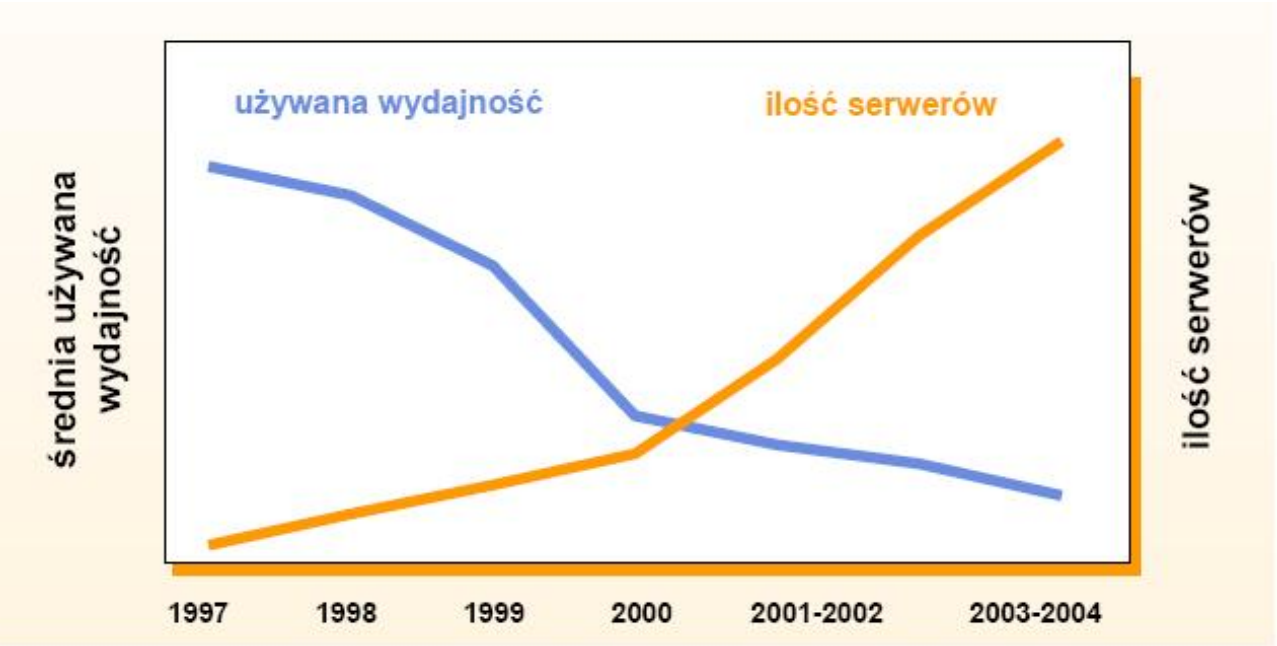
LPAR - logiczne partycjonowanie systemów

Mateusz Błażewicz Piotr Butryn Jan Sikora

MIMUW
20 grudnia 2007

Co to jest?

LPAR - logiczne partycjonowanie sprzętu
Dzielenie zasobów fizycznego serwera na niezależne, wirtualne maszyny
Wprowadzone w 1990 roku przez IBM dla mainframe'ow ESA/390



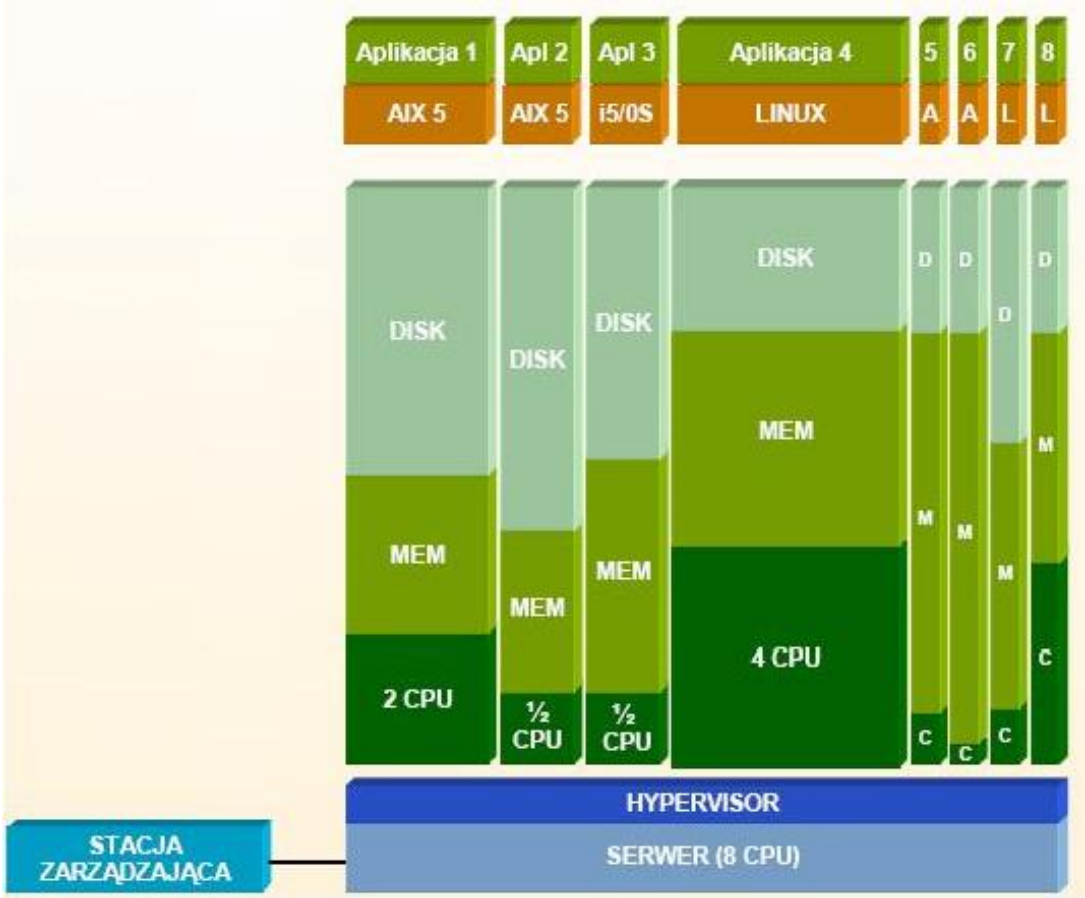
Konfiguracja

Parametry wirtualnego komputera:

- Moc i liczba procesorów
- Ilość pamięci RAM
- Fizyczne urządzenia
- Wirtualne urządzenia

Jak to działa?

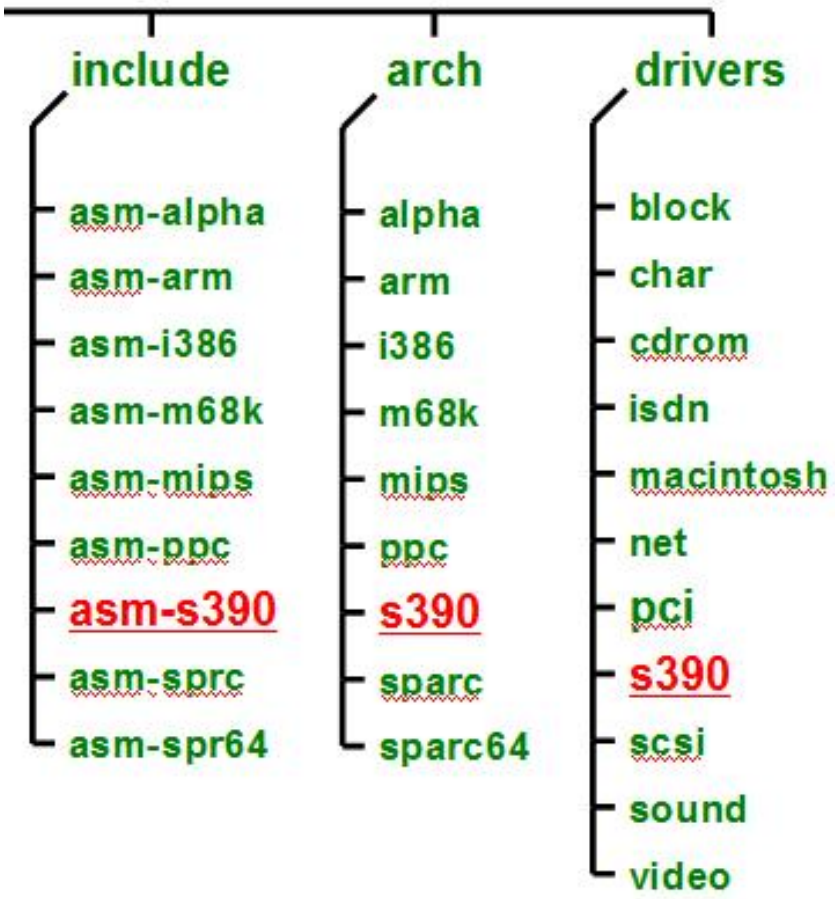
Parawirtualizacja - technika wirtualizacji w której system operacyjny-gość współpracuje ze środowiskiem operacyjnym komputera-hosta w zakresie obsługi niektórych urządzeń.



Wymagania

Wymagania względem OS

- Niewielkie zmiany w kodzie
- Sterowniki



Hypervisor

- Natywny - uruchamiany bezpośrednio na serwerze
- Hostowany - uruchamiany jako aplikacja (np. VMware Server)
- Pierwszy hypervisor: CP-40 firmy IBM (1967 rok)

Hypervisor

- Zarządca wirtualizacji, pośrednik między warstwą sprzętowa i systemową
- Umożliwia podział na logiczne partycje
- Zapewnia izolację LPARów
- Rozdziela zadania LPARu pomiędzy procesory
- Zapewnia komunikację między LPARami (Virtual SCSI, Virtual Ethernet)

DLPAR i mikro-partycjonowanie

DLPAR - dynamiczny LPAR

- Możliwość zmiany parametrów logicznej partycji w trakcie działania
- Wprowadzona w 2001 roku przez IBM dla platform opartych o procesor POWER4

