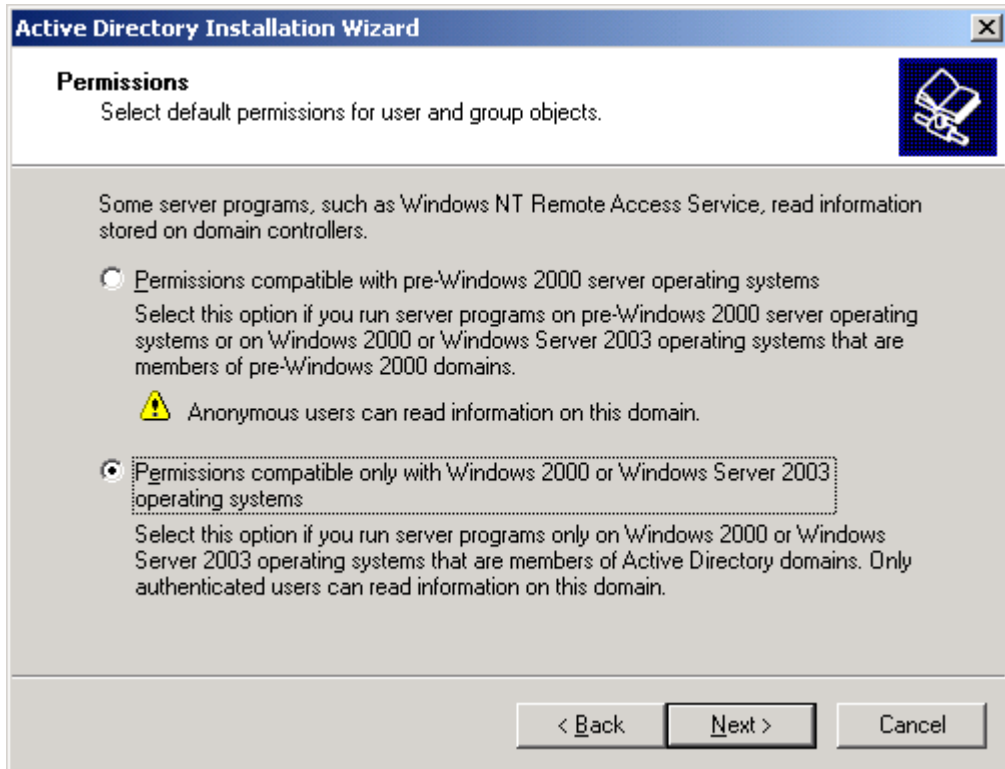


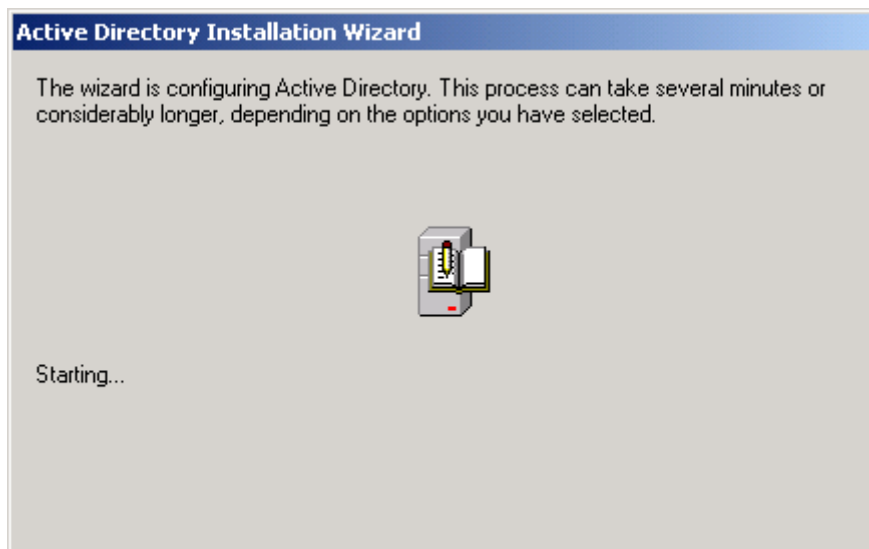
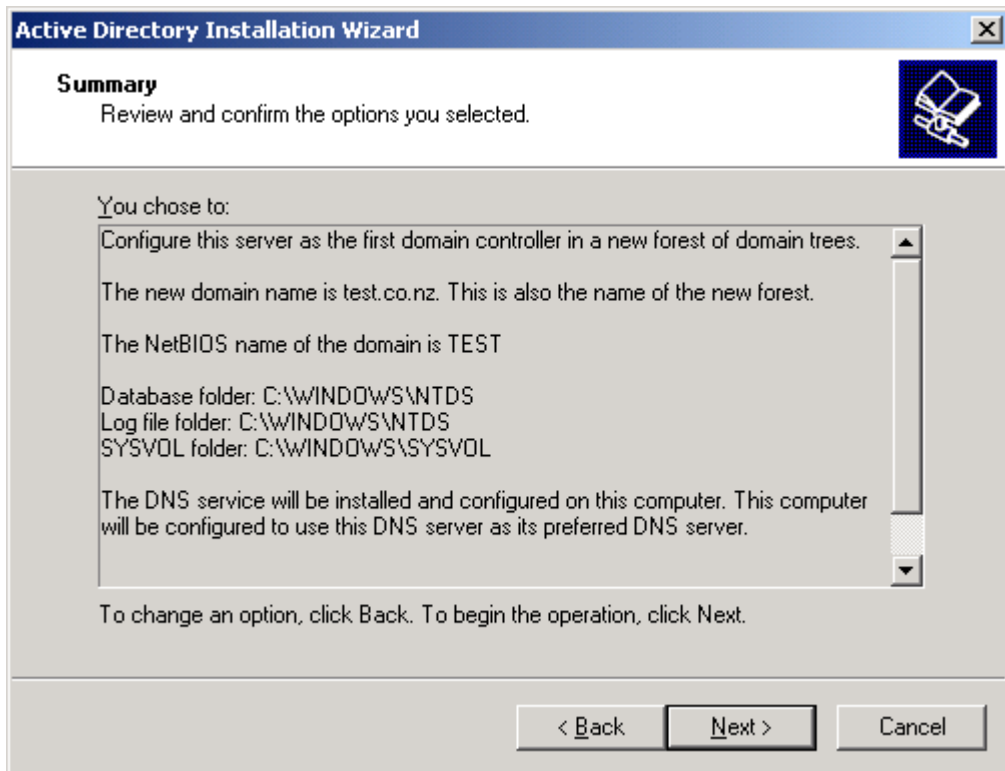


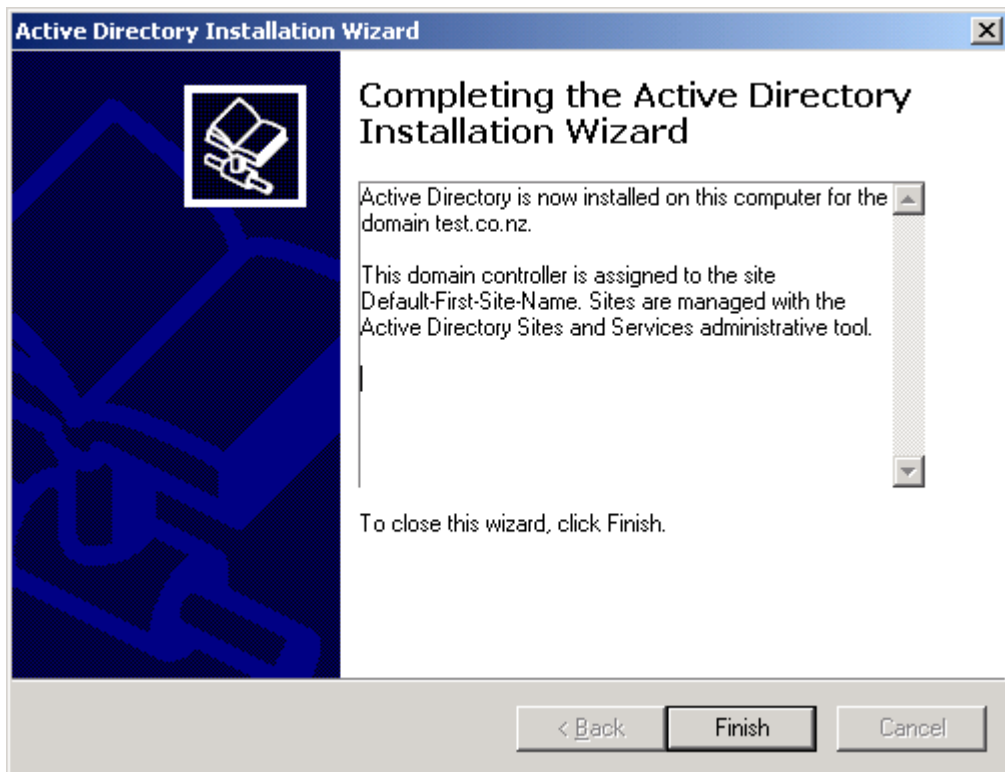
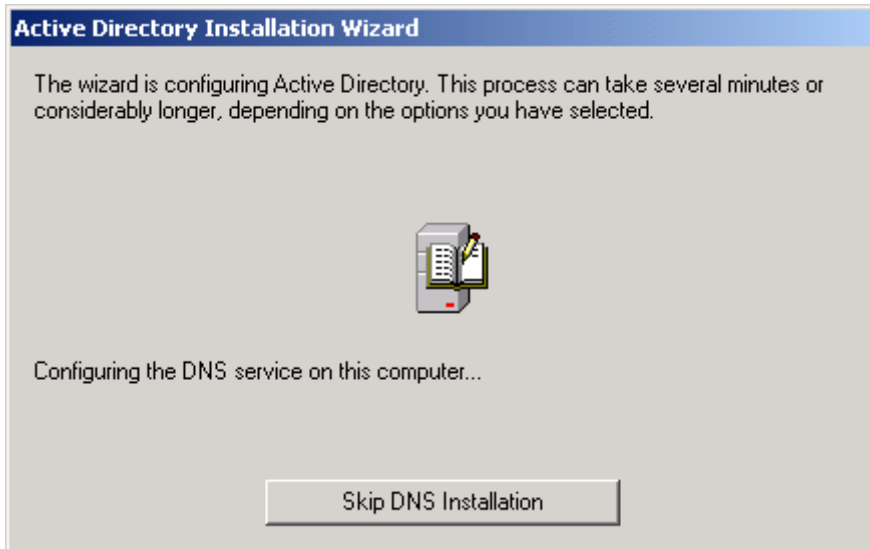




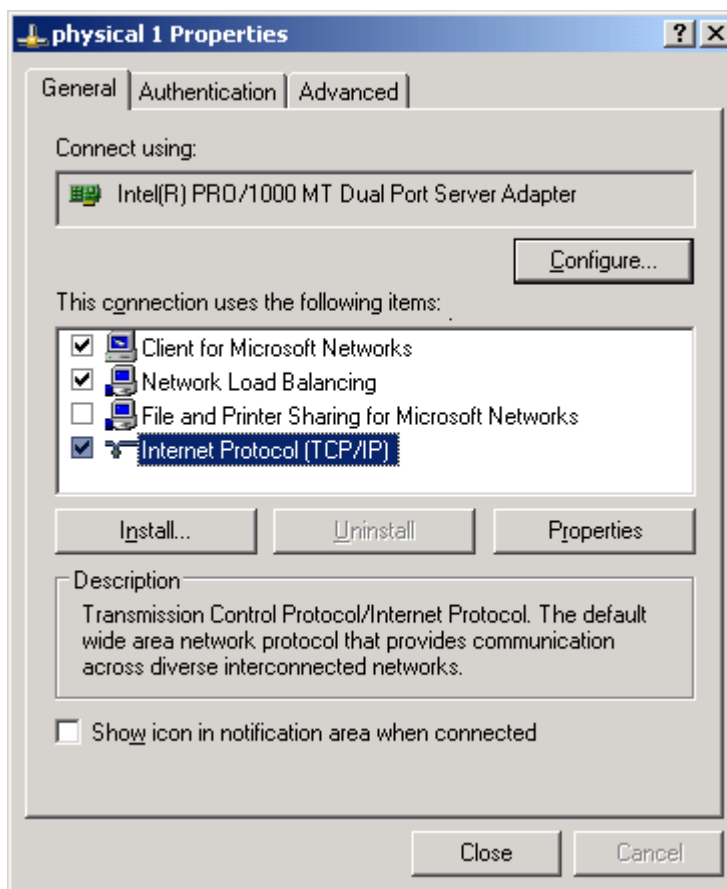
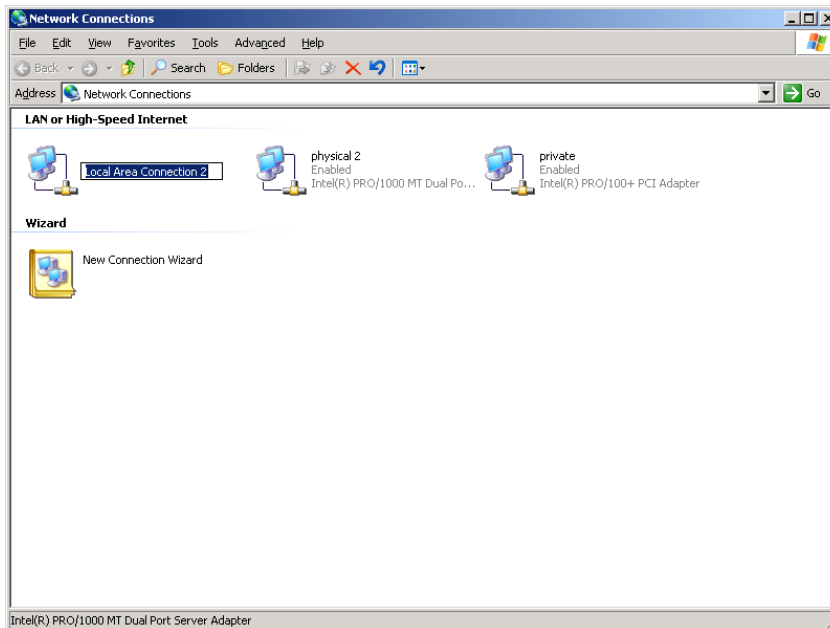