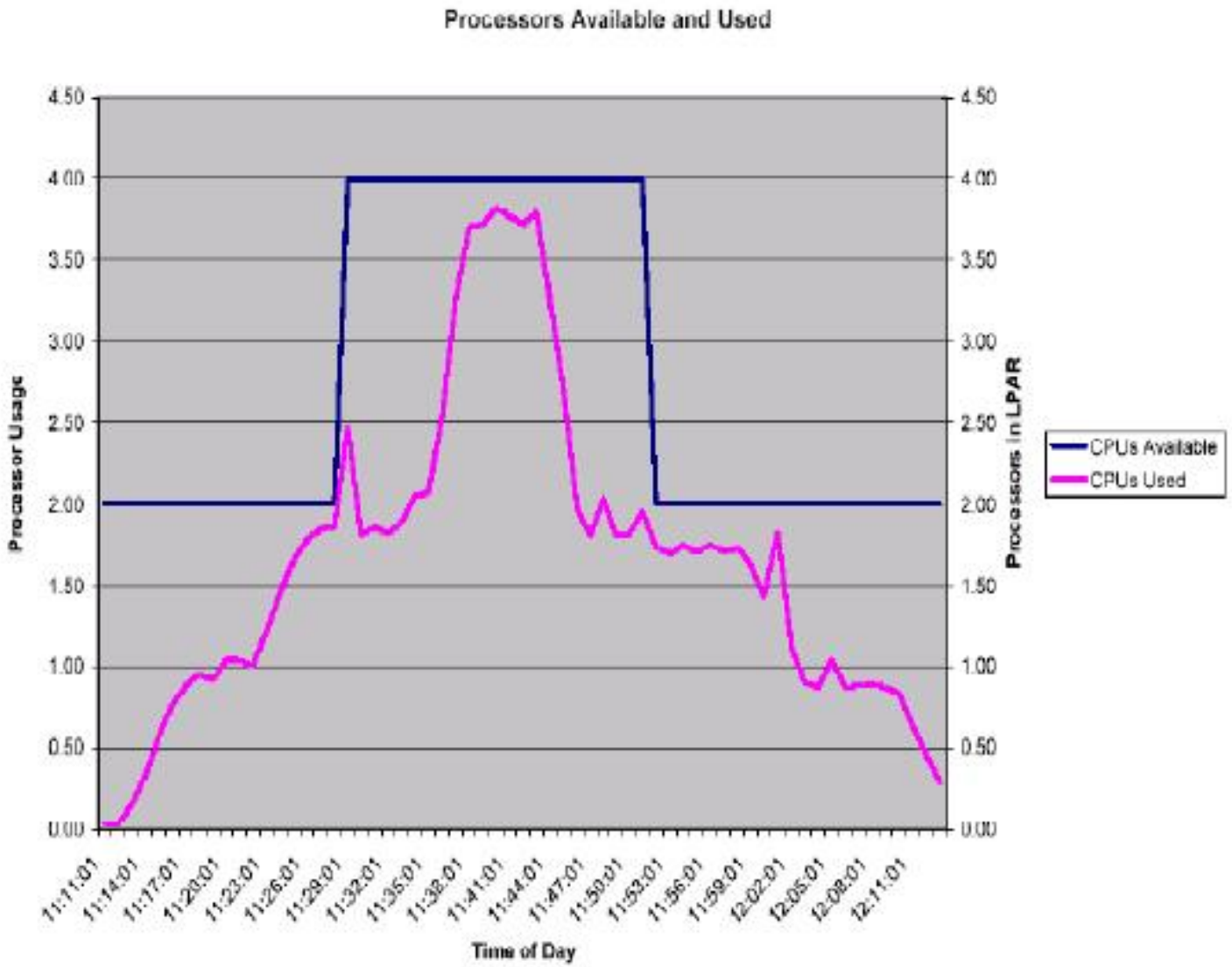
Mikro-partycjonowanie - rozdzielenie jednego procesora miedzy różne LPARy (POWER5) Maksimum 254 LPARów na jednej maszynie

DLPAR

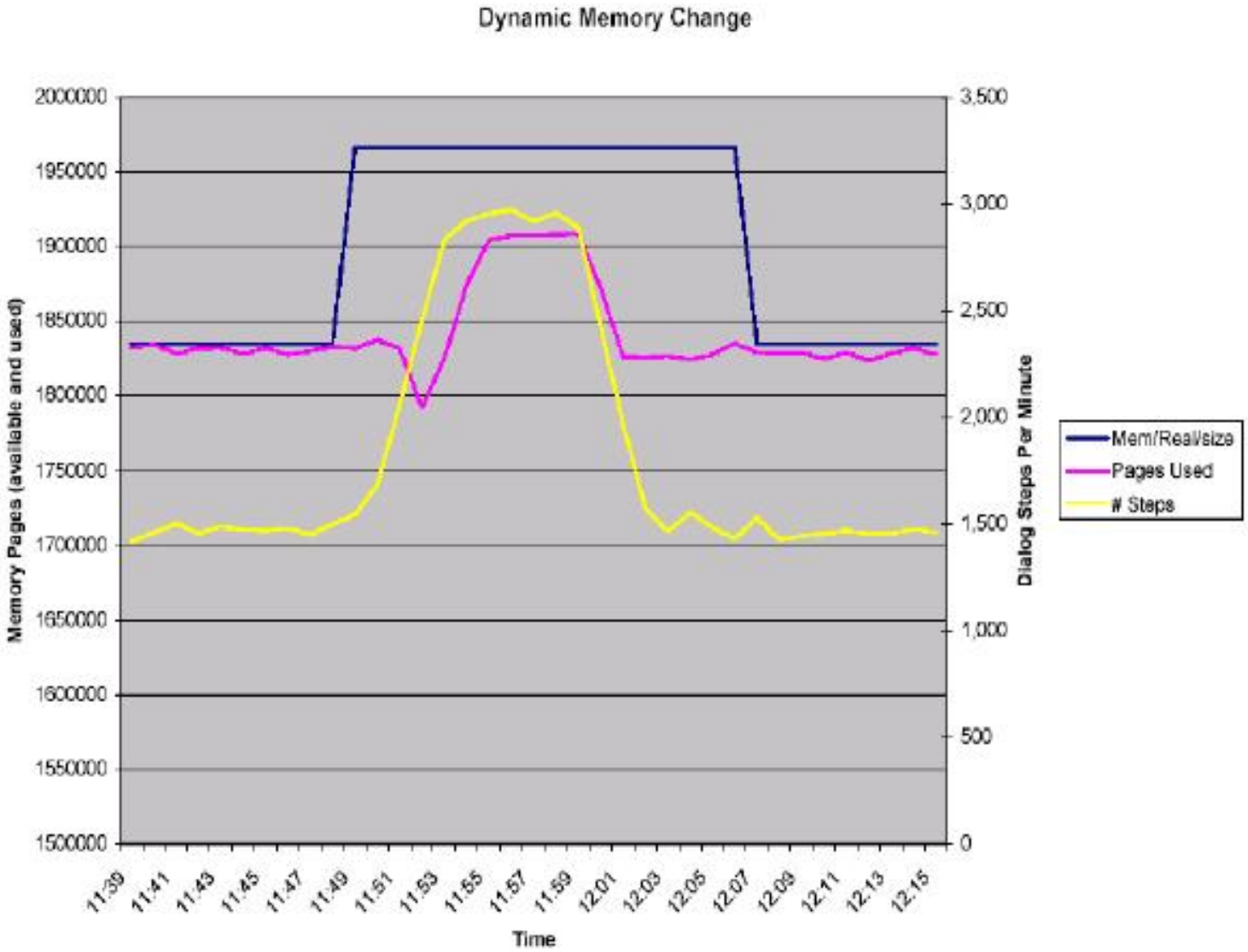
Dynamiczne operacje na LPARze:

- Zmiana mocy obliczeniowej
- Zmiana ilości pamięci RAM
- Dodanie /usunięcie urządzeń fizycznych
- Dodanie /usunięcie urządzeń wirtualnych

DLPAR


Processors Available and Used
Processor Usage
Processors in LPAR
Time of Day
CPUs Available
CPUs Used
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
11:11:01
11:14:01
11:17:01
11:20:01
11:23:01
11:26:01
11:29:01
11:32:01
11:35:01
11:38:01
11:41:01
11:44:01
11:47:01
11:50:01
11:53:01
11:56:01
11:59:01
12:02:01
12:05:01
12:08:01
12:11:01

DLPAR


Dynamic Memory Change
Memory Pages (available and used)
2000000
1950000
1900000
1850000
1800000
1750000
1700000
1650000
1600000
1550000
1500000
Dialog Steps Per Minute
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Mem/RealSize
Pages Used
Steps
11:39
11:41
11:43
11:45
11:47
11:49
11:51
11:53
11:55
11:57
11:59
12:01
12:03
12:05
12:07
12:09
12:11
12:13
12:15
Time

Wsparcie sprzętowe

- Mainframe'y oparte o procesor POWER4 i POWER5:
  - AIX
  - i5/OS
  - Linux
- Komputery zSeries
  - Linux on zSeries
- Komputery pSeries

Korzyści

- Lepsze wykorzystanie zasobów
- Oszczędność finansowa
- Bezpieczeństwo i mniejsza awaryjność
- Większa elastyczność
- Większa niezależność sprzętowa
- Dobra łączność między LPARami bez kilometrów kabli
- Łatwość tworzenia nowych instancji serwerów

Troche historii

LPAR na maszynach IBM:

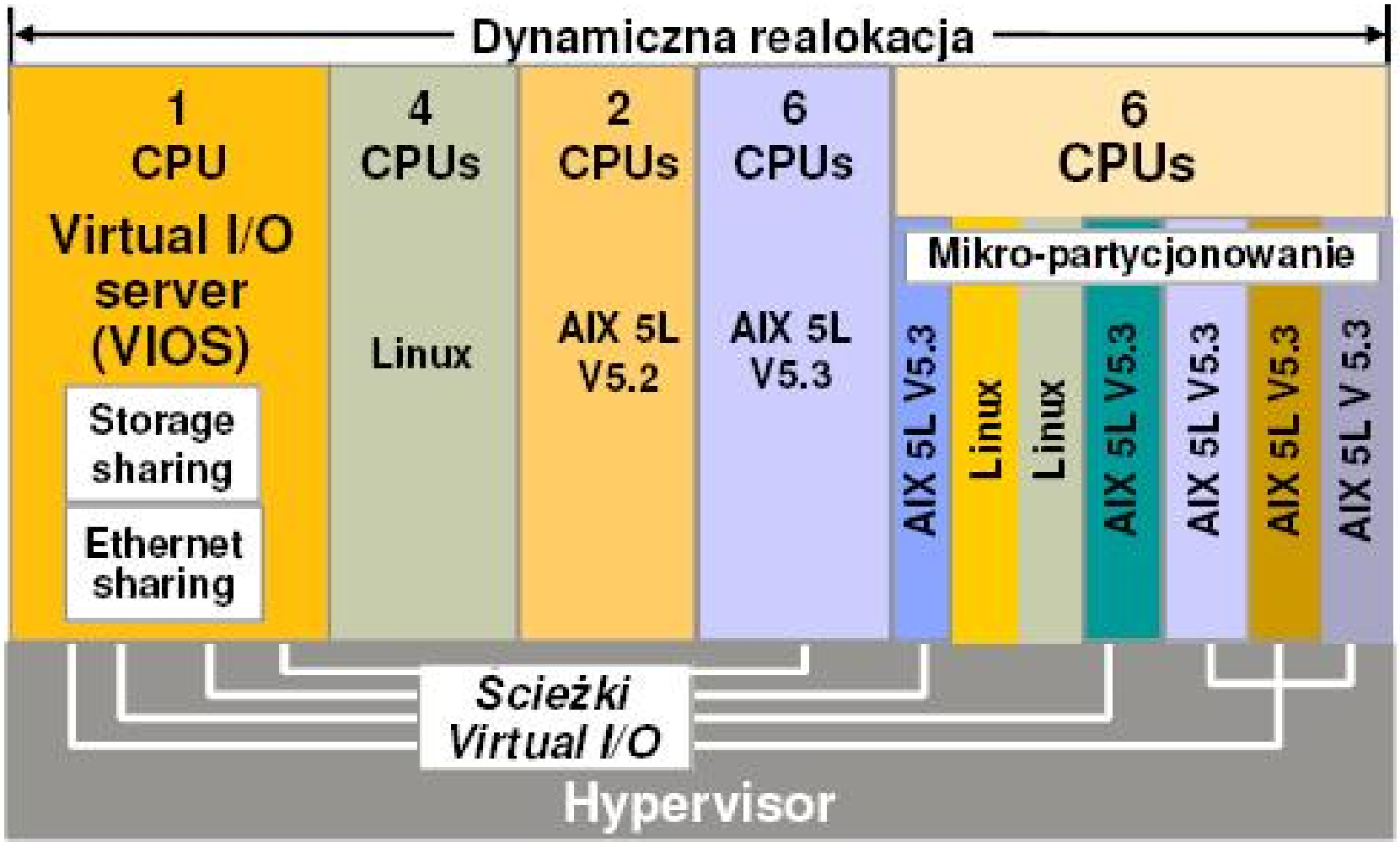
- ESA/390
- zSeries
- System 9
- iSeries
- pSeries

Zastosowane technologie

Serwery pSeries

- LPAR
- DLPAR
- CuOD - Dynamic Capacity on Demand
- Dynamic CPU Guard
- Virtual I/O
- Virtual Ethernet

Budowa


Dynamiczna realokacja
1 CPU
Virtual I/O server (VIOS)
Storage sharing
Ethernet sharing
4 CPUs
Linux
2 CPUs
AIX 5L V5.2
6 CPUs
AIX 5L V5.3
6 CPUs
Mikro-partycjonowanie
AIX 5L V5.3
Linux
Linux
AIX 5L V5.3
AIX 5L V5.3
AIX 5L V5.3
Ścieżki Virtual I/O
Hypervisor

Hypervisor

Hardware Management Console (HMC)

- dostęp do konsoli bezpośredni, lub zdalny: http://adres_konsoli/remote_client.html
- dane o konfiguracji LPAR-ów przechowywane na serwerze a nie konsoli
- backup danych o konfiguracji LPARów wykonywany za pośrednictwem konsoli HMC
- awaria konsoli nie ma wpływu na działanie serwera

Hypervisor

Integrated Virtualization Manager IVM

- alternatywa dla HMC do stosowania w "mniejszych" serwerach pSeries
- nie wymaga dedykowanej konsoli, instalowany wraz z oprogramowaniem Virtual I/O Serwera
- posiada interfejs WWW (https) + wiersz poleceń

Hypervisor

Różnice pomiędzy IVM a HMC

- brak możliwości tworzenie wielu profilów na LPAR
- brak pełnego wsparcia dla CoD
- brak możliwości zarządzania wieloma serwerami z jednego punktu
- brak możliwości przydziału fizycznych urządzeń do LPAR-a (wszystkie LPAR-y w pełni zwirtualizowane)
- nie można stosować jednocześnie konsoli HMC i IVM

Możliwość konfiguracji

Procesory

- ilość procesorów widziana przez system
- moc procesorów (minimalny kwant 1/10cpu)
- możliwość automatycznego dobierania mocy w miarę zapotrzebowania (procesory typu capped/uncapped)
- granice w jakich można dynamicznie zmieniać parametry procesorów dla danego LPAR-a

Możliwość konfiguracji

Procesory

- ilość procesorów widziana przez system
- moc procesorów (minimalny kwant 1/10cpu)
- możliwość automatycznego dobierania mocy w miarę zapotrzebowania (procesory typu capped/uncapped)
- granice w jakich można dynamicznie zmieniać parametry procesorów dla danego LPAR-a

Pamięć

- ilość pamięci przydzielonej LPAR-owi
- granice w jakich można dynamicznie przydzielać/odbierać pamięć

Możliwość konfiguracji

Procesory

- ilość procesorów widziana przez system
- moc procesorów (minimalny kwant 1/10cpu)
- możliwość automatycznego dobierania mocy w miarę zapotrzebowania (procesory typu capped/uncapped)
- granice w jakich można dynamicznie zmieniać parametry procesorów dla danego LPAR-a

Pamięć

- ilość pamięci przydzielonej LPAR-owi
- granice w jakich można dynamicznie przydzielać/odbierać pamięć

Fizyczne urządzenia przydzielone LPAR-owi (wszelkie karty rozszerzeń)

Możliwość konfiguracji

Procesory

- ilość procesorów widziana przez system
- moc procesorów (minimalny kwant 1/10cpu)
- możliwość automatycznego dobierania mocy w miarę zapotrzebowania (procesory typu capped/uncapped)
- granice w jakich można dynamicznie zmieniać parametry procesorów dla danego LPAR-a

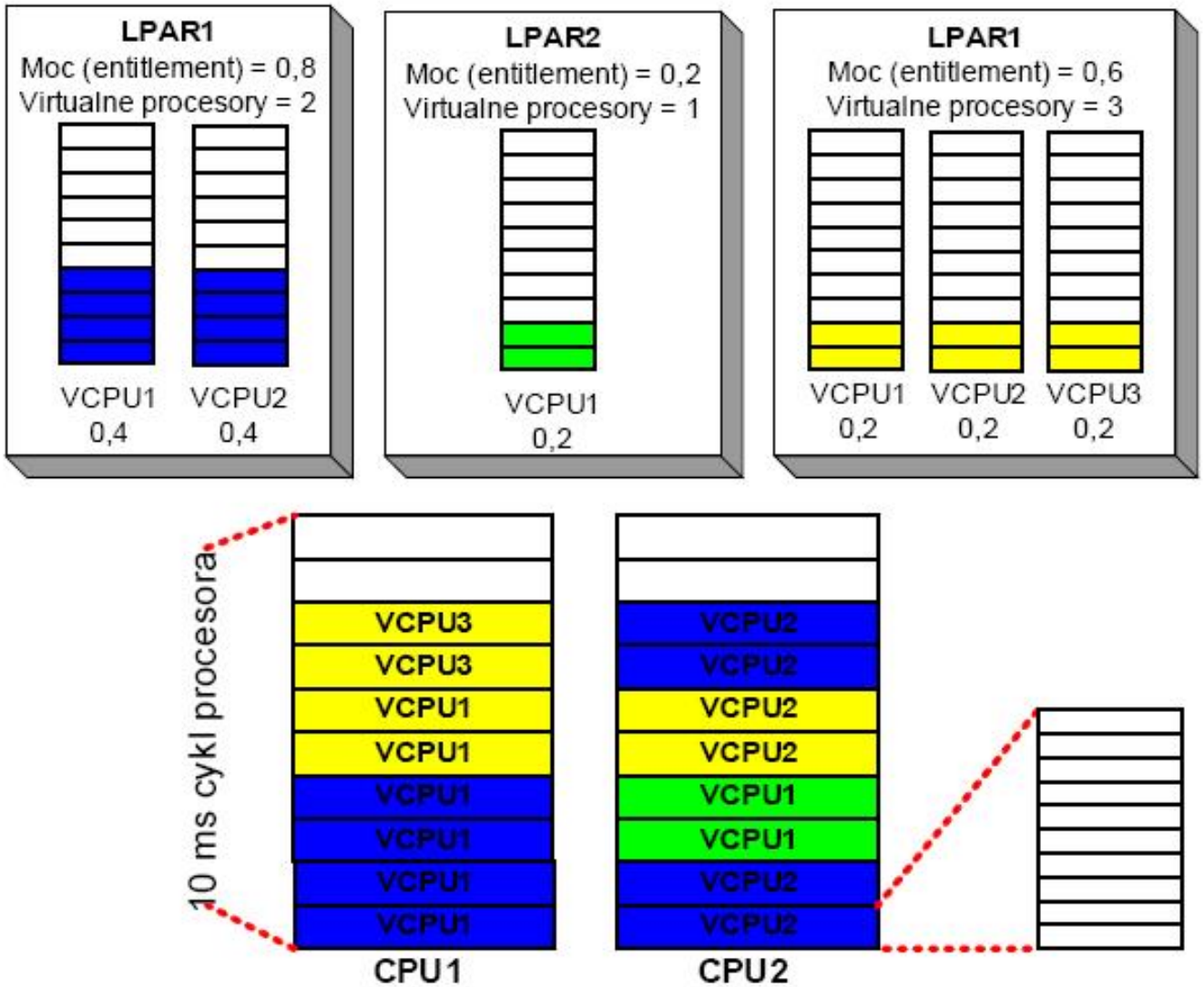
Pamięć

- ilość pamięci przydzielonej LPAR-owi
- granice w jakich można dynamicznie przydzielać/odbierać pamięć

Fizyczne urządzenia przydzielone LPAR-owi (wszelkie karty rozszerzeń)

Wirtualne urządzenia przydzielone LPAR-owi (Virtual SCSI/Network)

Przydział wirtualnych procesorów



Zmiany dynamiczne

Procesory

- dołożenie/odebranie mocy procesora (min - max)
- dołożenie/odebranie wirtualnych procesorów (min - max)
- zmiana trybu pracy procesora (capped/uncapped)

Zmiany dynamiczne

Procesory

- dołożenie/odebranie mocy procesora (min - max)
- dołożenie/odebranie wirtualnych procesorów (min - max)
- zmiana trybu pracy procesora (capped/uncapped)

Pamięć

- dołożenie/odebranie pamięci (min - max)

Zmiany dynamiczne

Procesory

- dołożenie/odebranie mocy procesora (min - max)
- dołożenie/odebranie wirtualnych procesorów (min - max)
- zmiana trybu pracy procesora (capped/uncapped)

Pamięć

- dołożenie/odebranie pamięci (min - max)

Urządzenia fizyczne

- dołożenie/odebranie dowolnych adapterów fizycznych

Zmiany dynamiczne

Procesory

- dołożenie/odebranie mocy procesora (min - max)
- dołożenie/odebranie wirtualnych procesorów (min - max)
- zmiana trybu pracy procesora (capped/uncapped)

Pamięć

- dołożenie/odebranie pamięci (min - max)

Urządzenia fizyczne

- dołożenie/odebranie dowolnych adapterów fizycznych

Urządzenia wirtualne

- dołożenie/odebranie wirtualnych kontrolerów SCSI
- dołożenie/odebranie przestrzeni dyskowej udostępnianej przez VIOS
- dołożenie/odebranie wirtualnego adaptera ethernet

Partition Load Manager

VIOS

Virtual I/O Serwer
Oprogramowanie umieszczone na dedykowanym LPAR-rze umożliwiające wirtualizację następujących zasobów:

- przestrzeń dyskowa
- dostęp do sieci Ethernet

VIOS

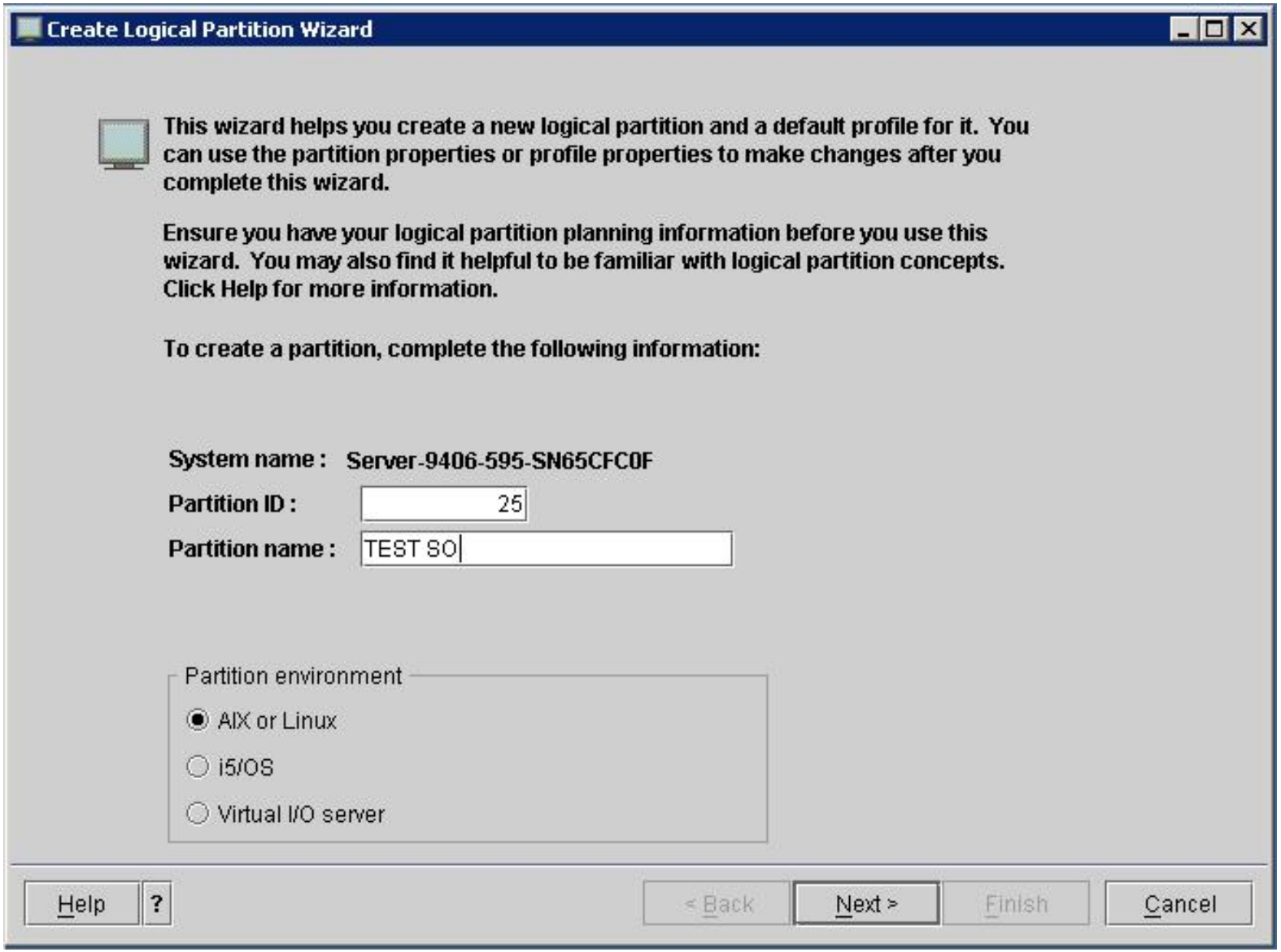
- wirtualizowanie wewnętrznych dysków SCSI
- wirtualizowanie logicznych woluminów dowolnej wielkości utworzonych na dyskach wewnętrznych
- wirtualizowanie zewnętrznych dysków SCSI (SAN)
- wirtualizowanie logicznych woluminów dowolnej wielkości utworzonych na dyskach zewnętrznych
- możliwość dynamicznego dodawania/usuwania urządzeń dyskowych do/z LPAR-a
- niewielki spadek prędkości obsługi urządzeń dyskowych - pojedyncze procenty
- niewielkie zaangażowanie mocy procesora w operacje dyskowe