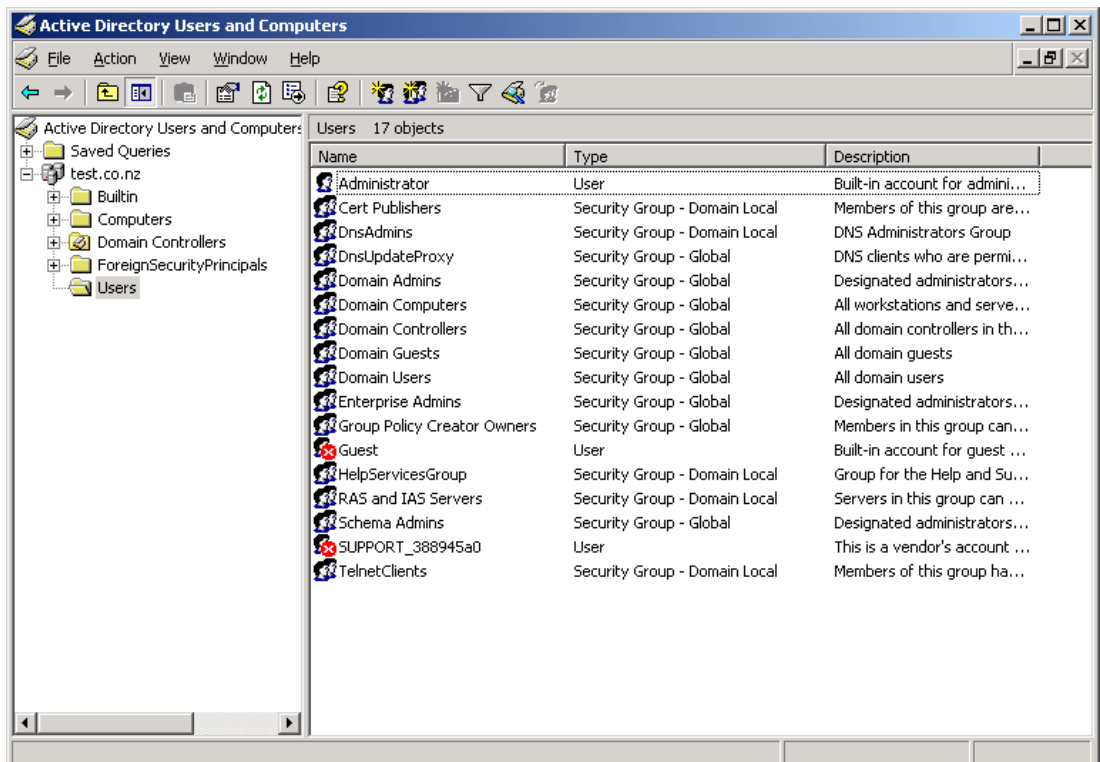
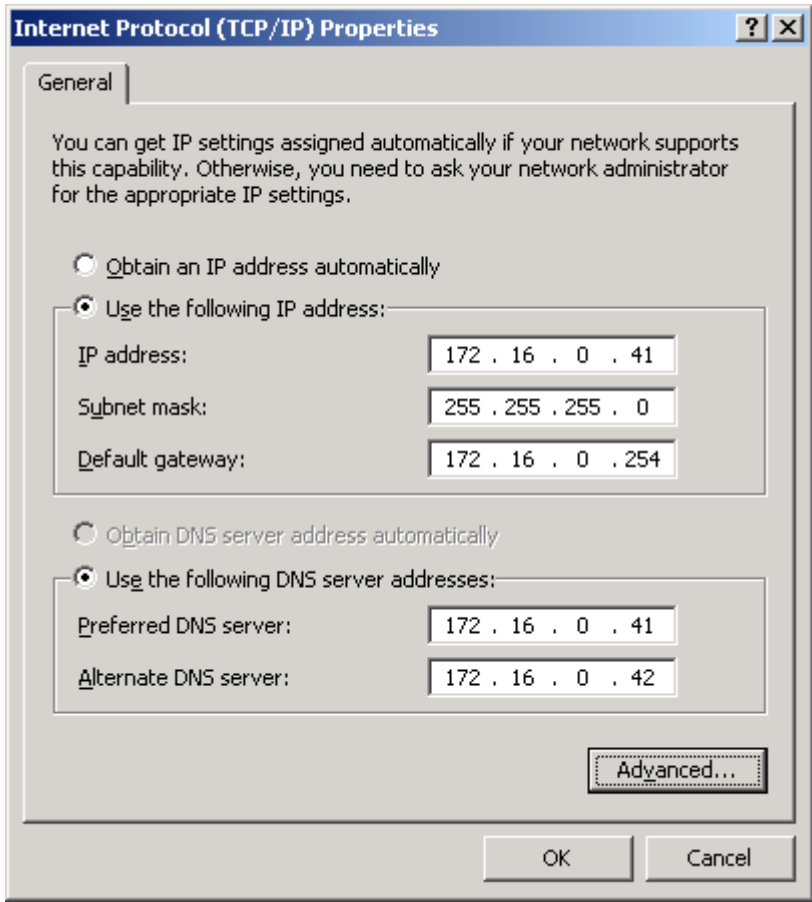




► Configure interfaces



**Note:** In the picture above the Network Load Balancing protocol appears in the list. Do not configure this here but make sure it is present and checked.



**New Object - User**

Create in: test.co.nz/Users

First name:  Initials:

Last name:

Full name:

User logon name:  @test.co.nz

User logon name (pre-Windows 2000):

< Back Next > Cancel

**New Object - User**

Create in: test.co.nz/Users

Password:

Confirm password:

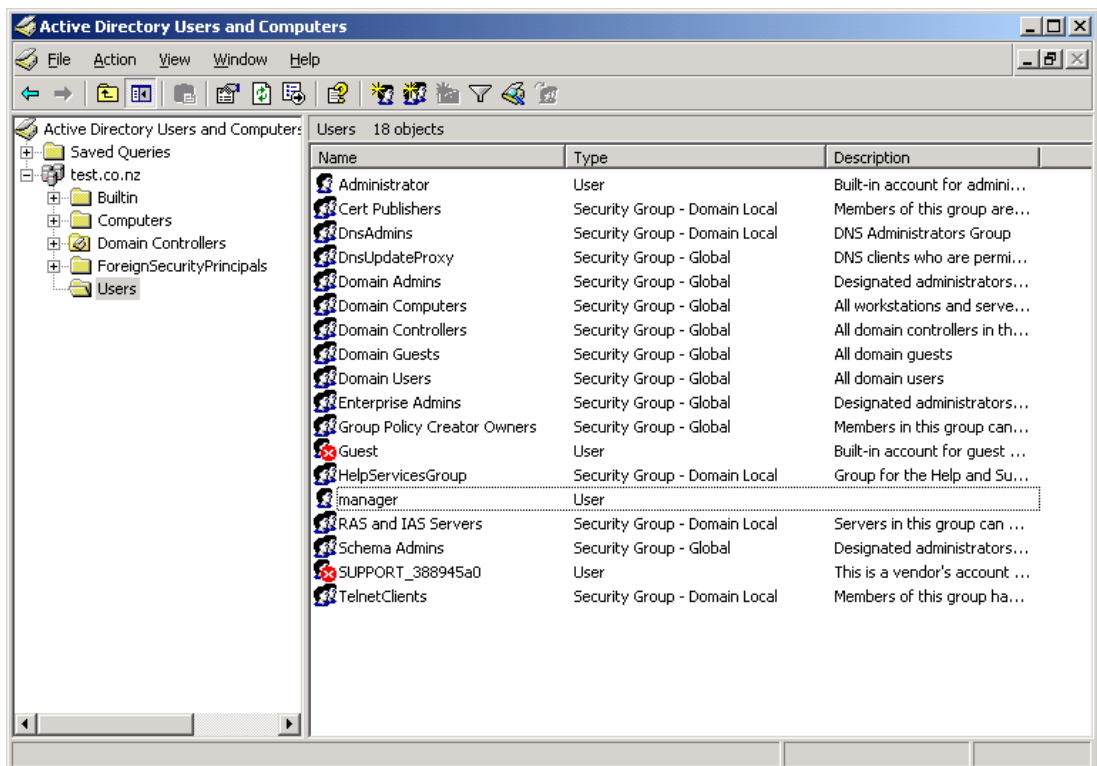
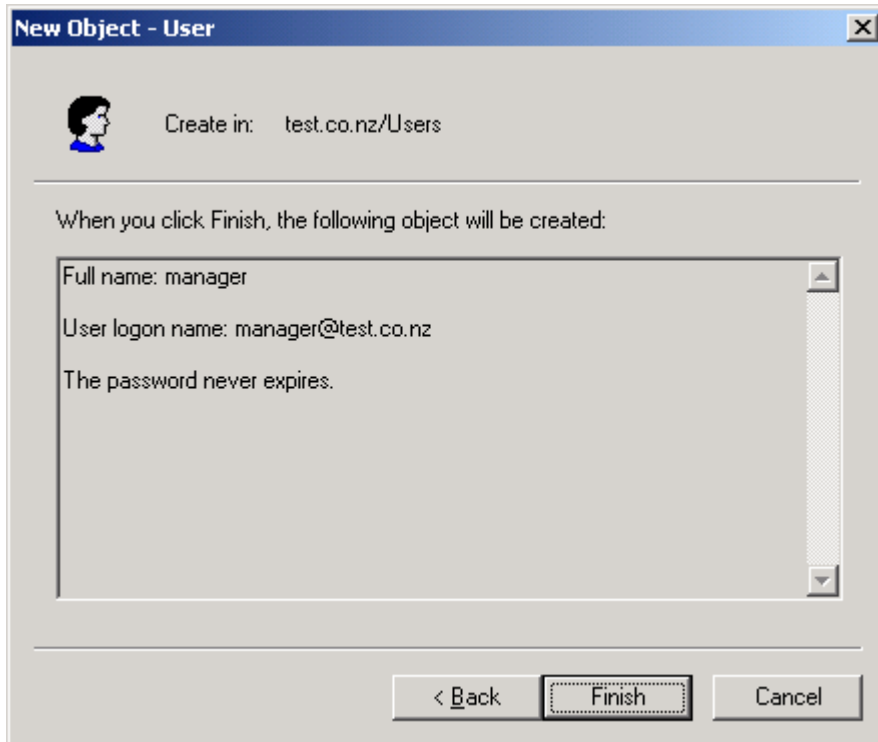
User must change password at next logon

User cannot change password

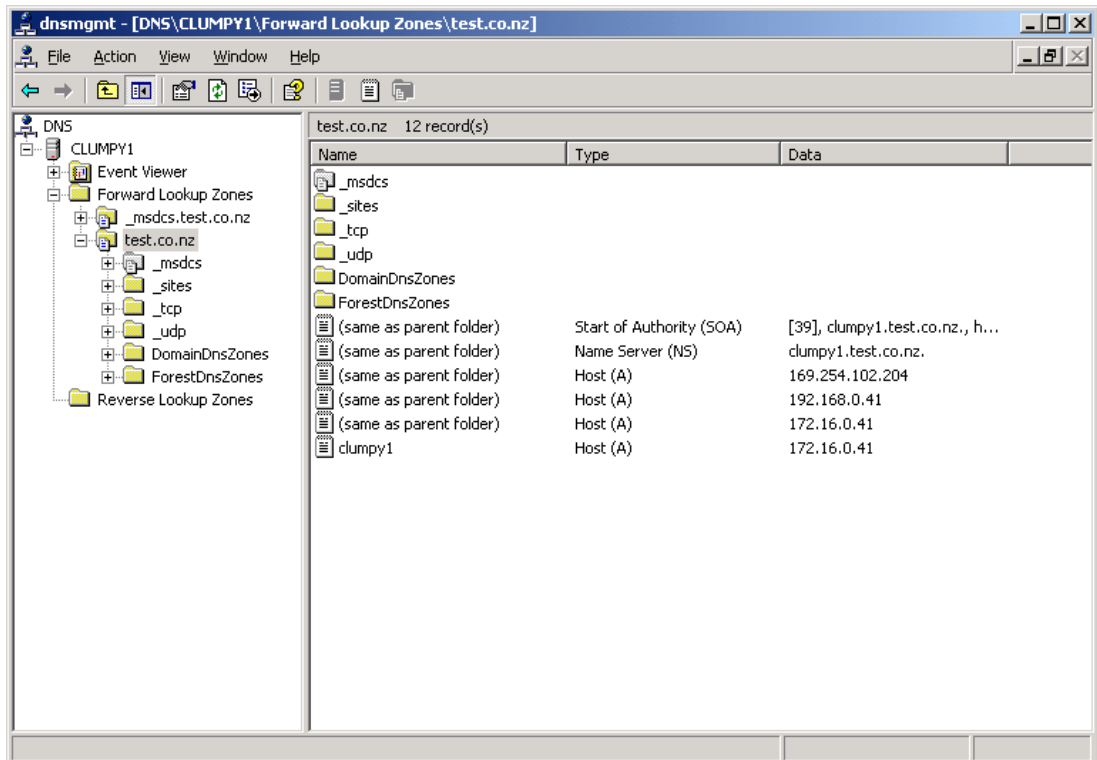
Password never expires

Account is disabled

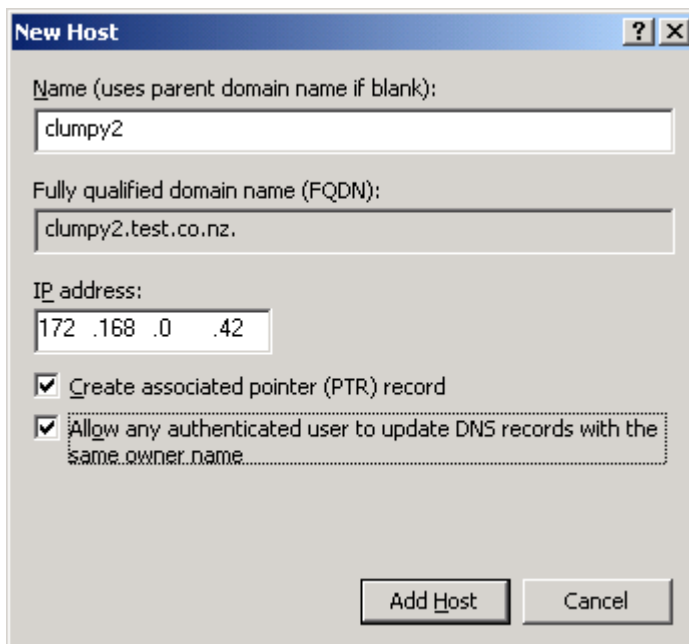
< Back Next > Cancel



► Create extra entries for the Cluster in DNS

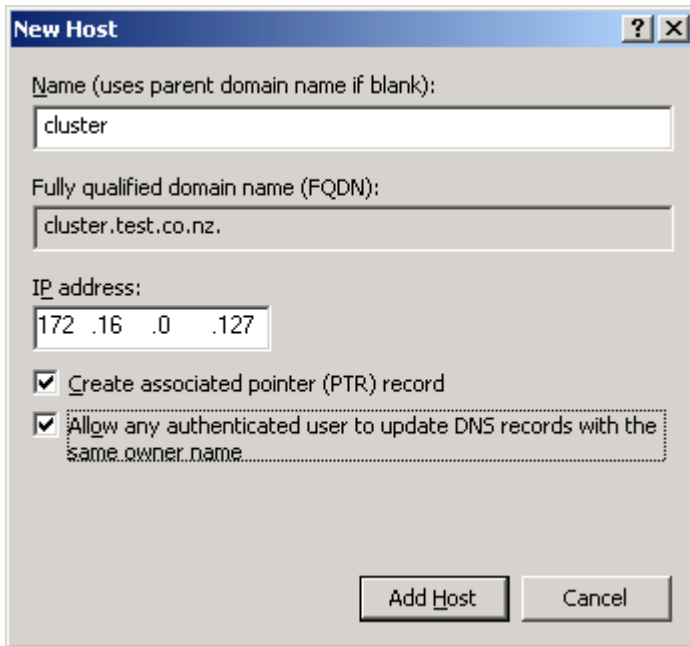


Right click in the right hand window and select “New Host (A)”. Clumpy1 is already there because that is the server we are using now. We need to add clumpy2, the gateway and the virtual IP entry for the cluster.

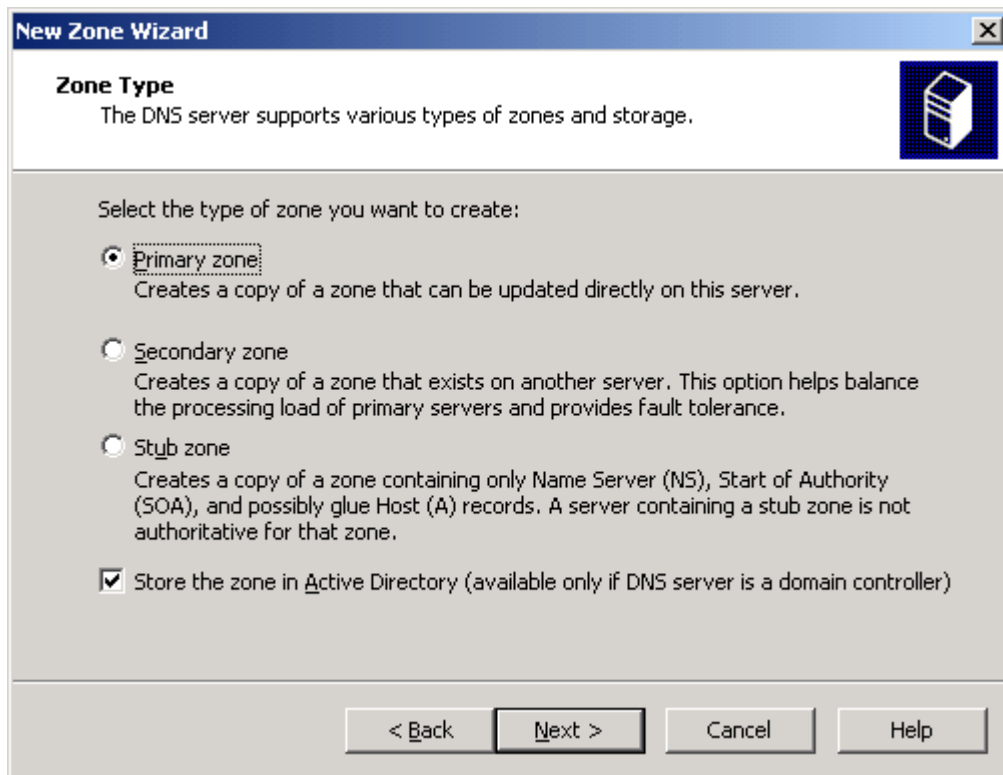


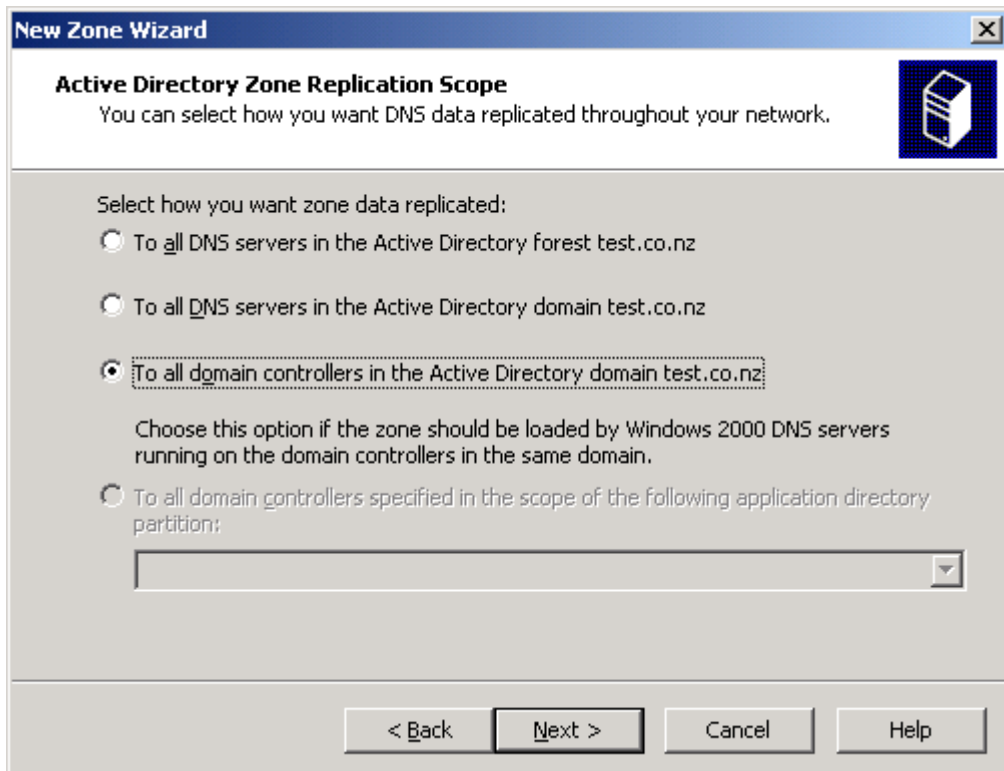
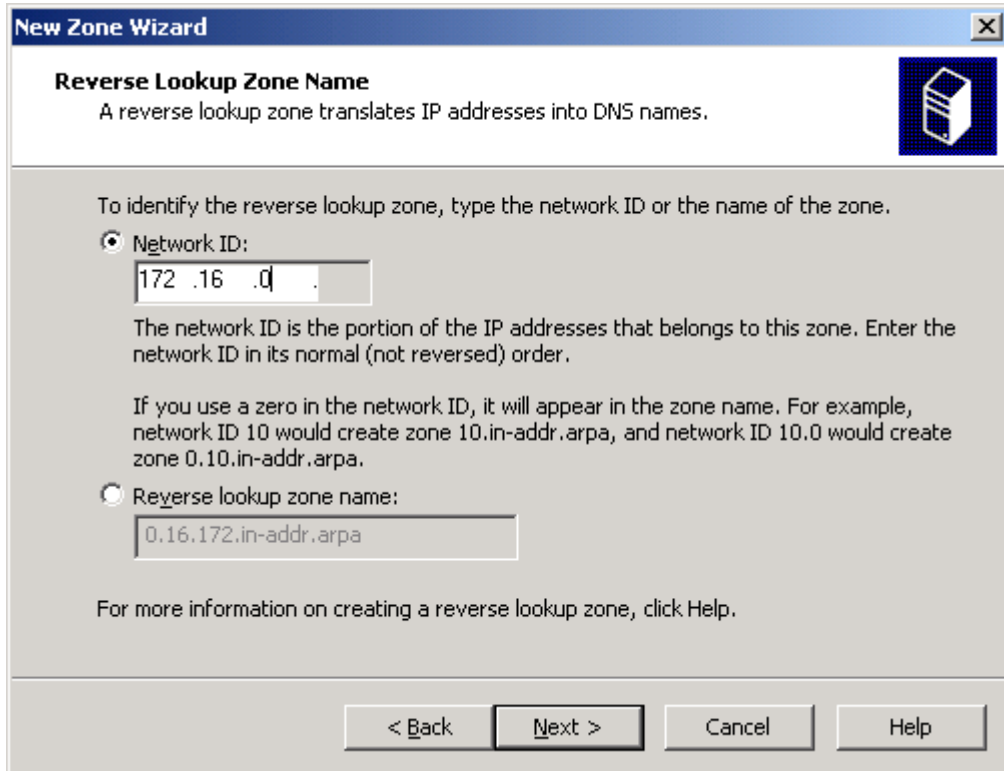


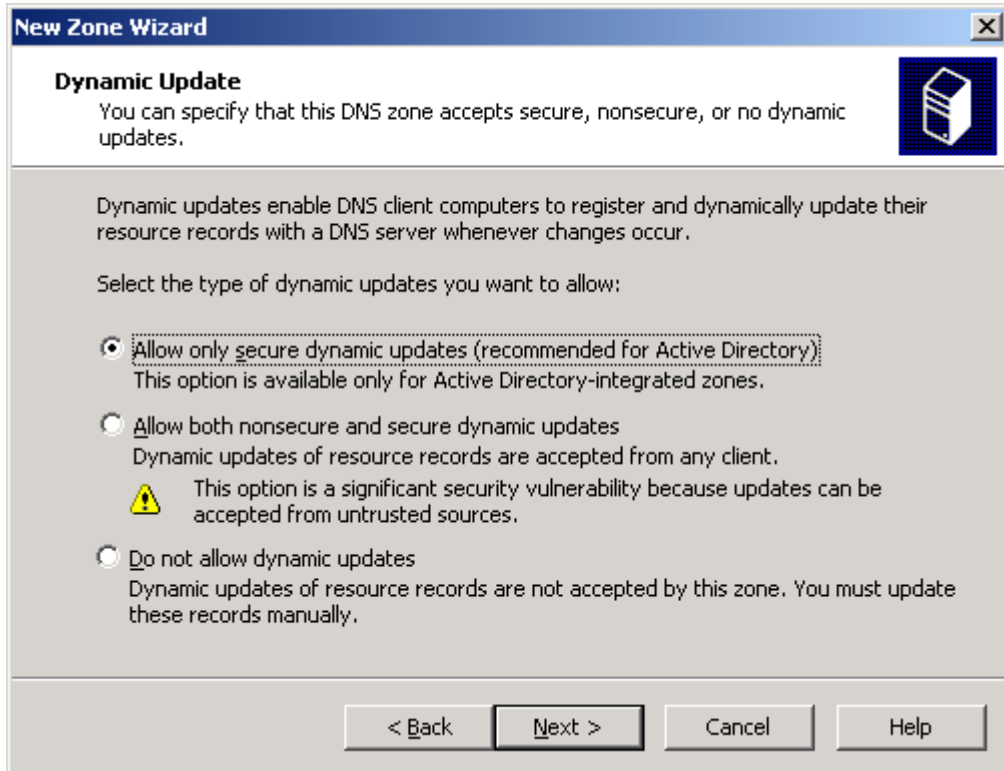
You get the following warning if the 172.16.0.0/24 (reverse) zone hasn't been created yet. You can ignore this and manually add an entry later to the reverse zone. It is probably better to create the reverse lookup zone before adding PTR records. This example will cover the creation of a reverse zone next

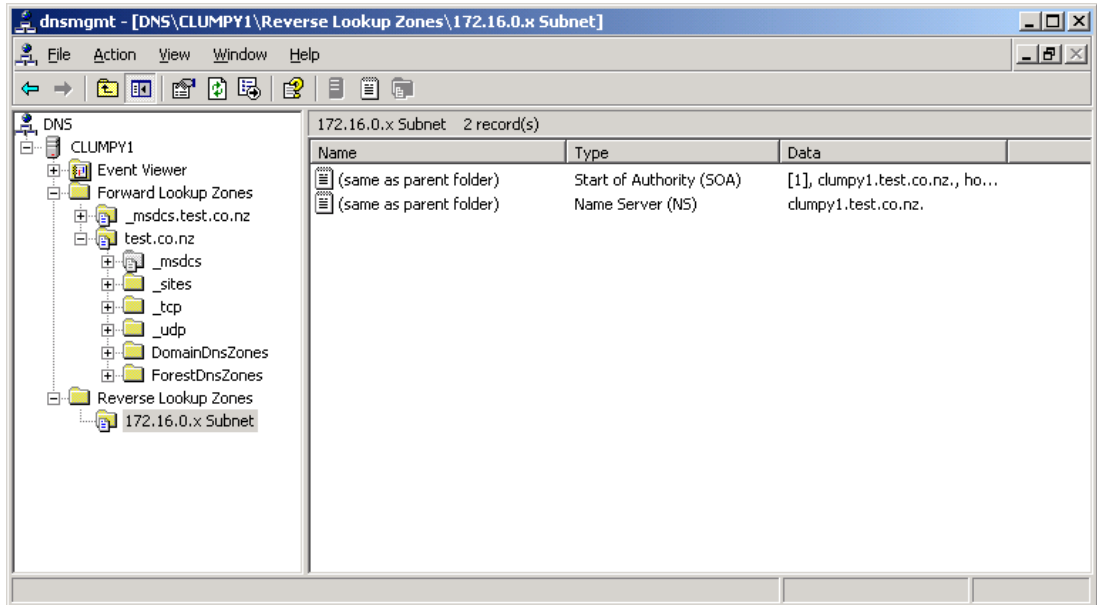


► Create a reverse zone for the test.co.nz domain



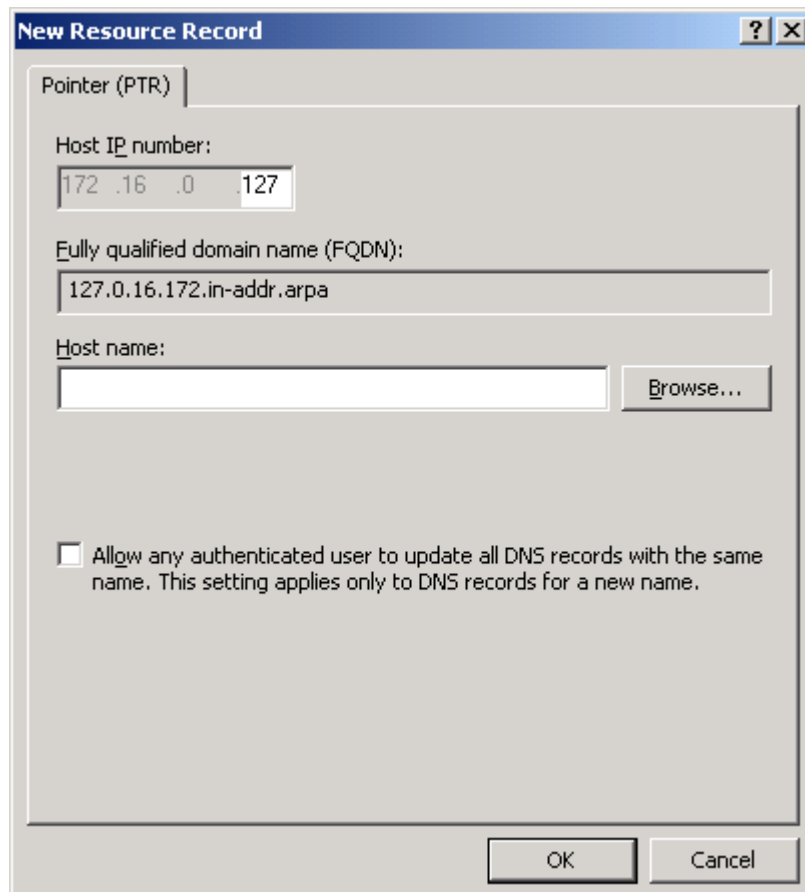




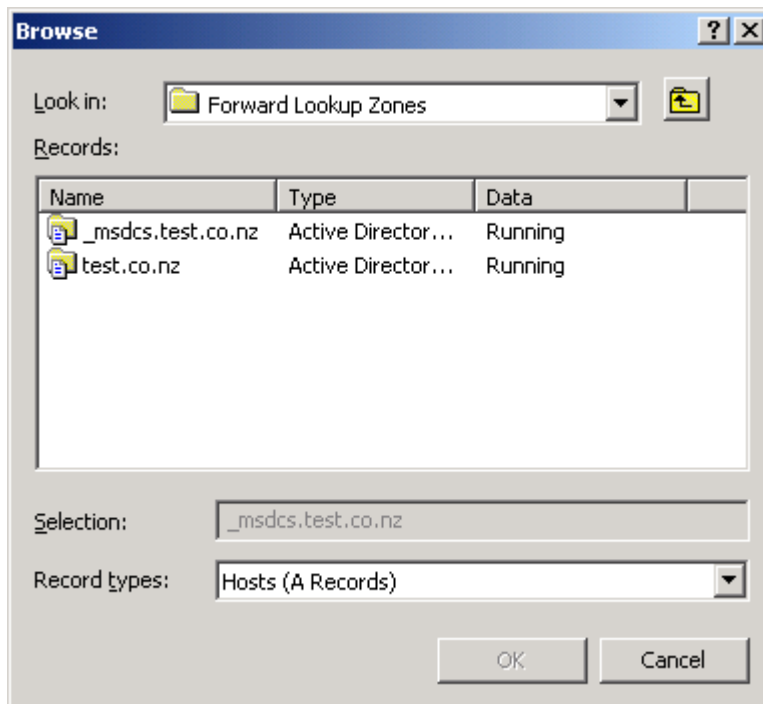
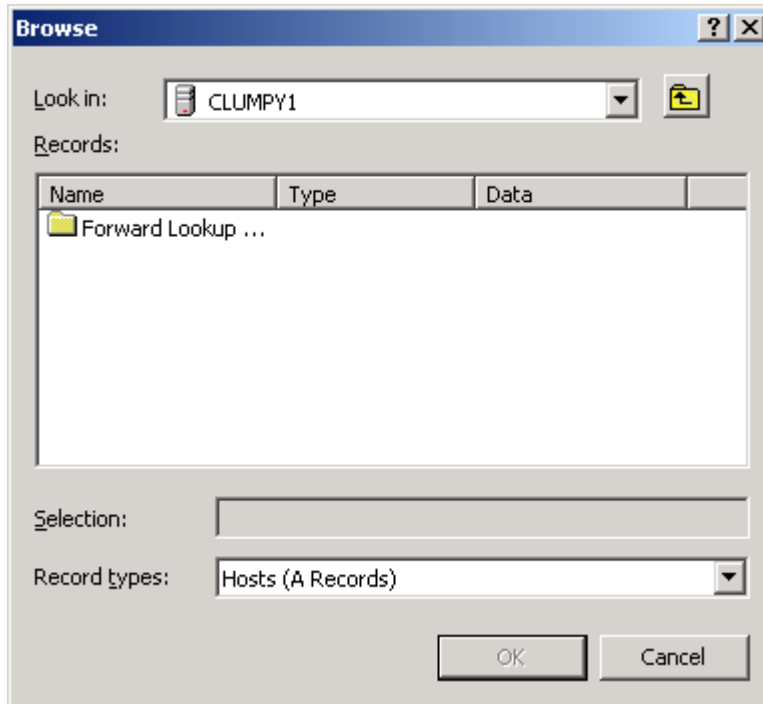