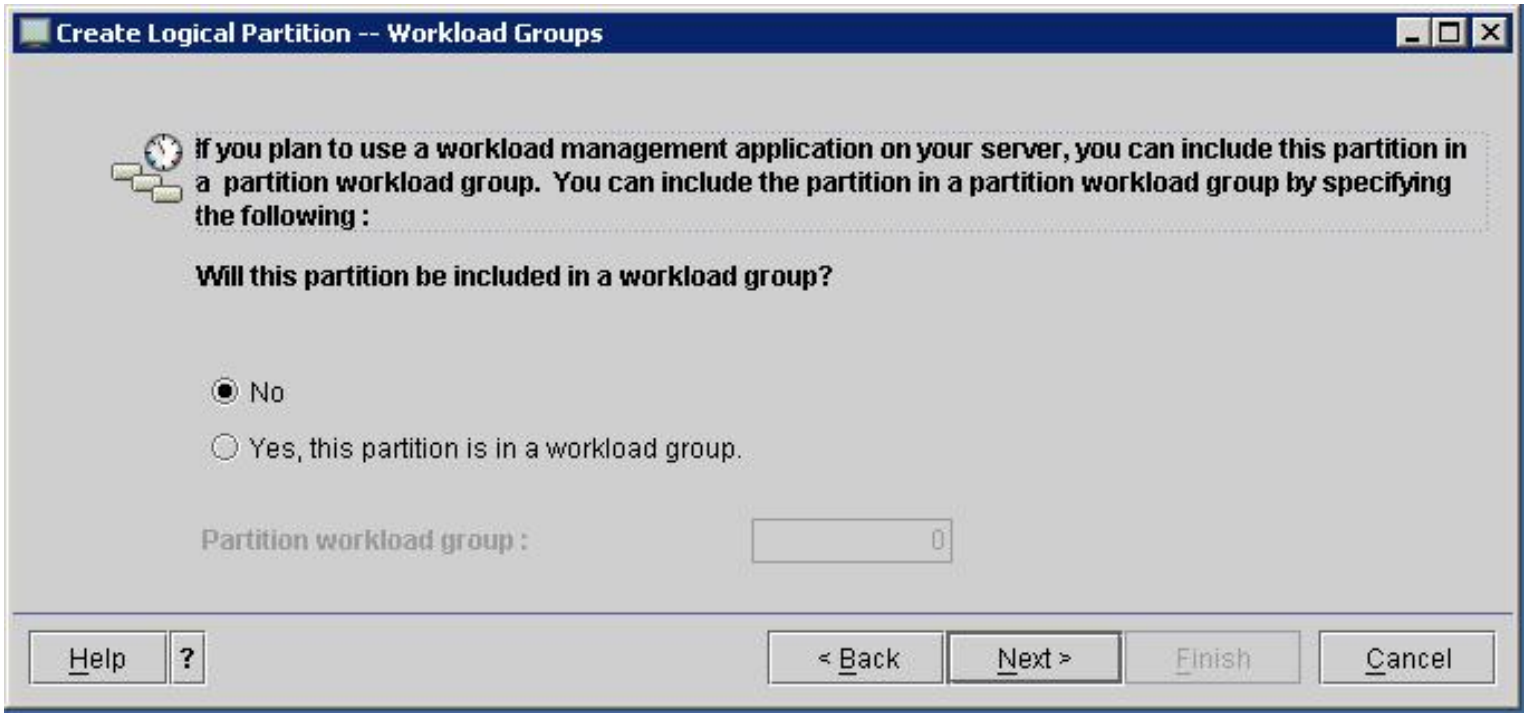
VIOS

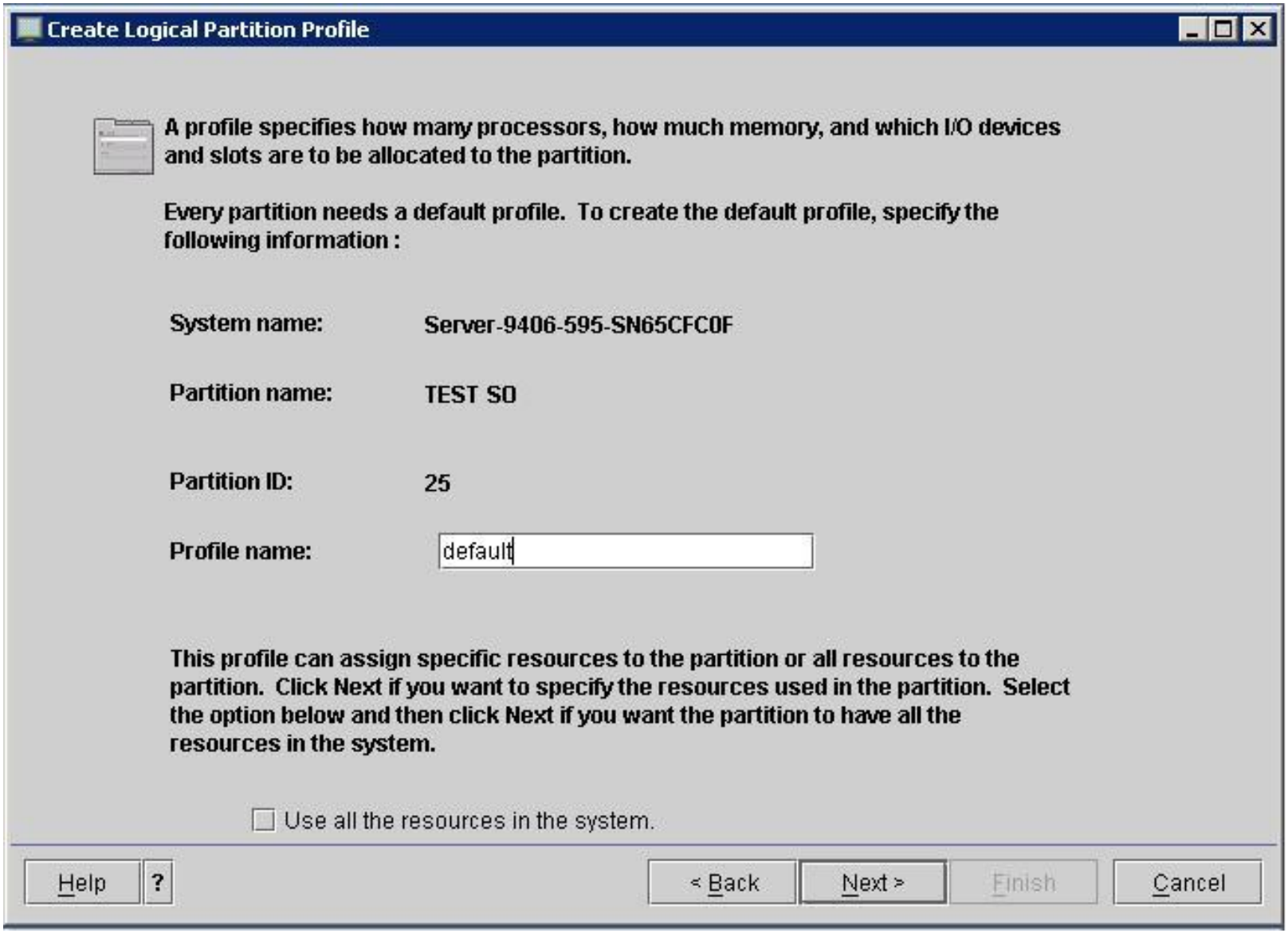
- umożliwia dostęp do sieci zewnętrznej poprzez wirtualne adaptery Ethernet
- stosunkowo duże zużycie mocy CPU na wirtualizację transferów sieciowych (w pełni nasycony interface 1GB = 1CPU "Power 5" o mocy 1,65 GHz)
- systemy w obrębie jednego serwera komunikują się poprzez wirtualne adaptery bez pośrednictwa VIOS-a
- nie należy wirtualizować systemów generujących dużą ilość ruchu sieciowego

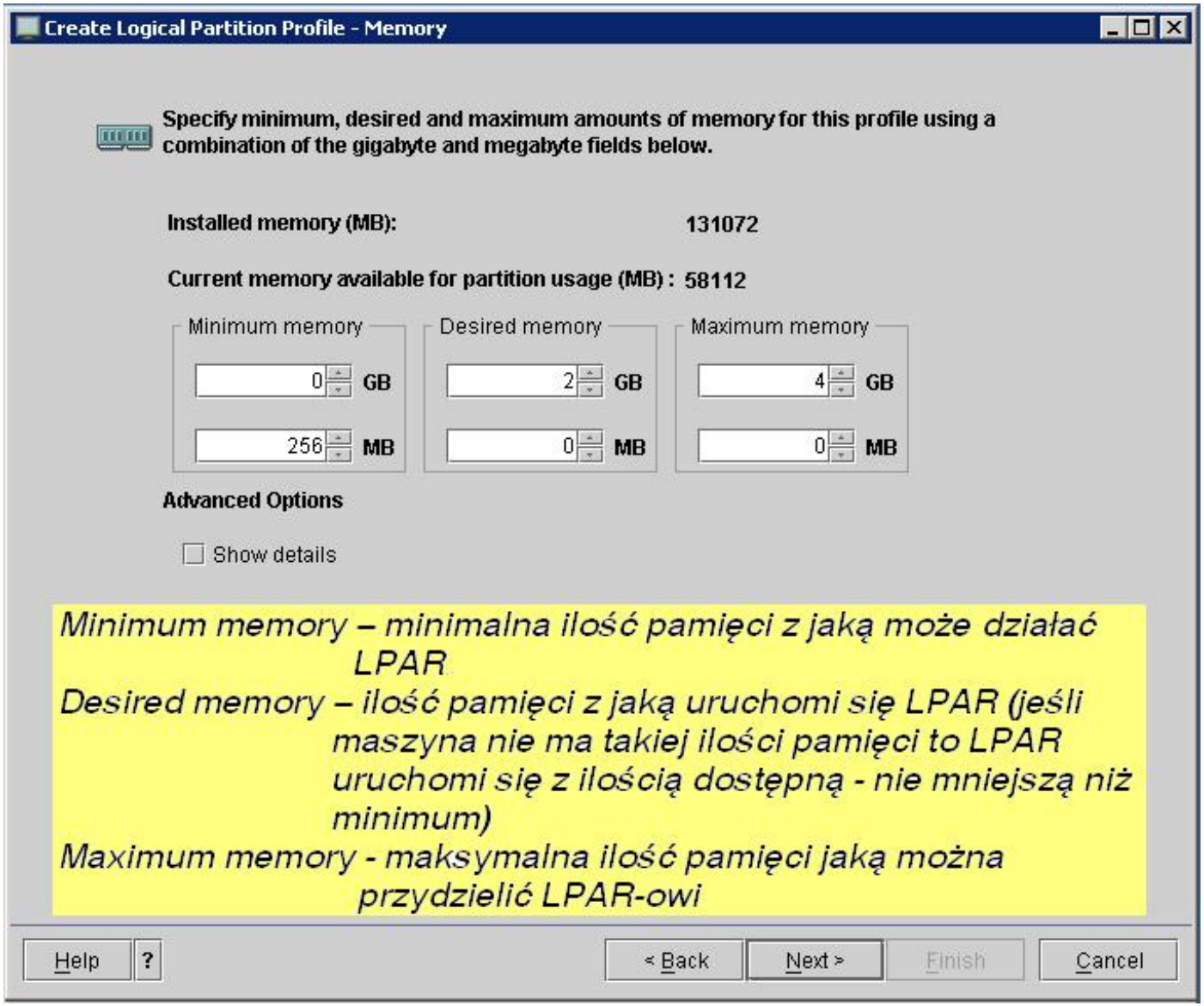
Ograniczenia

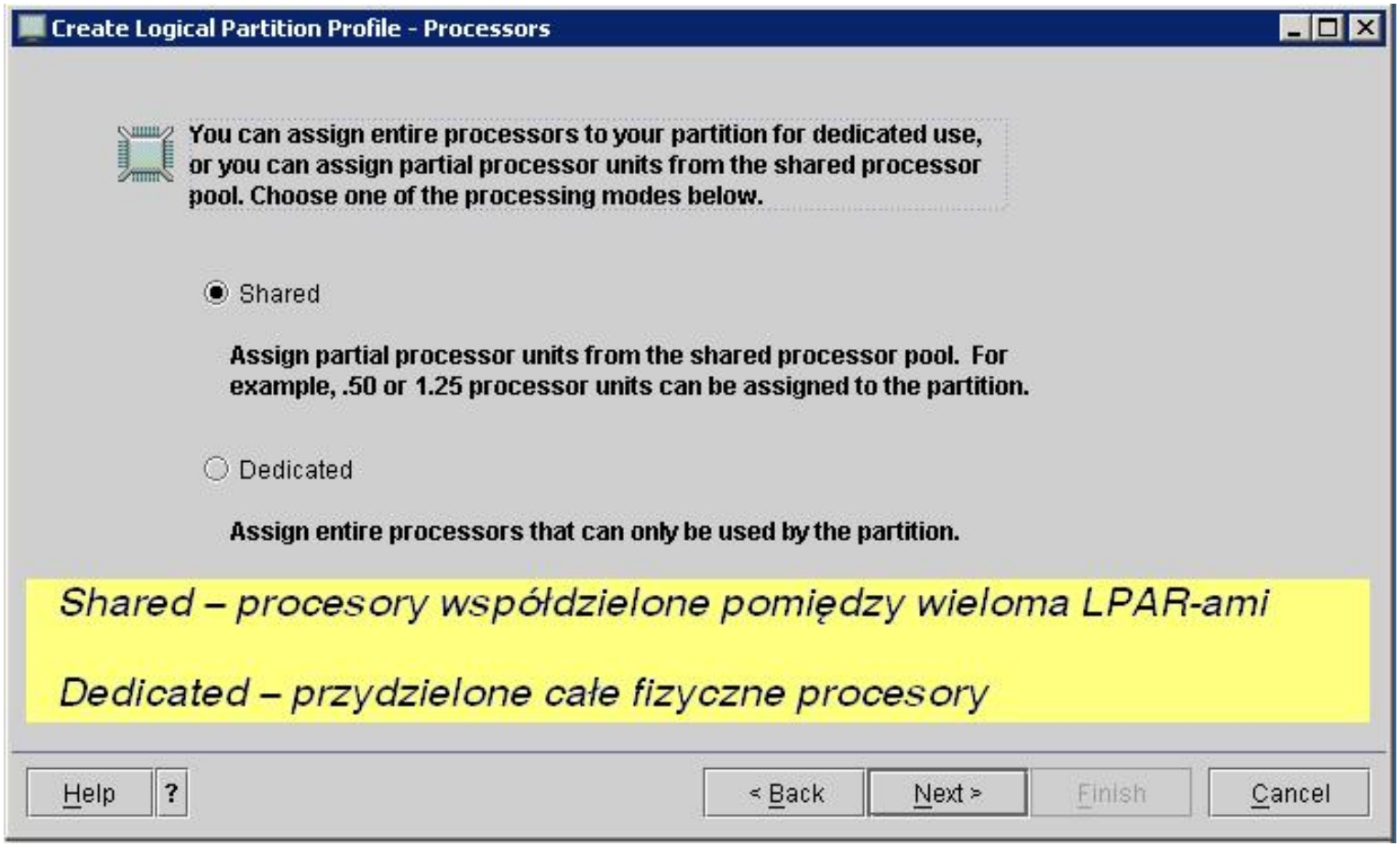
- maksymalna ilość LPAR-ów na serwer = 254
- maksymalna ilość wirtualnych procesorów na LPAR = 64
- minimalna ilość procesora na LPAR = 0,1
- maksymalna ilość mocy procesora na wirtualny procesor = 1
- nie można mieszać na LPAR-rze procesorów dedykowanych i współdzielonych

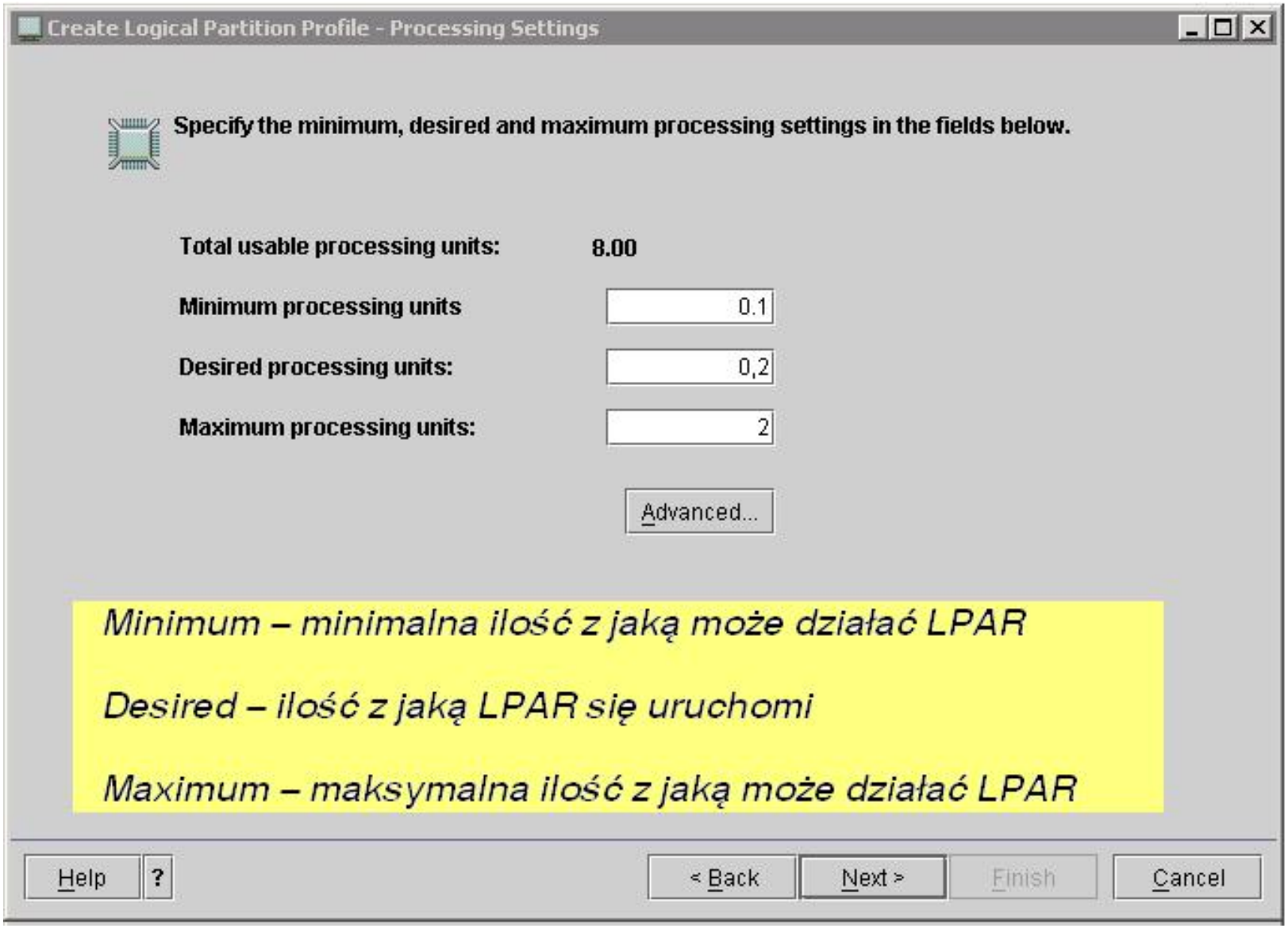

Create Logical Partition Wizard
This wizard helps you create a new logical partition and a default profile for it. You can use the partition properties or profile properties to make changes after you complete this wizard.
Ensure you have your logical partition planning information before you use this wizard. You may also find it helpful to be familiar with logical partition concepts. Click Help for more information.
To create a partition, complete the following information:
System name : Server-9406-595-SN65CFC0F
Partition ID : 25
Partition name : TEST SO
Partition environment
AIX or Linux
i5/OS
Virtual I/O server
Help
?
< Back
Next >
Finish
Cancel