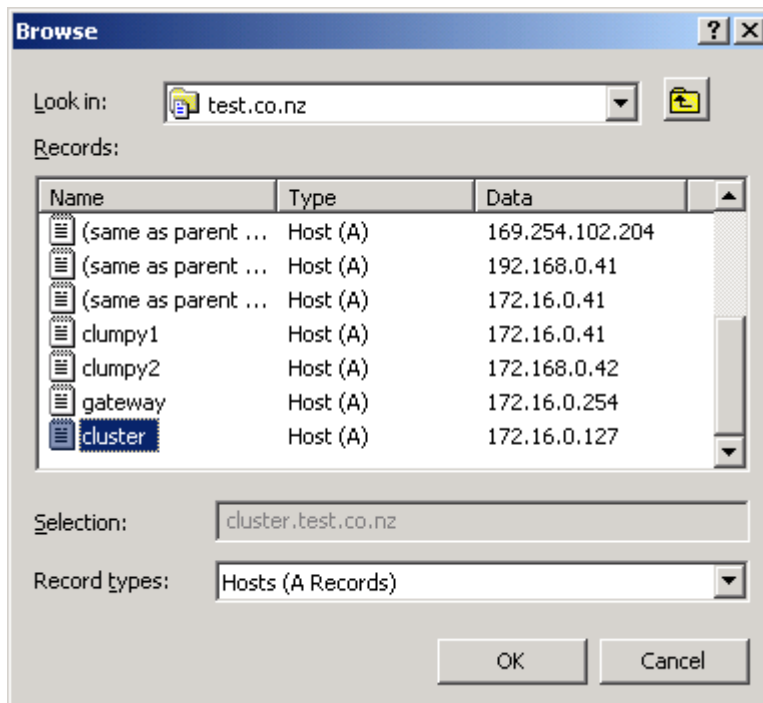
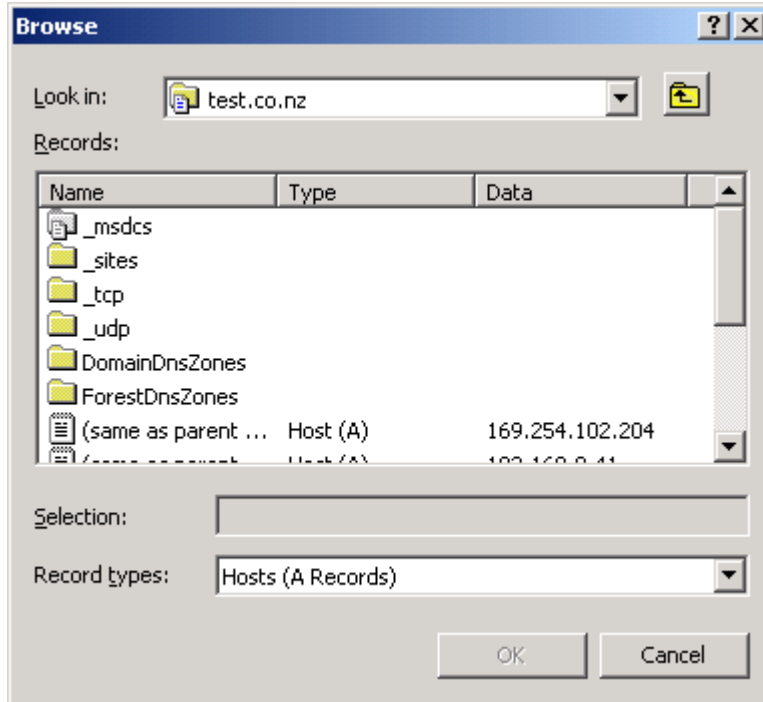
► Manually add the reverse entry for each name (clumpy1, clumpy2, cluster and gateway)

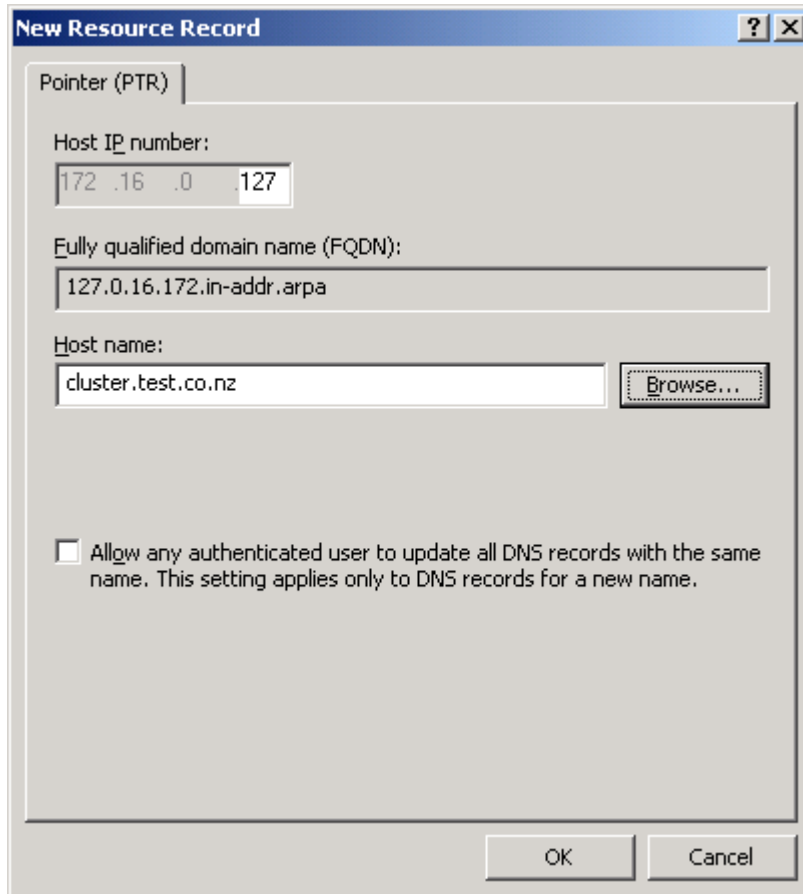
- I. Right click in the right hand pane of the windows above and select “New Pointer (PTR)”. Here is the example for the name “cluster” (for the cluster IP 172.16.0.127):



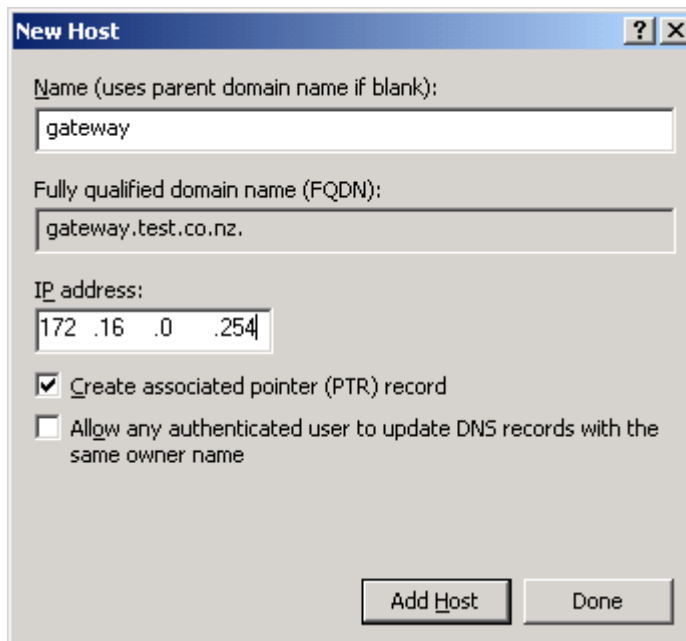
- Click “Browse” in the window above and find the name that you want the reverse entry to relate to.





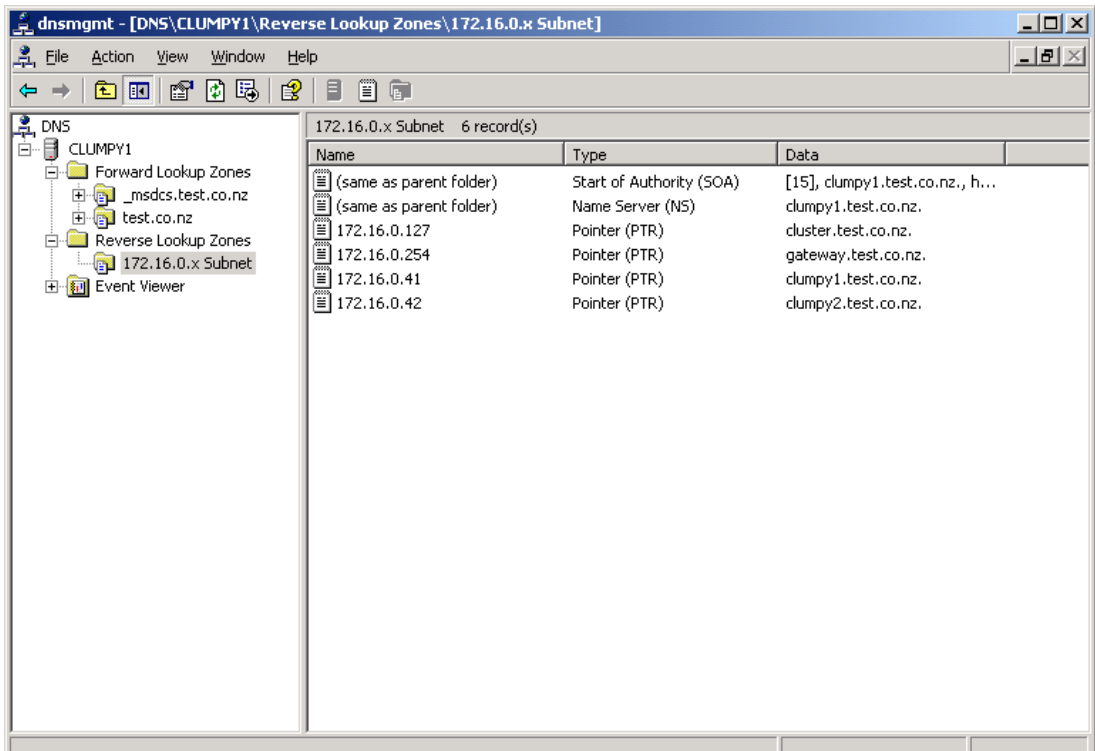
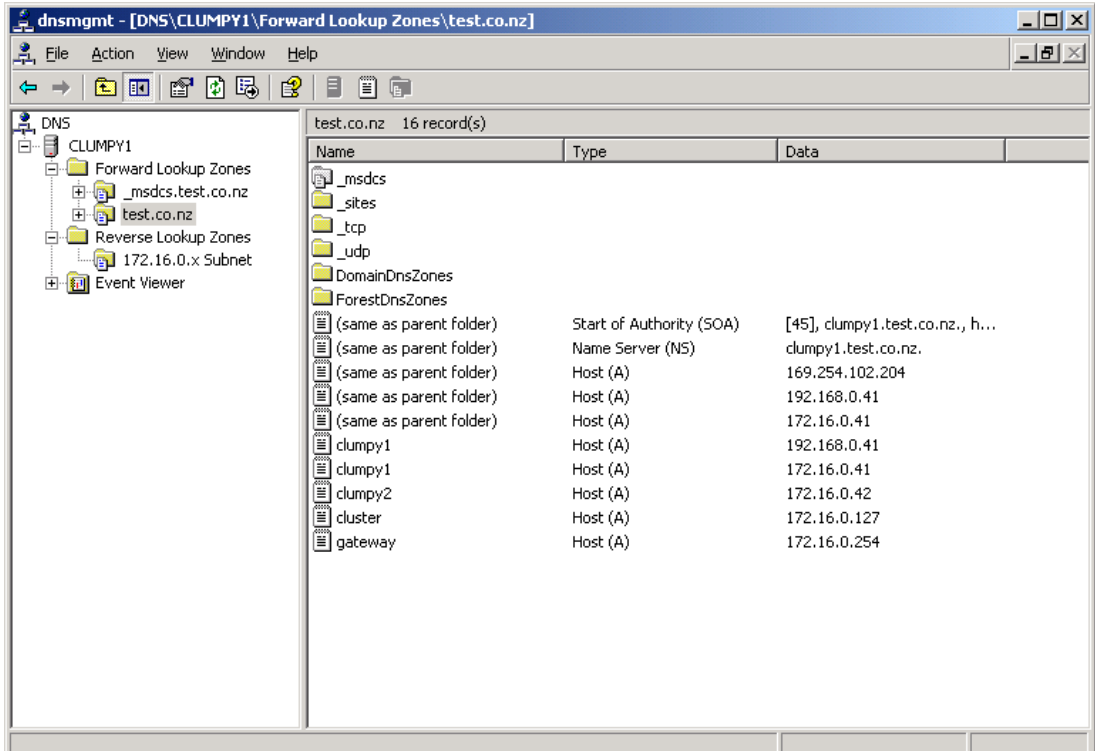


3. Do the same for the gateway address (have you added this to the forward zone yet?):





Appendix 1 —Quick Start Guide to setting up a Windows 2003 NLB cluster



► Set up the NLB Cluster

1. Start the NLB Manager from:  
Start > Administrative Tools > Network Load Balancing Manager
2. Start the wizard to create a new cluster:  
Cluster > New

The screenshot shows the 'Cluster Parameters' dialog box. It is divided into three main sections:

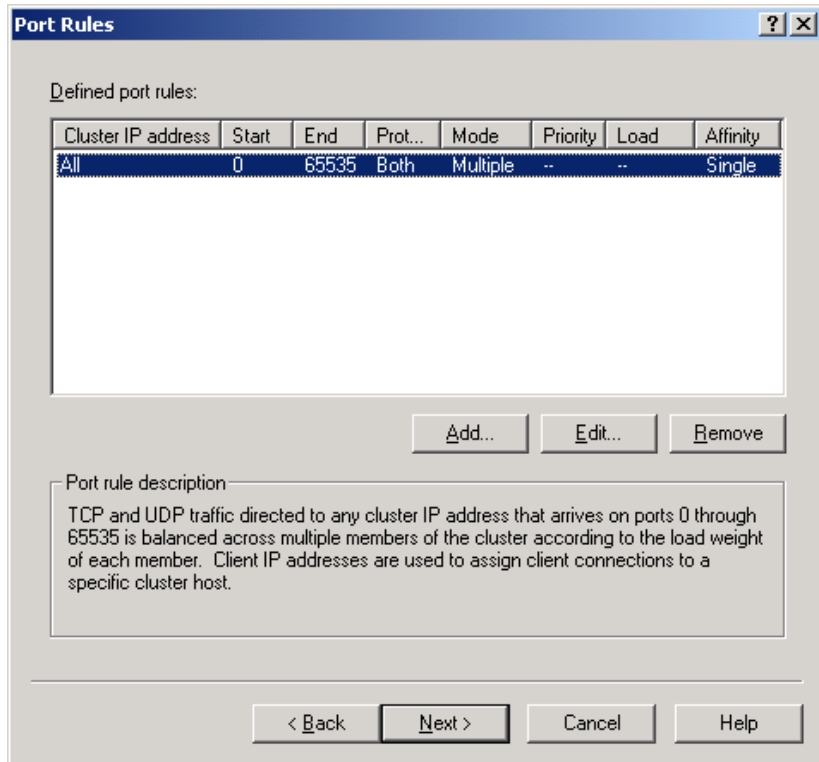
- Cluster IP configuration:** Contains four text input fields:
  - IP address: 172 . 16 . 0 . 127
  - Subnet mask: 255 . 255 . 255 . 0
  - Full Internet name: cluster.test.co.nz
  - Network address: 03-bf-ac-10-00-7f
- Cluster operation mode:** Contains three radio button options:
  - Unicast
  - Multicast
  - IGMP multicast
- Allow remote control:** Contains a checkbox labeled 'Allow remote control' which is unchecked, and two password input fields labeled 'Remote password' and 'Confirm password'.

At the bottom of the dialog are four buttons: '< Back', 'Next >', 'Cancel', and 'Help'.