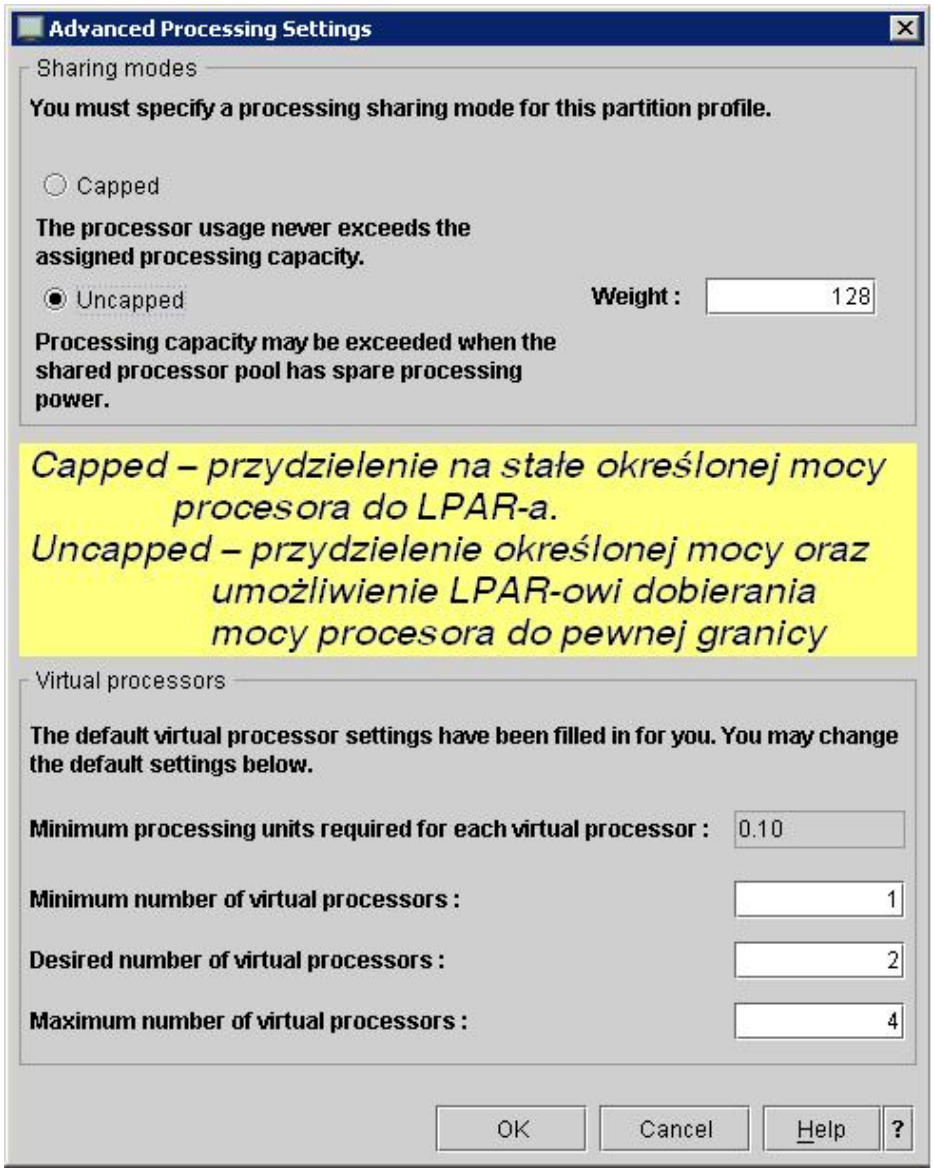

Create Logical Partition -- Workload Groups
If you plan to use a workload management application on your server, you can include this partition in a partition workload group. You can include the partition in a partition workload group by specifying the following :
Will this partition be included in a workload group?
No
Yes, this partition is in a workload group.
Partition workload group :
0
Help
?
< Back
Next >
Finish
Cancel

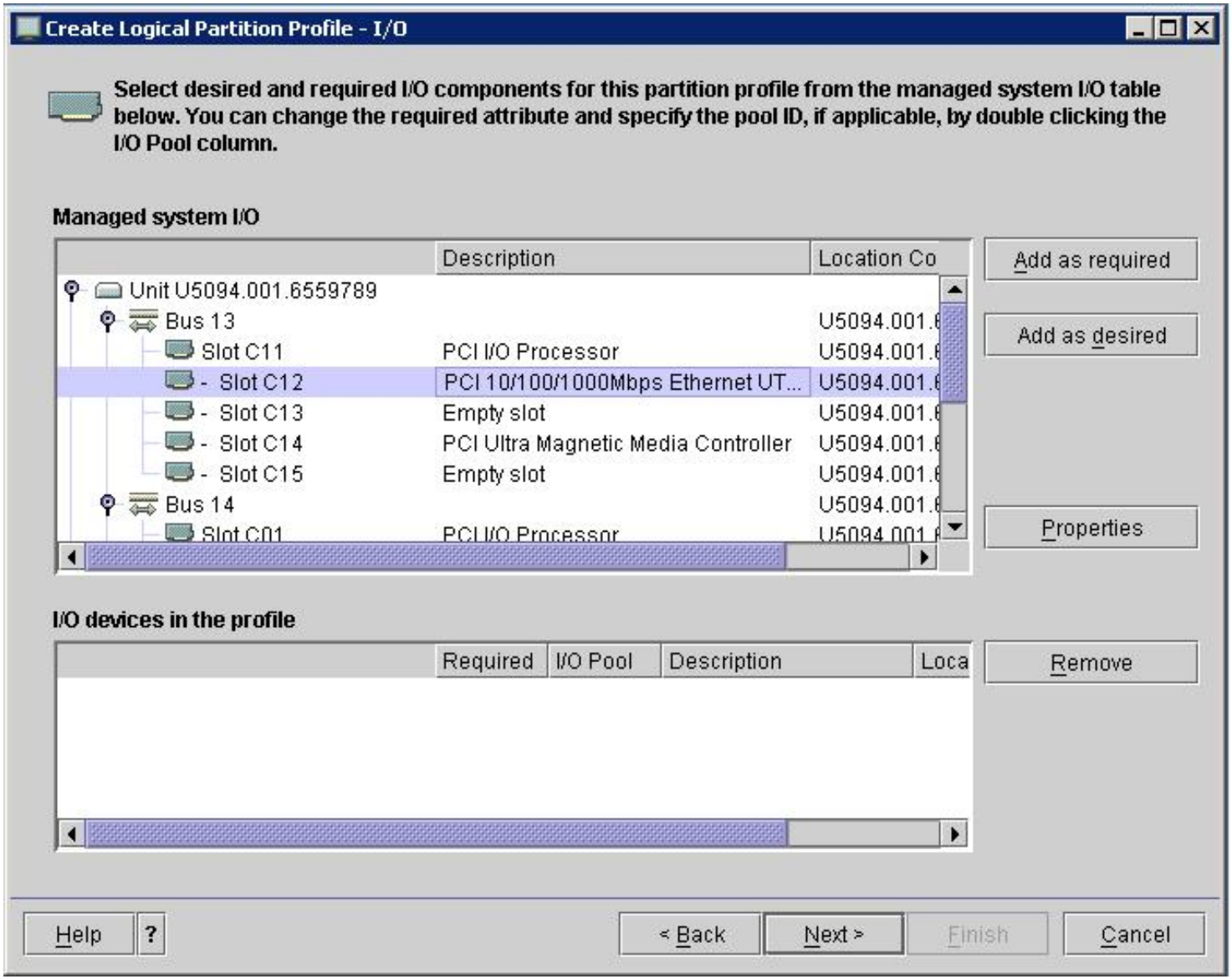

Create Logical Partition Profile
A profile specifies how many processors, how much memory, and which I/O devices and slots are to be allocated to the partition.
Every partition needs a default profile. To create the default profile, specify the following information :
System name:
Server-9406-595-SN65CFC0F
Partition name:
TEST S0
Partition ID:
25
Profile name:
default
This profile can assign specific resources to the partition or all resources to the partition. Click Next if you want to specify the resources used in the partition. Select the option below and then click Next if you want the partition to have all the resources in the system.
Use all the resources in the system.
Help
?
< Back
Next >
Finish
Cancel

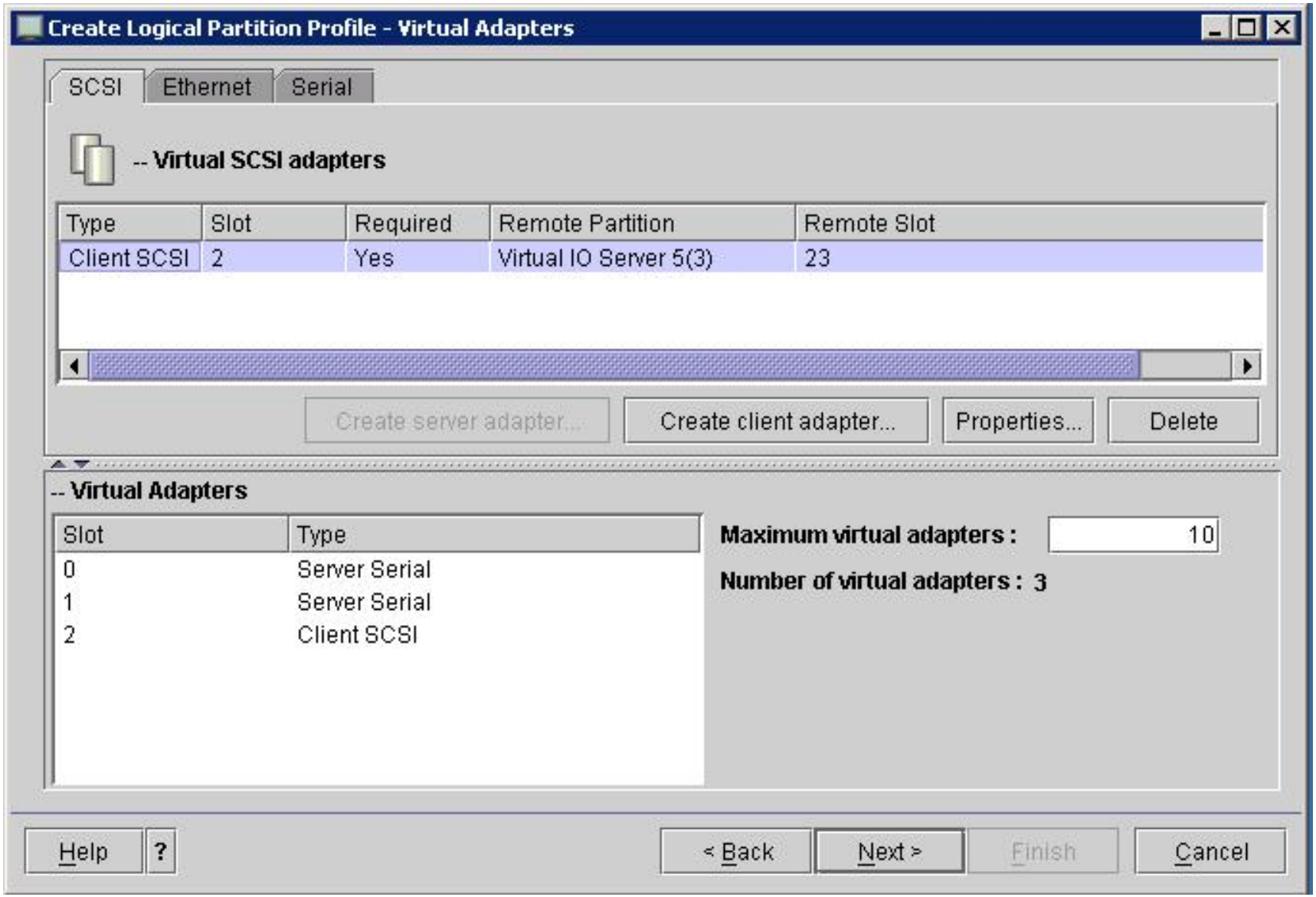

Create Logical Partition Profile - Memory
Specify minimum, desired and maximum amounts of memory for this profile using a combination of the gigabyte and megabyte fields below.
Installed memory (MB): 131072
Current memory available for partition usage (MB) : 58112
Minimum memory
0 GB
256 MB
Desired memory
2 GB
0 MB
Maximum memory
4 GB
0 MB
Advanced Options
Show details
Minimum memory – minimalna ilość pamięci z jaką może działać LPAR
Desired memory – ilość pamięci z jaką uruchomi się LPAR (jeśli maszyna nie ma takiej ilości pamięci to LPAR uruchomi się z ilością dostępną - nie mniejszą niż minimum)
Maximum memory - maksymalna ilość pamięci jaką można przydzielić LPAR-owi
Help
?
< Back
Next >
Finish
Cancel