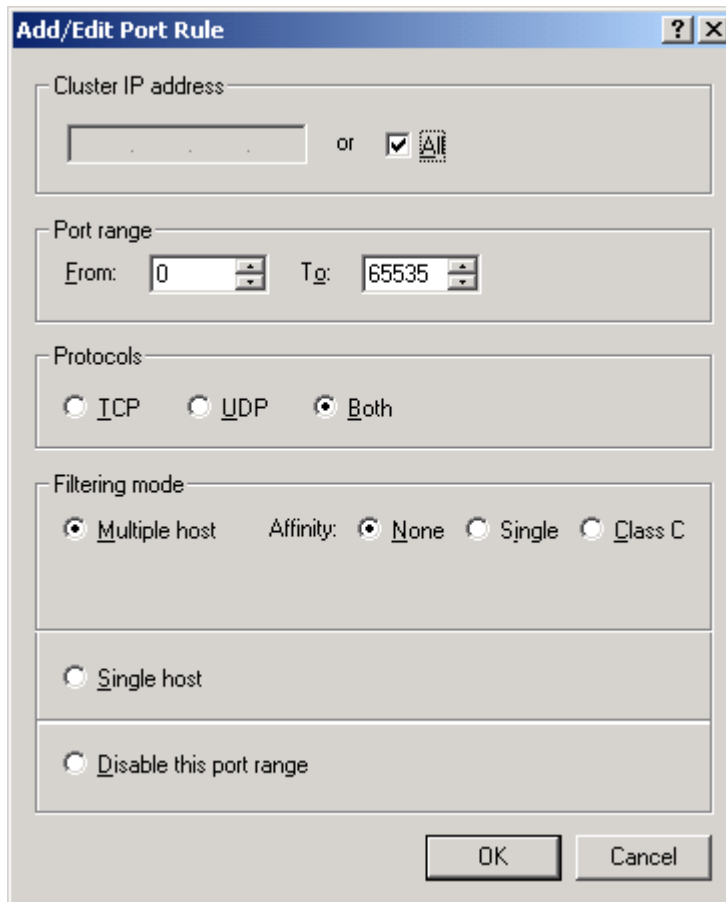
The screenshot shows the 'Cluster IP Addresses' dialog box. It is divided into two main sections:

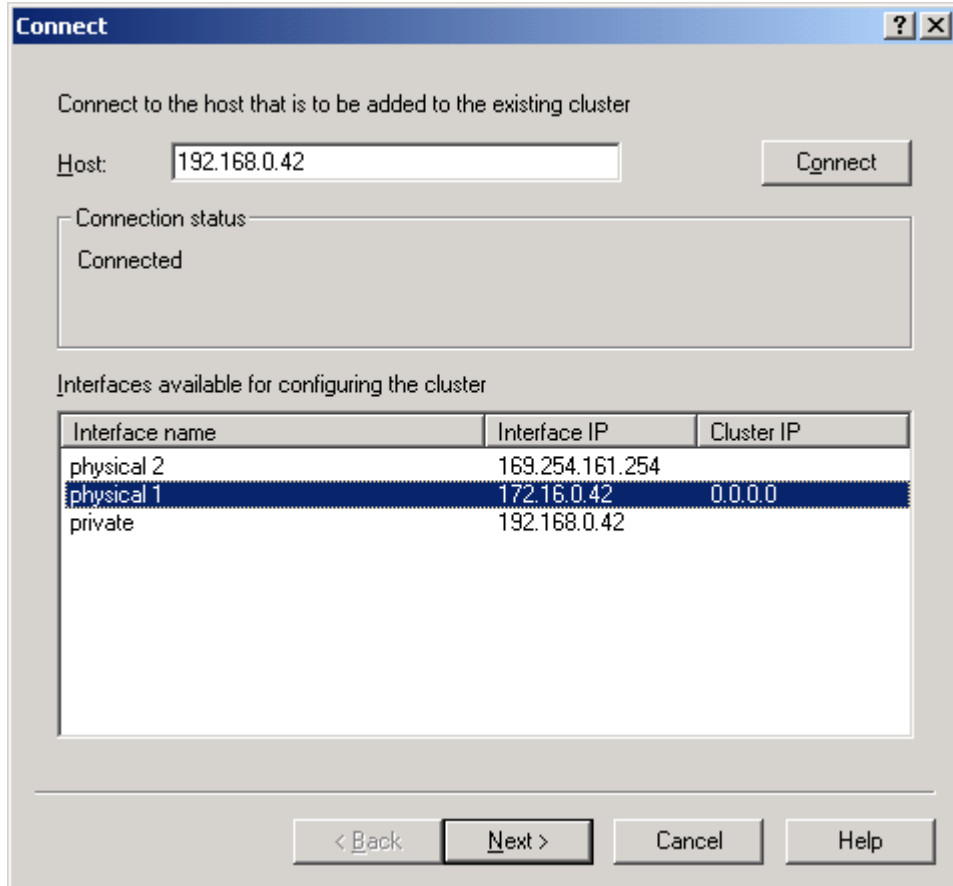
- Primary cluster IP address:** Contains two text input fields:
  - IP address: 172 . 16 . 0 . 127
  - Subnet mask: 255 . 255 . 255 . 0
- Additional cluster IP addresses:** Contains a table with two columns: 'IP address' and 'Subnet mask'. The table is currently empty. Below the table are three buttons: 'Add...', 'Edit...', and 'Remove'.

At the bottom of the dialog are four buttons: '< Back', 'Next >', 'Cancel', and 'Help'.



3. Change the filtering mode to affinity “None”.

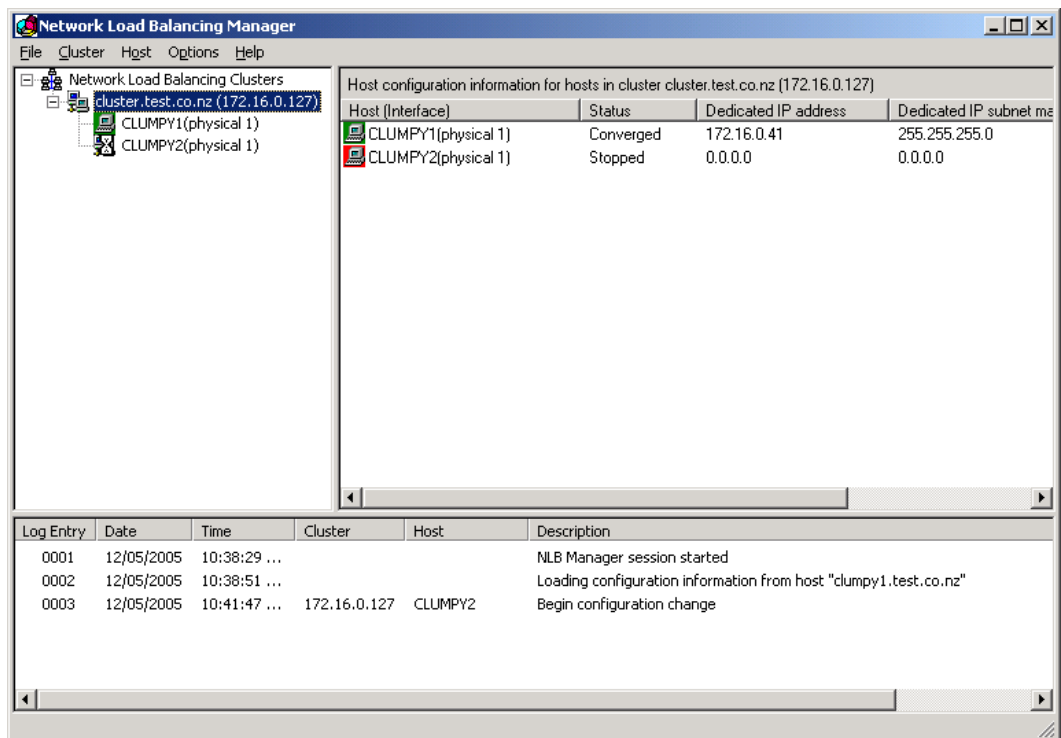
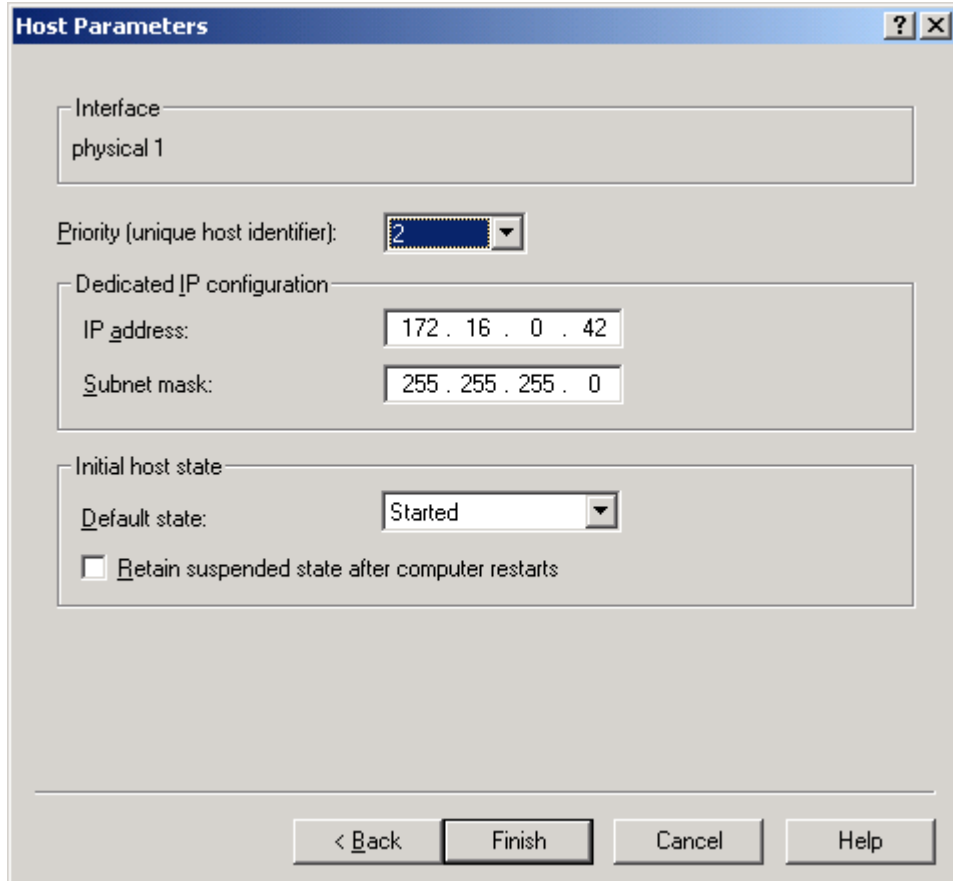




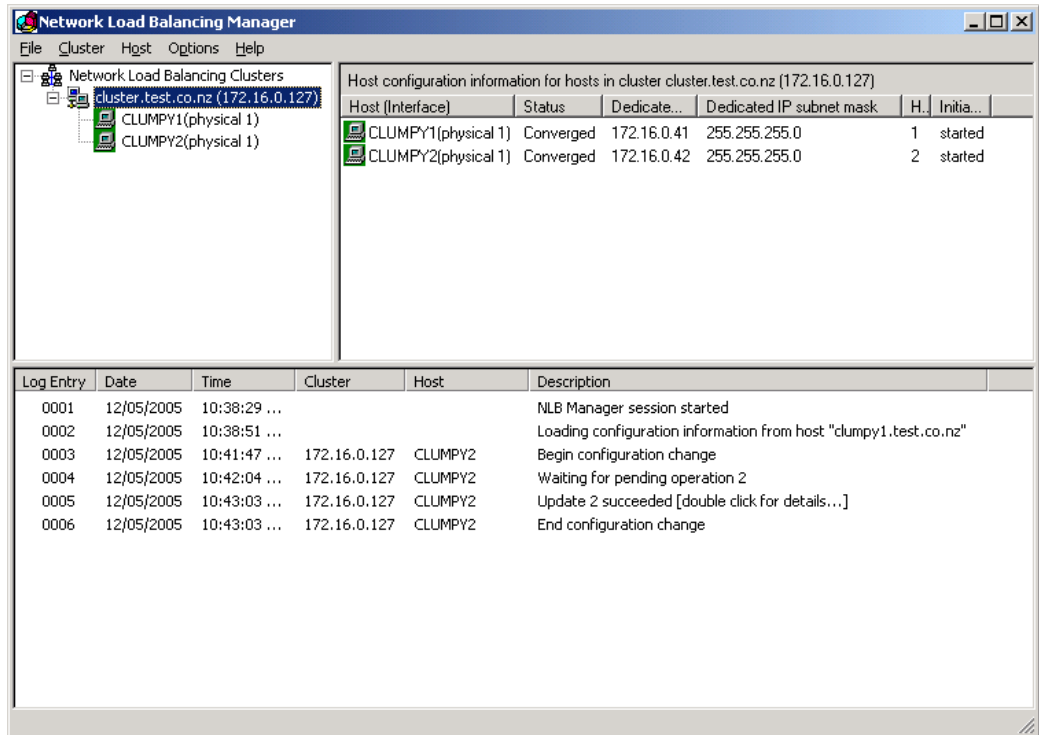
4. Enter the local IP in the “Host” box and click on the “Connect” button. Wait for as long as it takes and then select the cluster side IP and click on Next. The box above illustrates adding the second node, which you should do after you have added the first node. For the second node you enter the second NIC's private network IP address in the “Host” box. For the second node it may take some time after pressing the “Connect” button for the interfaces to appear in the available interfaces box. Be patient.

Notice that there are already IP addresses assigned to the other interfaces of clumpy2. If you have not configured interfaces other than the private NIC of clumpy2 that is okay. On unconfigured interfaces your clumpy2 will have IP addresses of the form 169.254.0.0 or perhaps 0.0.0.0. All you need is an IP on the private NIC of clumpy2 and a physical connection, to ensure basic IP connectivity.

The box below illustrates that for the second node you must change the priority to 2.

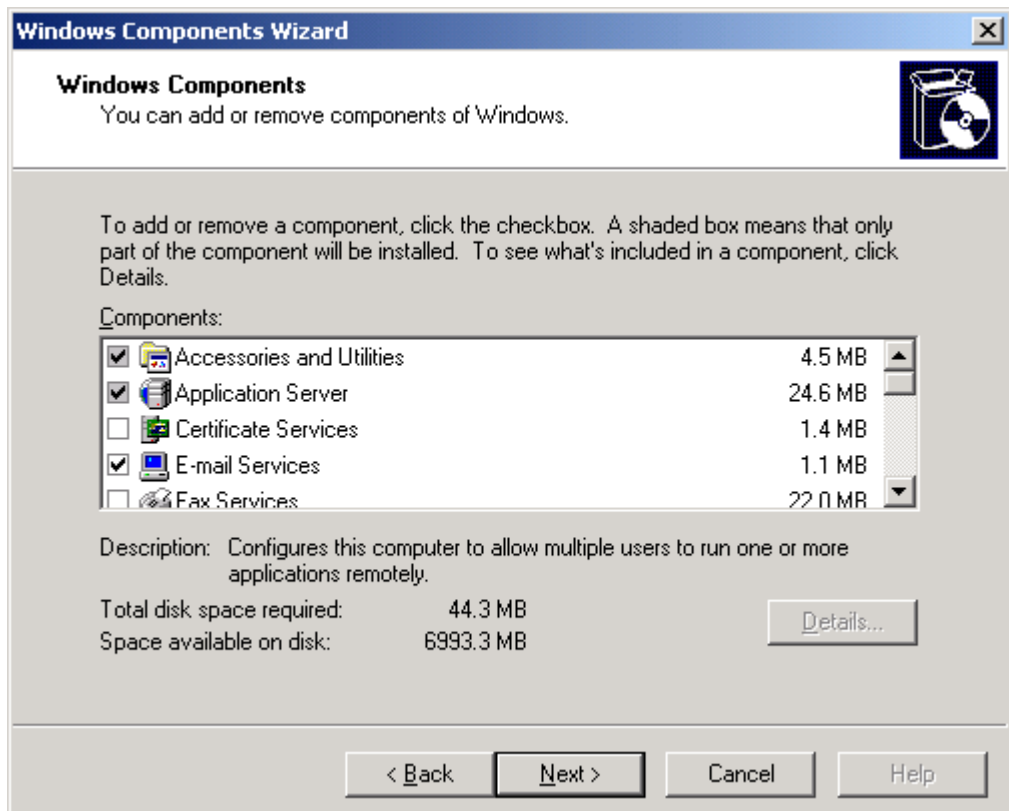


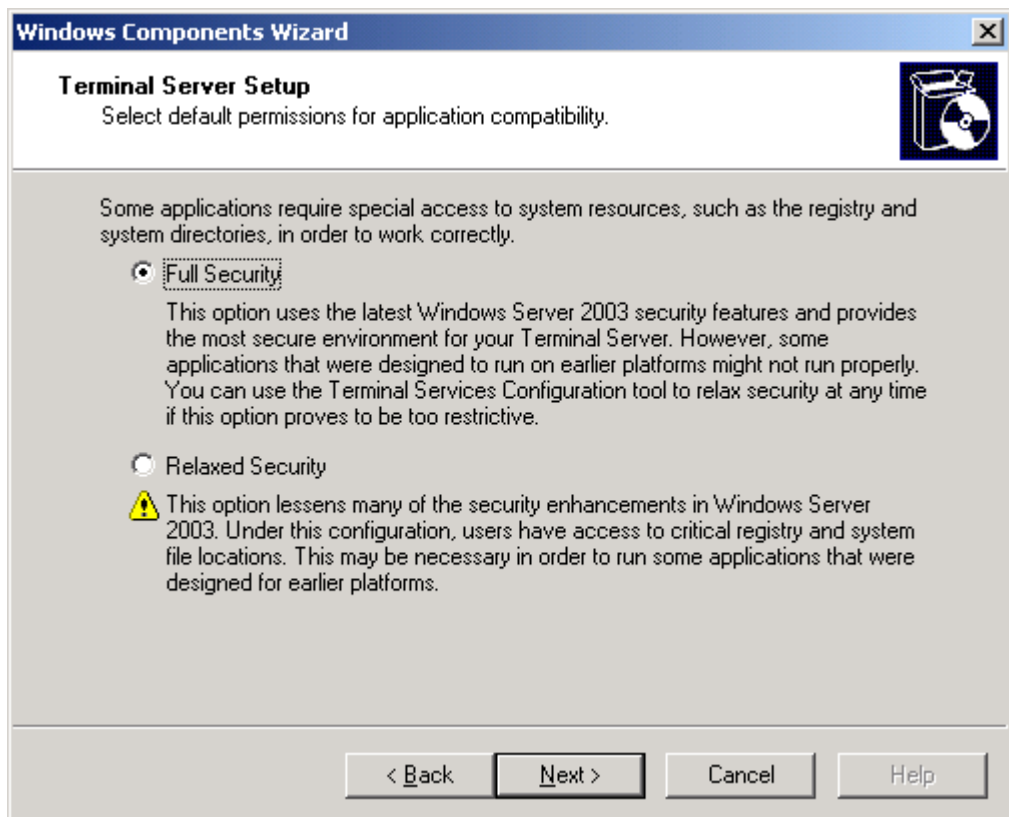
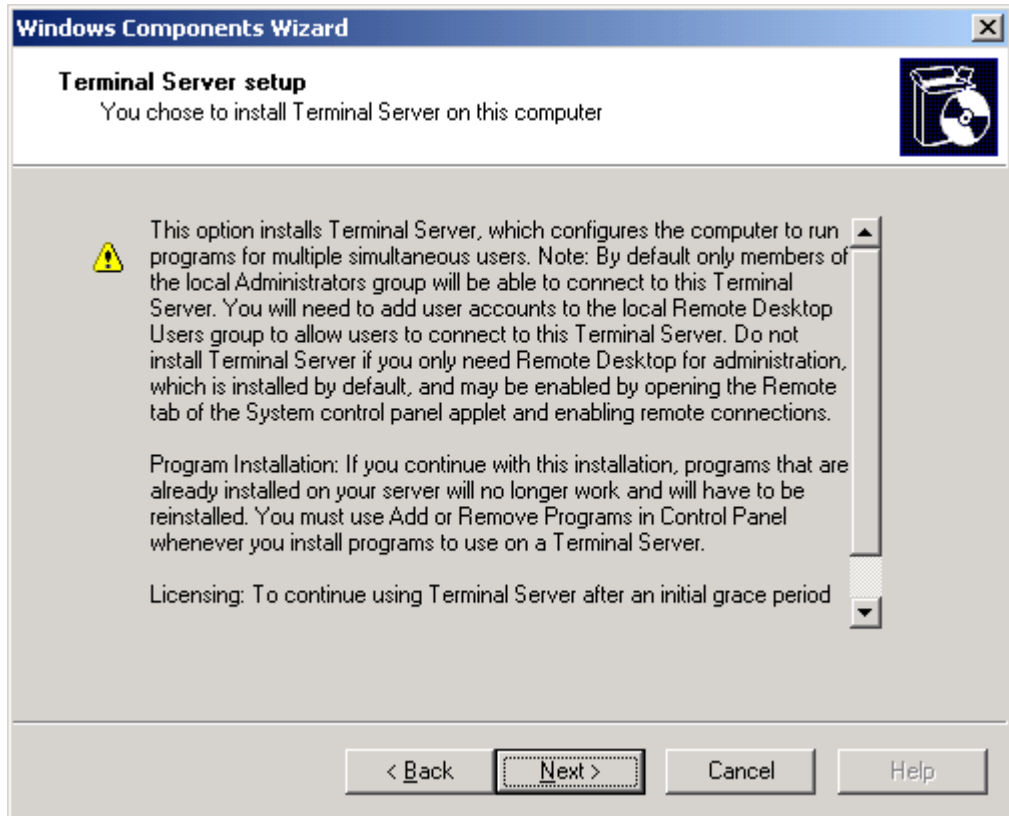
It takes quite a long time sometimes for the nodes to converge. Be patient.

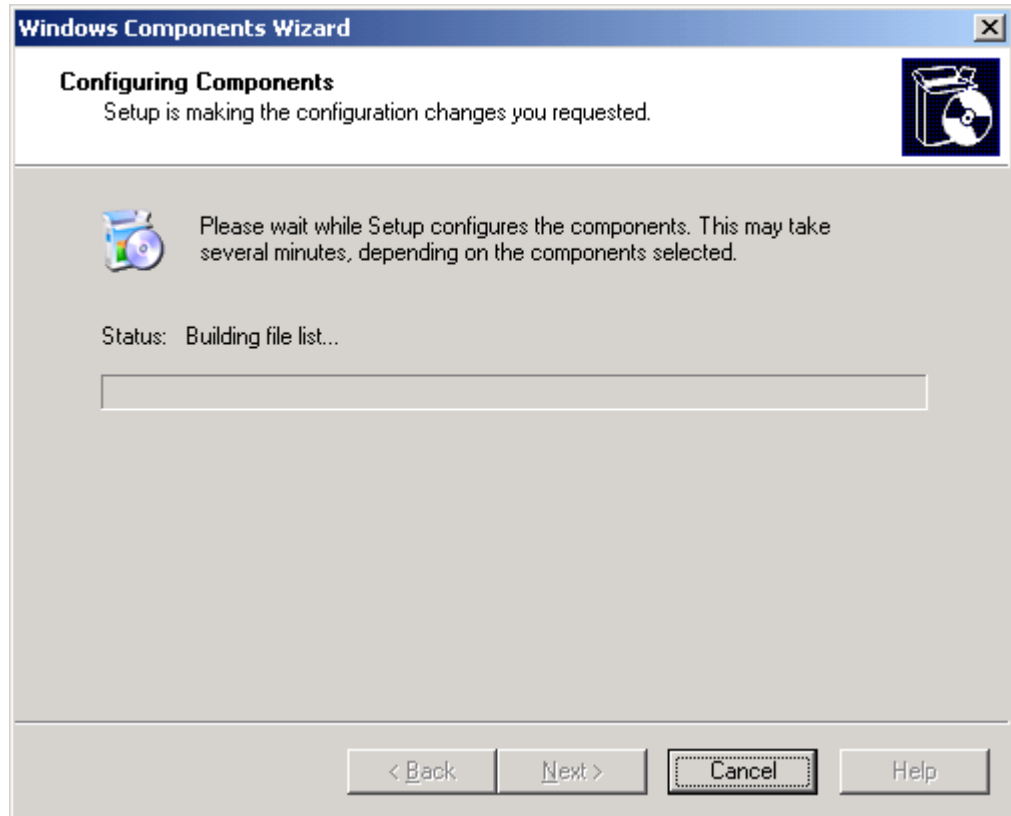


► Add IIS to enable node as a Web server

1. From the Control Panel select “Add or Remove Programmes” and then “Add/Remove Windows Components”.







## Installation of second NLB server - clumpy 2

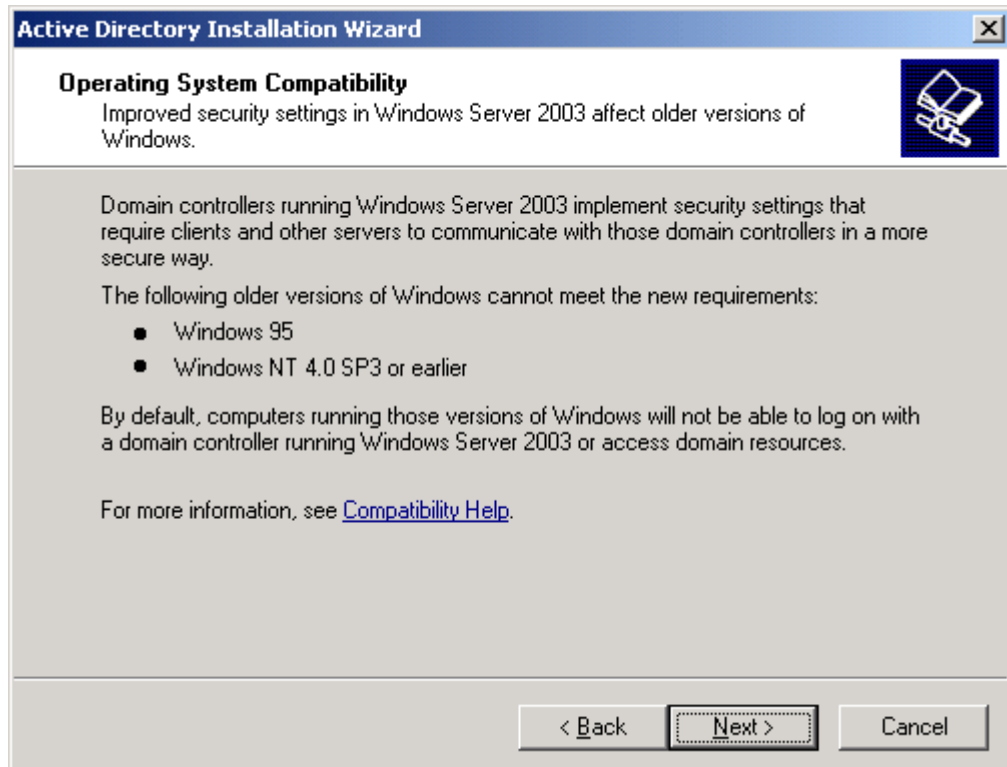
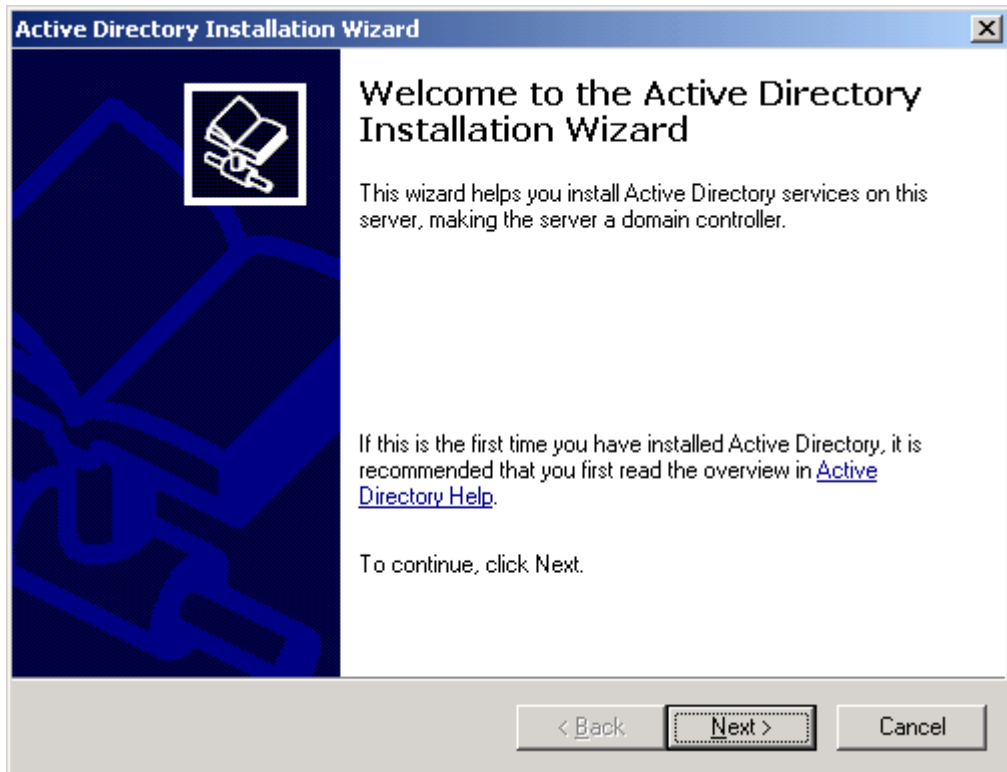
- ▶ Convert your file system to NTFS if it is not already

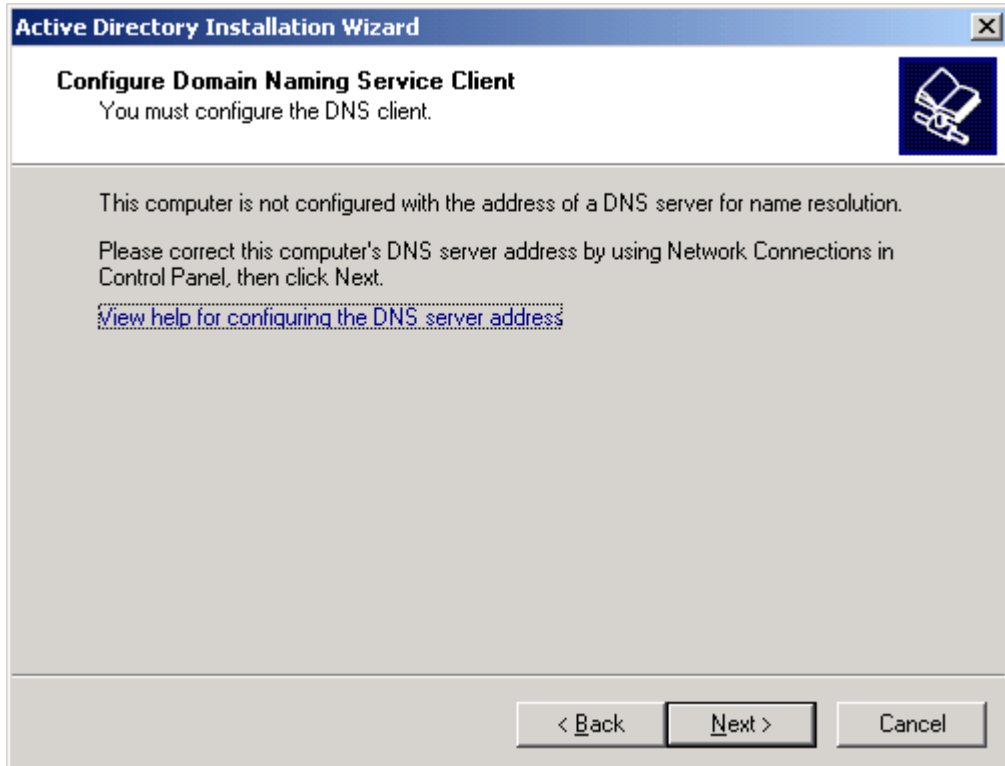
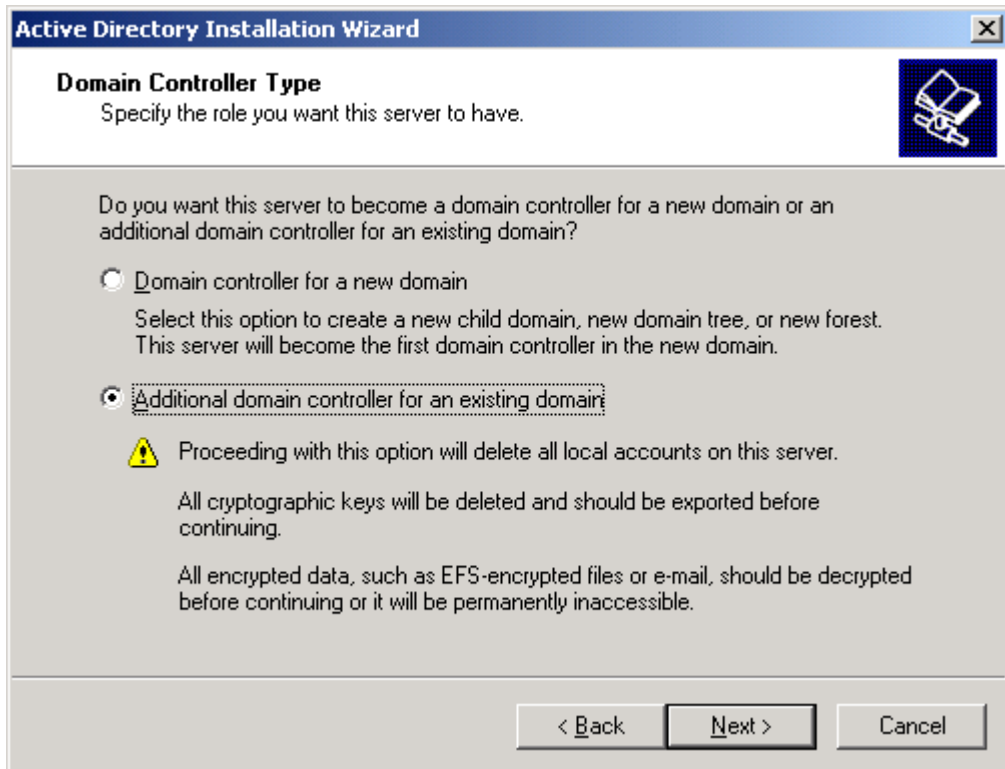