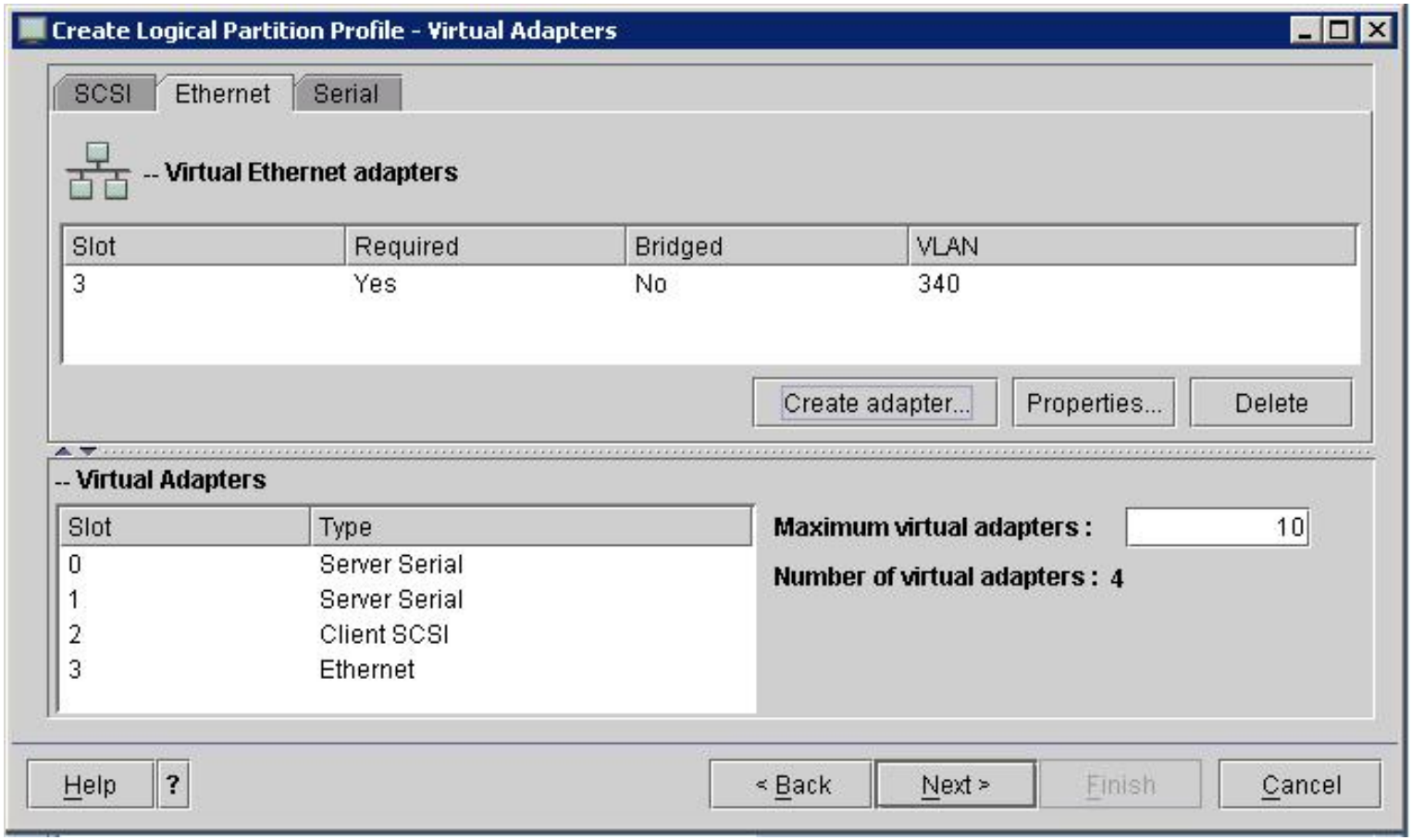

Create Logical Partition Profile - Processors
You can assign entire processors to your partition for dedicated use, or you can assign partial processor units from the shared processor pool. Choose one of the processing modes below.
Shared
Assign partial processor units from the shared processor pool. For example, .50 or 1.25 processor units can be assigned to the partition.
Dedicated
Assign entire processors that can only be used by the partition.
Shared – procesory współdzielone pomiędzy wieloma LPAR-ami
Dedicated – przydzielone całe fizyczne procesory
Help
?
< Back
Next >
Finish
Cancel

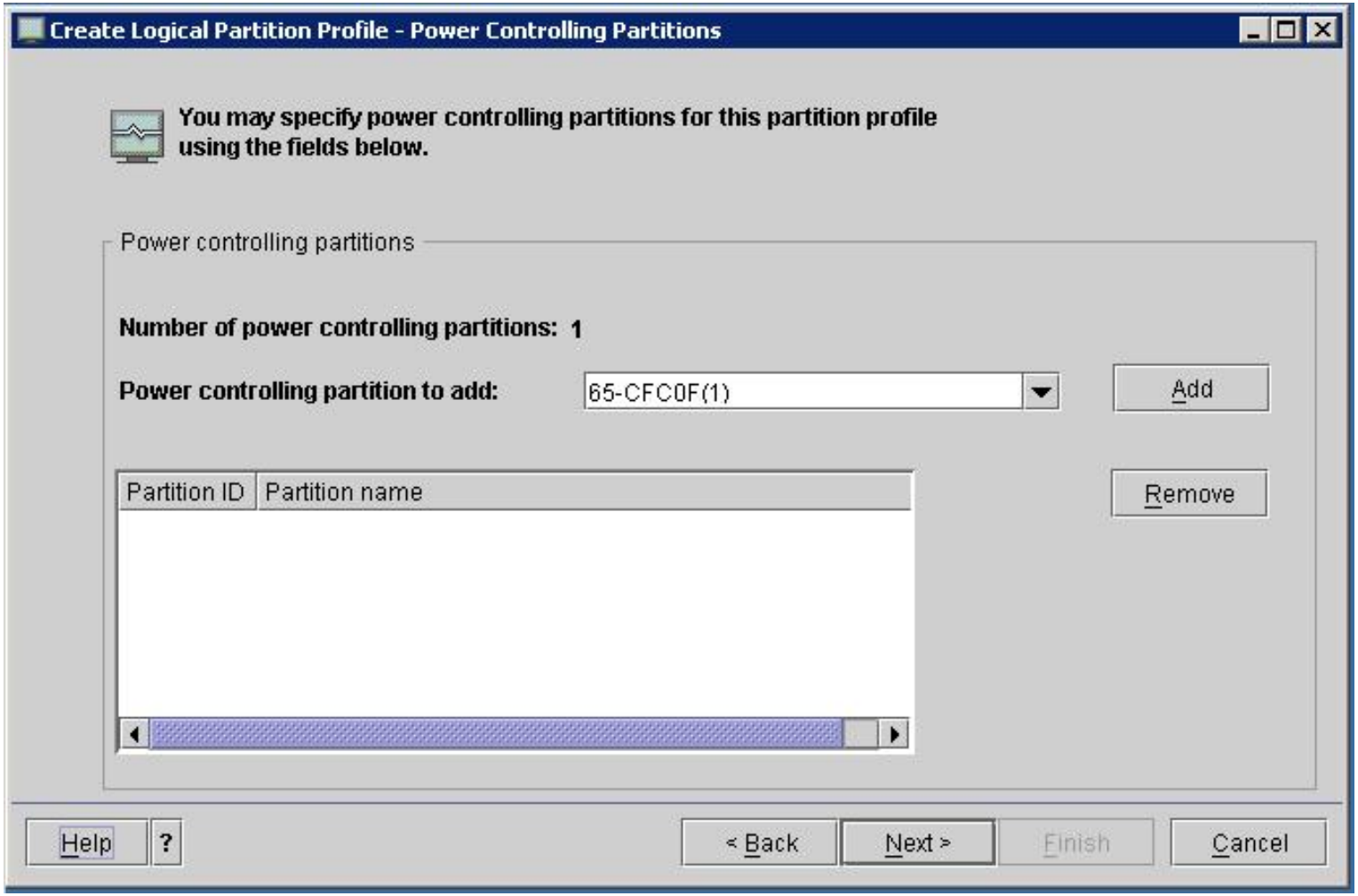

Create Logical Partition Profile - Processing Settings
Specify the minimum, desired and maximum processing settings in the fields below.
Total usable processing units:
8.00
Minimum processing units
0.1
Desired processing units:
0,2
Maximum processing units:
2
Advanced...
Minimum – minimalna ilość z jaką może działać LPAR
Desired – ilość z jaką LPAR się uruchomi
Maximum – maksymalna ilość z jaką może działać LPAR
Help
?
< Back
Next >
Finish
Cancel

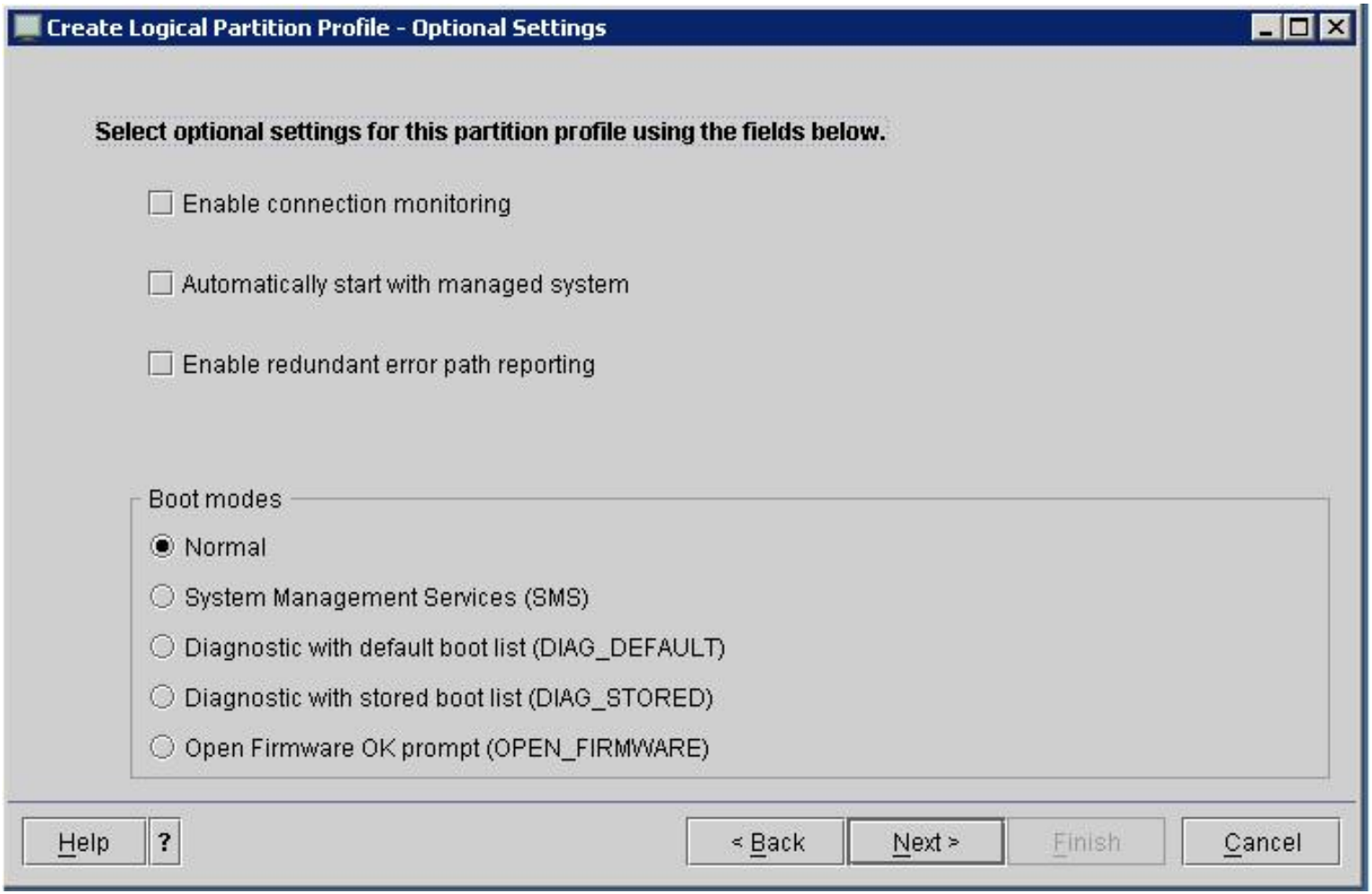

Advanced Processing Settings
Sharing modes
You must specify a processing sharing mode for this partition profile.
Capped
The processor usage never exceeds the assigned processing capacity.
Uncapped
Weight :
128
Processing capacity may be exceeded when the shared processor pool has spare processing power.
Capped – przydzielenie na stałe określonej mocy procesora do LPAR-a.
Uncapped – przydzielenie określonej mocy oraz umożliwienie LPAR-owi dobierania mocy procesora do pewnej granicy
Virtual processors
The default virtual processor settings have been filled in for you. You may change the default settings below.
Minimum processing units required for each virtual processor :
0.10
Minimum number of virtual processors :
1
Desired number of virtual processors :
2
Maximum number of virtual processors :
4
OK
Cancel
Help
?