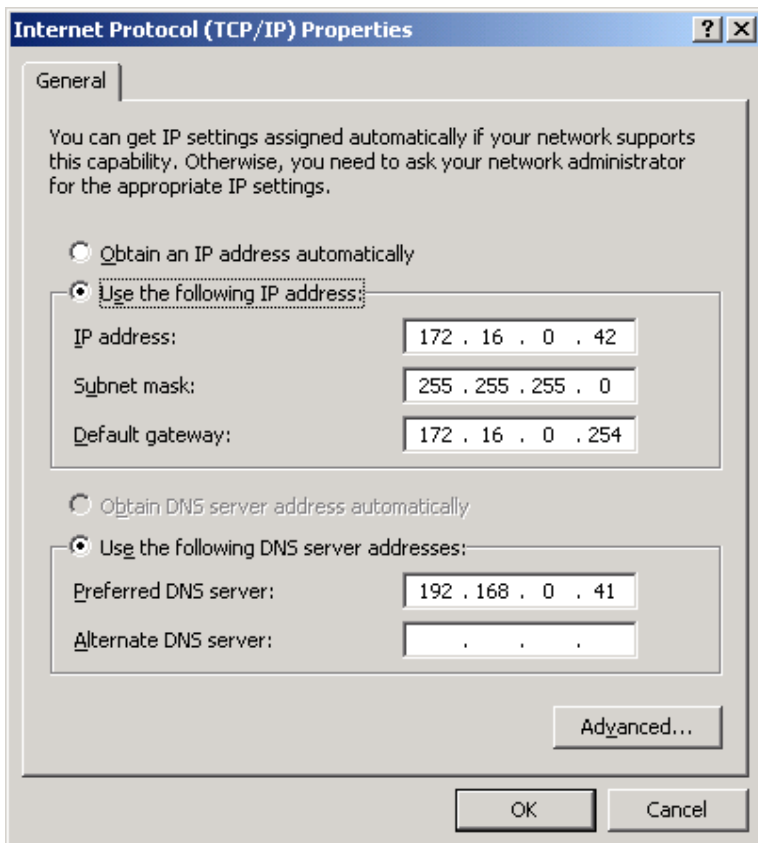
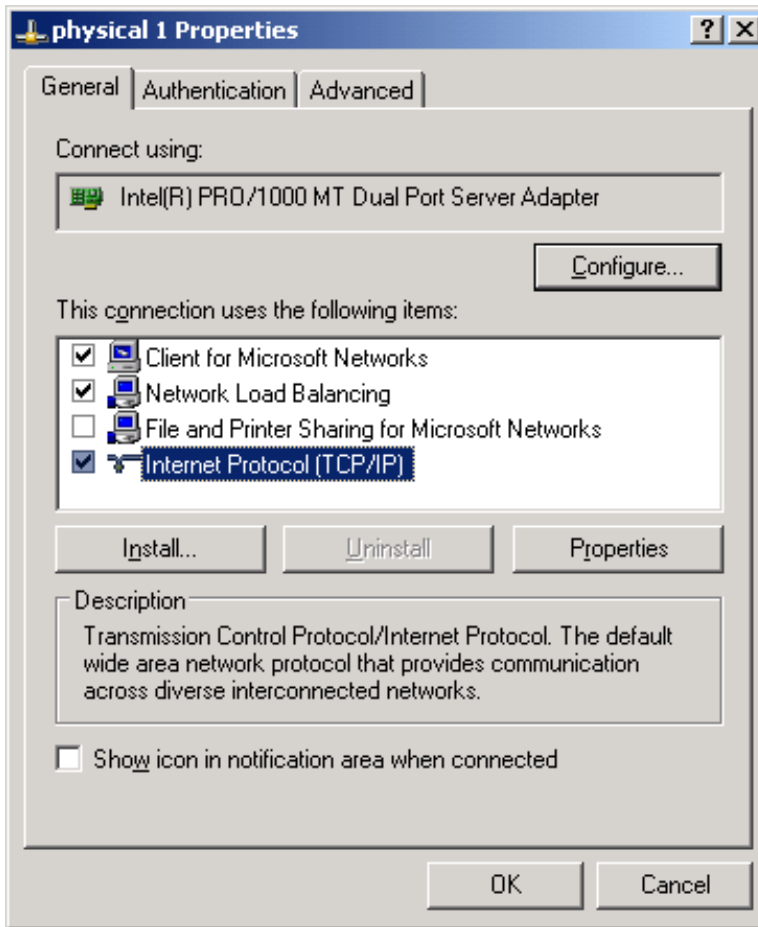
► Install the Domain Controller on clumpy2

Don't set the second server as a backup DNS server—that's too much extra bother at this stage. Use the private network for DNS in the adapter settings.

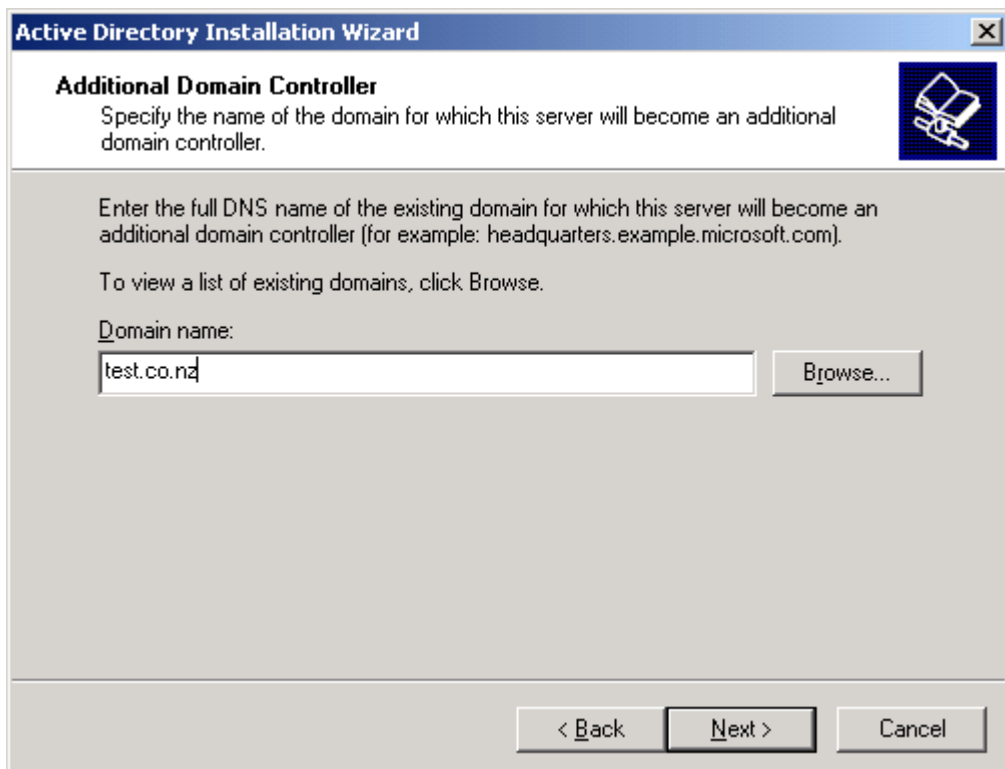
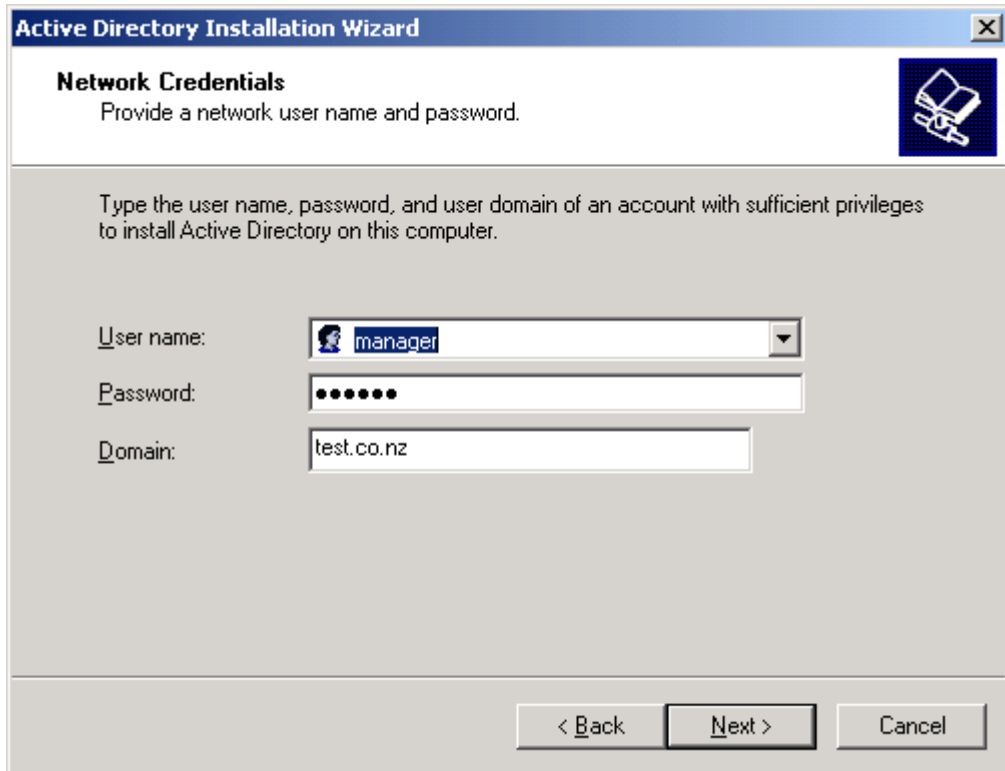


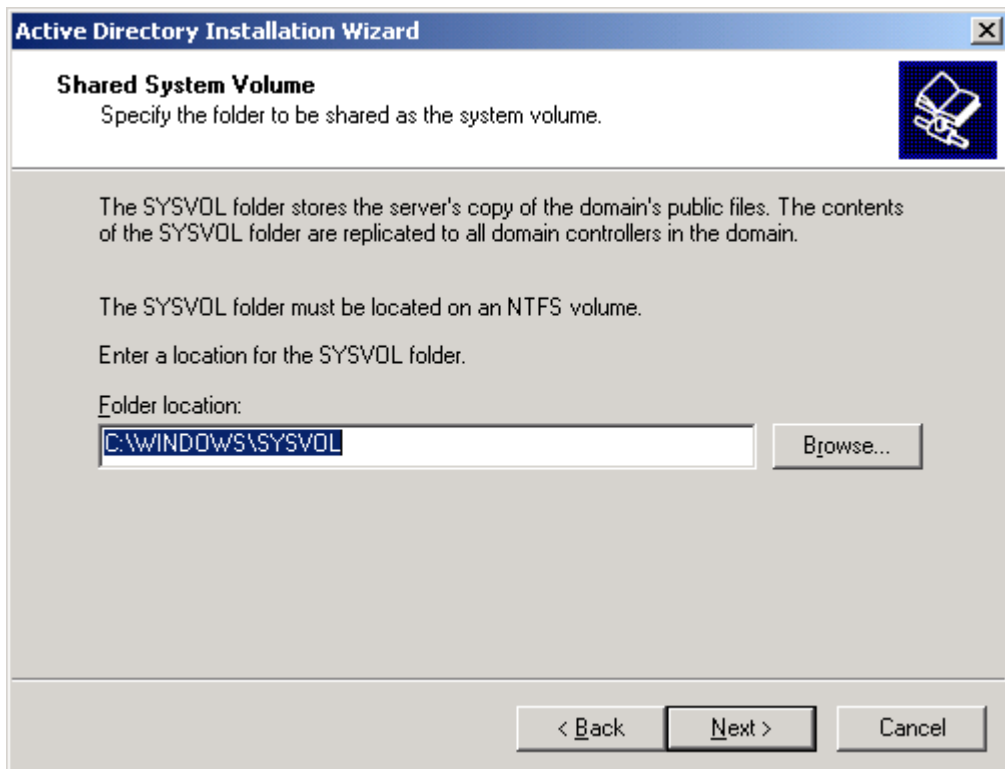
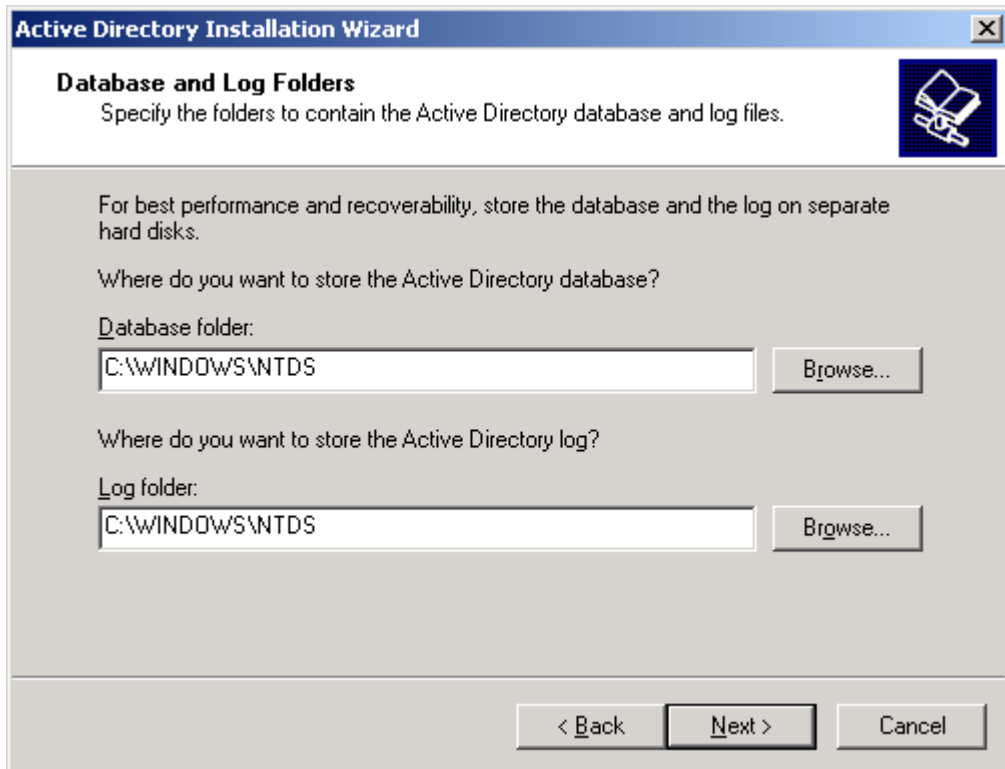


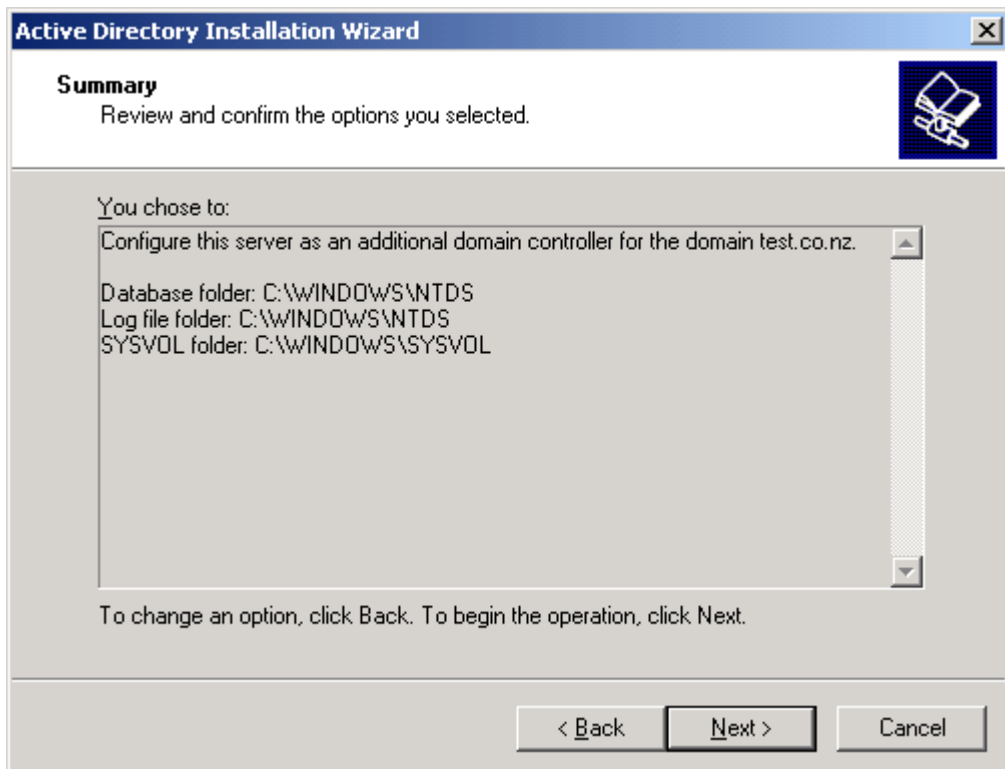
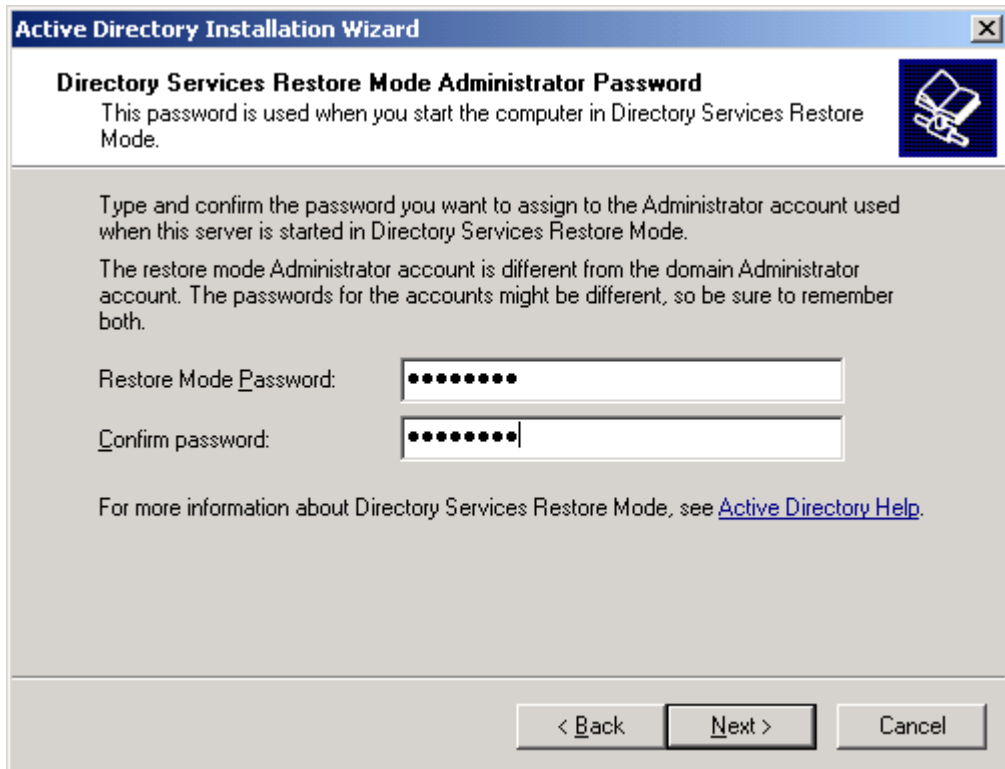
If you get the message above, go back and set up a DNS server through the Network Connections settings in the Control Panel

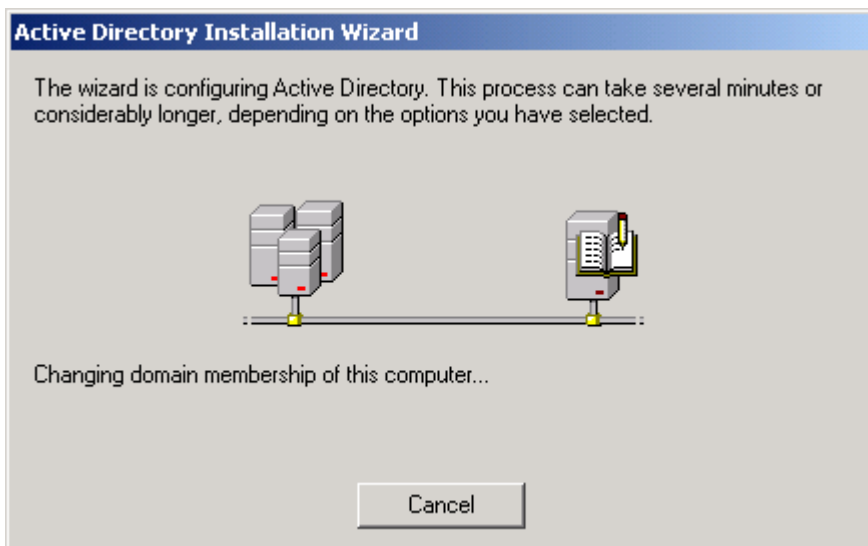
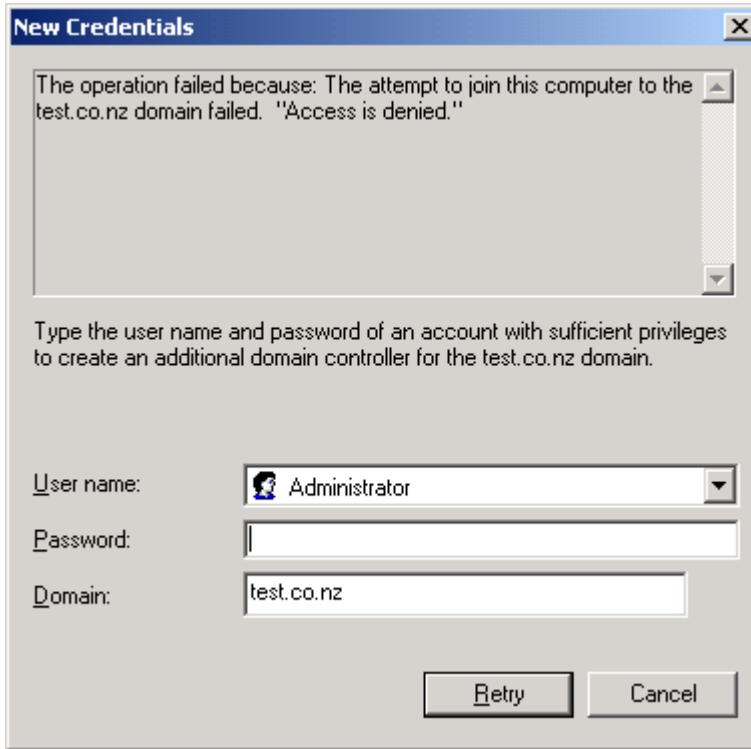


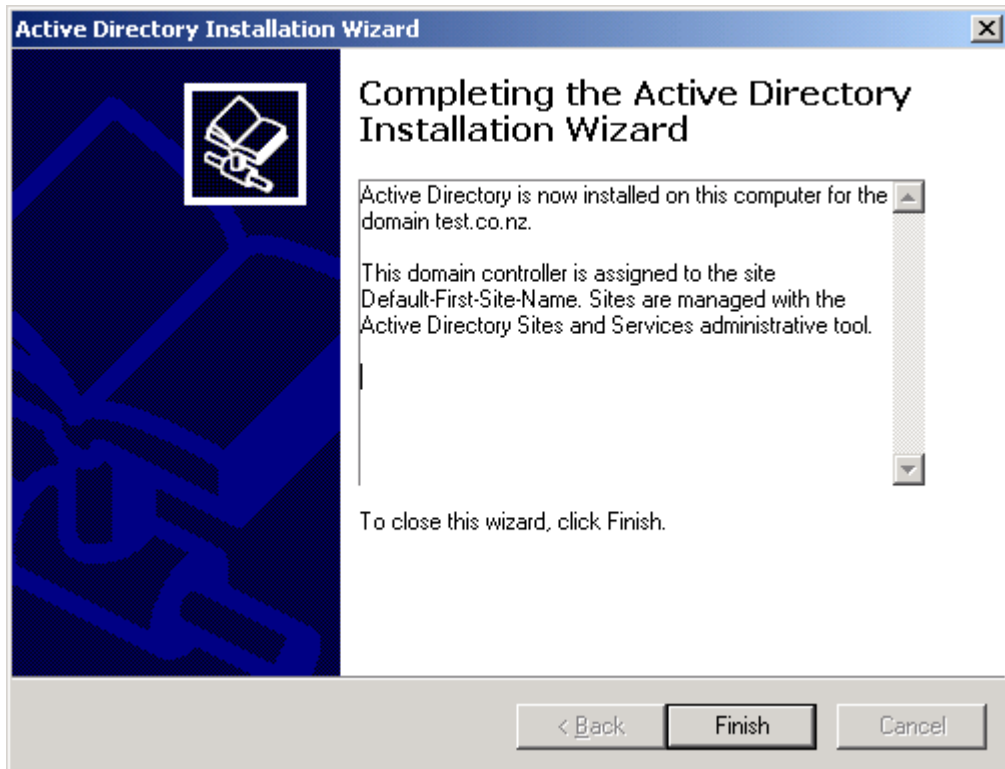
Use the Administrator account, in this example, you can see that manager was used and then subsequently rejected a few steps down the line.











Install Windows IIS if you haven't already (refer to the clumpy1 setup).

You do not need to do anything to clumpy2 for it to join the NLB cluster. The earlier use of the NLB Manager on clumpy1 should have configured clumpy2 accurately.

You will need to make many tweaks to the NLB policy to suit your environment. Use the NLB Manager to do that. Also, you may need to spend time convincing Windows 2003 to use the private NIC for cluster heartbeat messages.

This document should at least leave you with a basic, but working cluster.



## Appendix 2: IGMP snooping port membership information in switch memory tables

The simplest way to find out which ports belong to a group is to use the command **show igmpsnooping**, as described in "Multicast mode with the IGMP option selected" on page 5.

Additionally, you can find specific information about which ports belong to a group in the layer 2 multicast tables. This section describes how to interpret the table output.

**SwitchBlade and AT-9800** On a Switchblade or AT-9800 series switch, use the following command to match ports to multicast groups (represented by the multicast MAC address):

```
show switch memory=l2broadcast
```

This produces two tables, as the following example shows for a 12-port AT-9800 switch. Instructions on how to interpret the output is after this example.

```

Manager > sh swi tab=l2broadcast
CXE L2 multicast and VLAN broadcast table:

```

Index	MAC	VID	L2MC-Group	MC Hits	PV	Prio	Perm	DIS	DM	Expire
0	55-55-55-55-55-55	1	4096	1 0	0	0	1	0	0	0
<b>1</b>	<b>01-00-5e-7f-00-7f 1</b>	<b>5</b>	<b>5</b>	<b>1 5</b>	<b>0</b>	<b>0</b>	<b>1</b>	<b>0</b>	<b>0</b>	<b>0</b>
2	01-00-5e-00-01-18	1	6	1 6	0	0	1	0	0	0
3	01-00-5e-7f-ff-fe	1	7	1 7	0	0	1	0	0	0
4	01-00-5e-00-01-16	1	9	1 9	0	0	1	0	0	0
5	01-00-5e-00-00-00	1	4097	1 1	0	0	1	0	0	0
6	01-00-5e-00-00-00	2	4098	1 2	0	0	1	0	0	0
7	01-00-5e-00-00-00	28	4124	1 26	0	0	1	0	0	0
8	01-00-5e-00-00-00	1	1	1 1	0	0	1	0	0	0
9	01-00-5e-00-00-00	2	1	1 1	0	0	1	0	0	0
10	01-00-5e-00-00-00	28	1	1 1	0	0	1	0	0	0
11	33-33-00-00-00-00	1	2	1 2	0	0	1	0	0	0
12	33-33-00-00-00-00	2	3	1 3	0	0	1	0	0	0
13	33-33-00-00-00-00	28	4	1 4	0	0	1	0	0	0
14	01-00-00-00-00-00	1	4097	1 1	0	0	1	0	0	0
15	01-00-00-00-00-00	2	4098	1 2	0	0	1	0	0	0
16	01-00-00-00-00-00	28	4124	1 26	0	0	1	0	0	0

Index	Segment	L2MC-Group	TagMask	OutportMask	Hits	BCSC-Id	BCSC-Enable	Perm
0	77	0	00000000	00000000	0	0	0	1
<b>1</b>	<b>77</b>	<b>5</b>	<b>00000000</b>	<b>00000021</b>	<b>5</b>	<b>0</b>	<b>0</b>	<b>1</b>
2	77	6	00000000	00000021	6	0	0	1
3	77	7	00000000	00000021	7	0	0	1
4	77	9	00000000	00000008	9	0	0	1
5	83	1	80000000	80000000	1	0	0	1
6	85	2	80000000	8000003b	2	0	0	1
7	85	3	80000000	80000000	3	0	0	1
8	85	4	80000000	80000800	4	0	0	1
9	87	8190	00000000	00000000	0	0	0	1
10	87	8189	00000000	00000000	0	0	0	1
11	87	4097	00000000	0000003b	0	0	0	1
12	87	4098	00000000	00000000	0	0	0	1
13	87	4124	00000000	00000800	0	0	0	1

To work out which ports are members of a group:

1. In the first table, select the row that has the MAC address of the L2 multicast group that you are interested in. Note the number in its “L2MC Group” field and the VID. In the example in bold above, the L2MC Group field value is 5 and the VID is 1, so this is VLAN1.
2. Find the row in the second table that has the same value in its L2MC Group field.
3. In that row, note the OutportMask field. This is a series of hexadecimal digits (0x21 in the example). The right-hand most digits each represent 4 concurrent ports (in the range 1-4 and 5-8 in the example, because they are the two right-most digits).
4. Convert each hexadecimal digit to a 4-digit binary number (0010 and 0001 in the example).
5. Counting from the right, note the positions of the digit 1 in each binary number. Each 1 corresponds to a port that belongs to the target group. Therefore, in this example, ports 1 and 6 are members of the group 01-00-5e-7f-00-7f in VLAN1, as the following table shows.

...	ports 12-9	ports 8-5	ports 4-1	
0	0	2	1	hexadecimal
0000	0000	0010	0001	binary
-	-	6	1	port number

**Rapier,  
AT-8800,  
AT-8600**

On these switches, use the following command to match ports to multicast groups (represented by the multicast MAC address):

```
show switch memory=mar1
```

This produces a single line of output per group, as the following example shows.

```
Manager > sh swi memory=mar1

Multicast Table-0,0 3/(1-255)
-----
1. Mac=01-00-5e-00-01-18 VLAN=1 ports=00000011 utports=03ffffff7 cosDst=0
2. Mac=01-00-5e-7f-00-7f VLAN=1 ports=00000011 utports=03ffffff7 cosDst=0
3. Mac=01-00-5e-7f-ff-fe VLAN=1 ports=00000011 utports=03ffffff7 cosDst=0
```

To work out which ports are members of a group:

1. Select the row that has the MAC address of the L2 multicast group that you are interested in.
2. In that row, note the “ports=” value. This is a hexadecimal number that indicates the member ports (0x11 in the example in bold above).
3. Convert that hexadecimal number to binary (10001 in the example).
4. Counting from the right, note the positions of the digit 1 in the binary number. Each 1 corresponds to a port that belongs to the target group. Therefore, in this example, ports 1 and 5 are members of the group 01-00-5e-7f-00-7f in VLAN1.

**x900, AT-9900, AT-8948** On these switches, use the following command to match ports to multicast groups (represented by the multicast MAC address):

```
show switch memory=l2multicast
```

This lists the ports for each group, as the following example shows. It is much easier to tell which groups are in each multicast group than it is with the previous switch series.

```

Manager > sh switch table=l2multicast

L2 Multicast Group and VLAN Broadcast table:
-----
                Total number of entries = 6
-----

Index  MAC                VID  MCGroup  CPU_MEM  NumPorts
      PORT_LIST
-----
0      01-00-5e-7f-ff-fa 1    4097     0        1
      Ports = 4
1      01-00-5e-7f-ff-fe 1    4099     0        2
      Ports = 2,6
2      01-00-5e-7f-00-7f 1    4102     0        2
      Ports = 2,6
3      01-00-5e-00-01-18 1    4098     0        2
      Ports = 2,6
4      01-00-5e-00-01-16 1    4096     0        1
      Ports = 4
5      03-bf-ac-10-00-7f 1    1        1        1
      Ports = 12

```