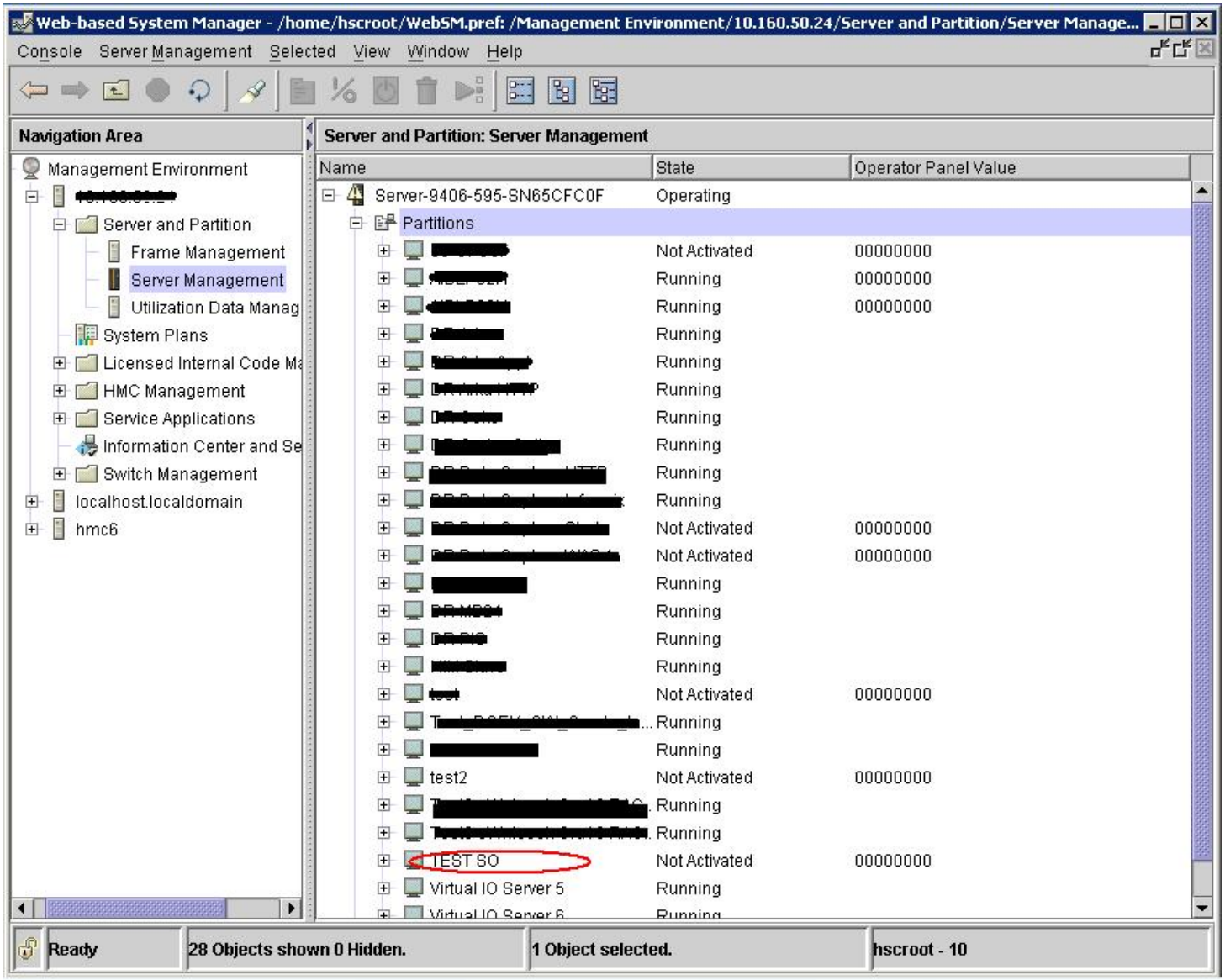

Create Logical Partition Profile - I/O
Select desired and required I/O components for this partition profile from the managed system I/O table below. You can change the required attribute and specify the pool ID, if applicable, by double clicking the I/O Pool column.
Managed system I/O
Description
Location Co
Unit U5094.001.6559789
Bus 13
Slot C11
PCI I/O Processor
Slot C12
PCI 10/100/1000Mbps Ethernet UT...
Slot C13
Empty slot
Slot C14
PCI Ultra Magnetic Media Controller
Slot C15
Empty slot
Bus 14
Add as required
Add as desired
Properties
I/O devices in the profile
Required
I/O Pool
Description
Loca
Remove
Help
?
< Back
Next >
Finish
Cancel


Create Logical Partition Profile - Virtual Adapters
SCSI
Ethernet
Serial
-- Virtual SCSI adapters
Type
Slot
Required
Remote Partition
Remote Slot
Client SCSI
2
Yes
Virtual IO Server 5(3)
23
Create server adapter...
Create client adapter...
Properties...
Delete
-- Virtual Adapters
Slot
Type
0
Server Serial
1
Server Serial
2
Client SCSI
Maximum virtual adapters :
10
Number of virtual adapters : 3
Help
?
< Back
Next >
Finish
Cancel


Create Logical Partition Profile - Virtual Adapters
SCSI
Ethernet
Serial
-- Virtual Ethernet adapters
Slot
Required
Bridged
VLAN
3
Yes
No
340
Create adapter...
Properties...
Delete
-- Virtual Adapters
Slot
Type
0
Server Serial
1
Server Serial
2
Client SCSI
3
Ethernet
Maximum virtual adapters :
10
Number of virtual adapters : 4
Help
?
< Back
Next >
Finish
Cancel


Create Logical Partition Profile - Power Controlling Partitions
You may specify power controlling partitions for this partition profile using the fields below.
Power controlling partitions
Number of power controlling partitions: 1
Power controlling partition to add:
65-CFC0F(1)
Add
Partition ID
Partition name
Remove
Help
?
< Back
Next >
Finish
Cancel

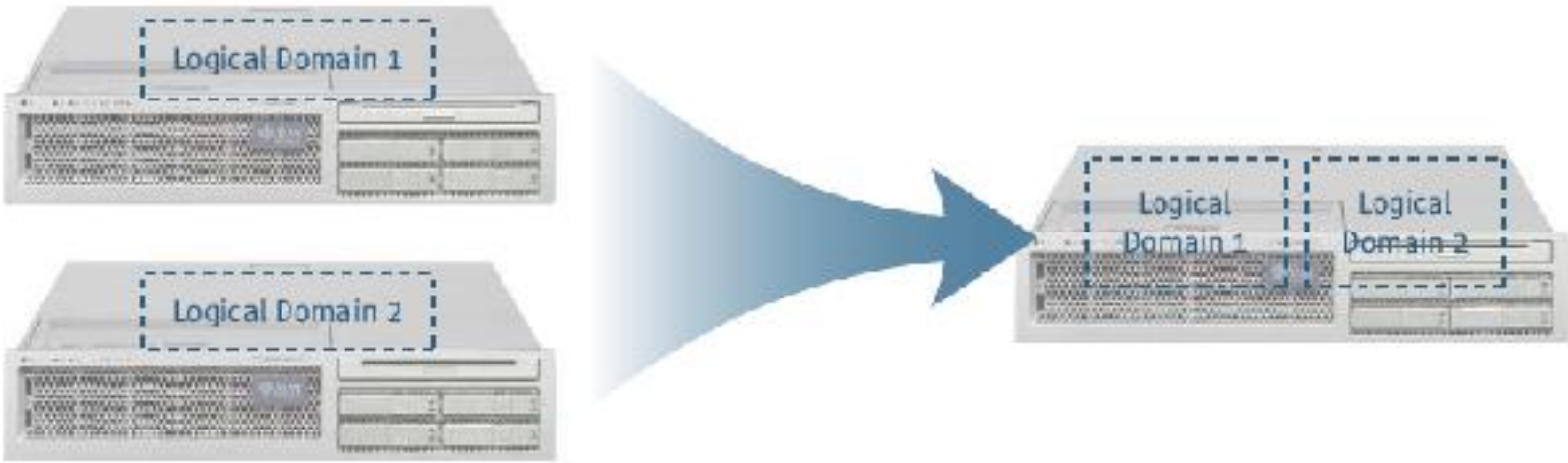

Create Logical Partition Profile - Optional Settings
Select optional settings for this partition profile using the fields below.
Enable connection monitoring
Automatically start with managed system
Enable redundant error path reporting
Boot modes
Normal
System Management Services (SMS)
Diagnostic with default boot list (DIAG_DEFAULT)
Diagnostic with stored boot list (DIAG_STORED)
Open Firmware OK prompt (OPEN_FIRMWARE)
Help
?
< Back
Next >
Finish
Cancel


Web-based System Manager - /home/hscroot/WebSM.pref: /Management Environment/10.160.50.24/Server and Partition/Server Manage...
Console Server Management Selected View Window Help
Navigation Area
Management Environment
Server and Partition
Frame Management
Server Management
Utilization Data Manag
System Plans
Licensed Internal Code Ma
HMC Management
Service Applications
Information Center and Se
Switch Management
localhost.localdomain
hmc6
Server and Partition: Server Management
Name
State
Operator Panel Value
Server-9406-595-SN65CFC0F
Operating
Partitions
Not Activated
Running
test
test2
TEST SO
Virtual IO Server 5
Ready
28 Objects shown 0 Hidden.
1 Object selected.
hscroot - 10

Co to jest LDOM?

Nowoczesna technologia partycjonowania
Rozwijana przez SUN MICROSYSTEMS
Partycje (domeny)

- fizycznie w jednej maszynie
- logicznie – niezależne jednostki



Technologia

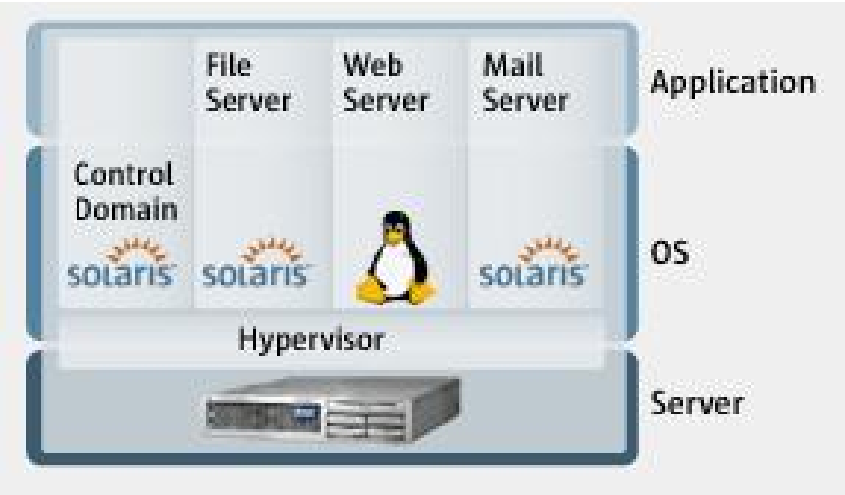
Nowatorska konstrukcja

- Każda partycja jest kompletną wirtualną maszyną ze swoim OS
- Partycje z podziałem na role
- Stabilny HyperViser

Rekonfigurowalność dynamiczna

Efekty

- większa wydajność
- mniejsze koszty
- kompleksowość
- bezpieczeńswo



Serwery

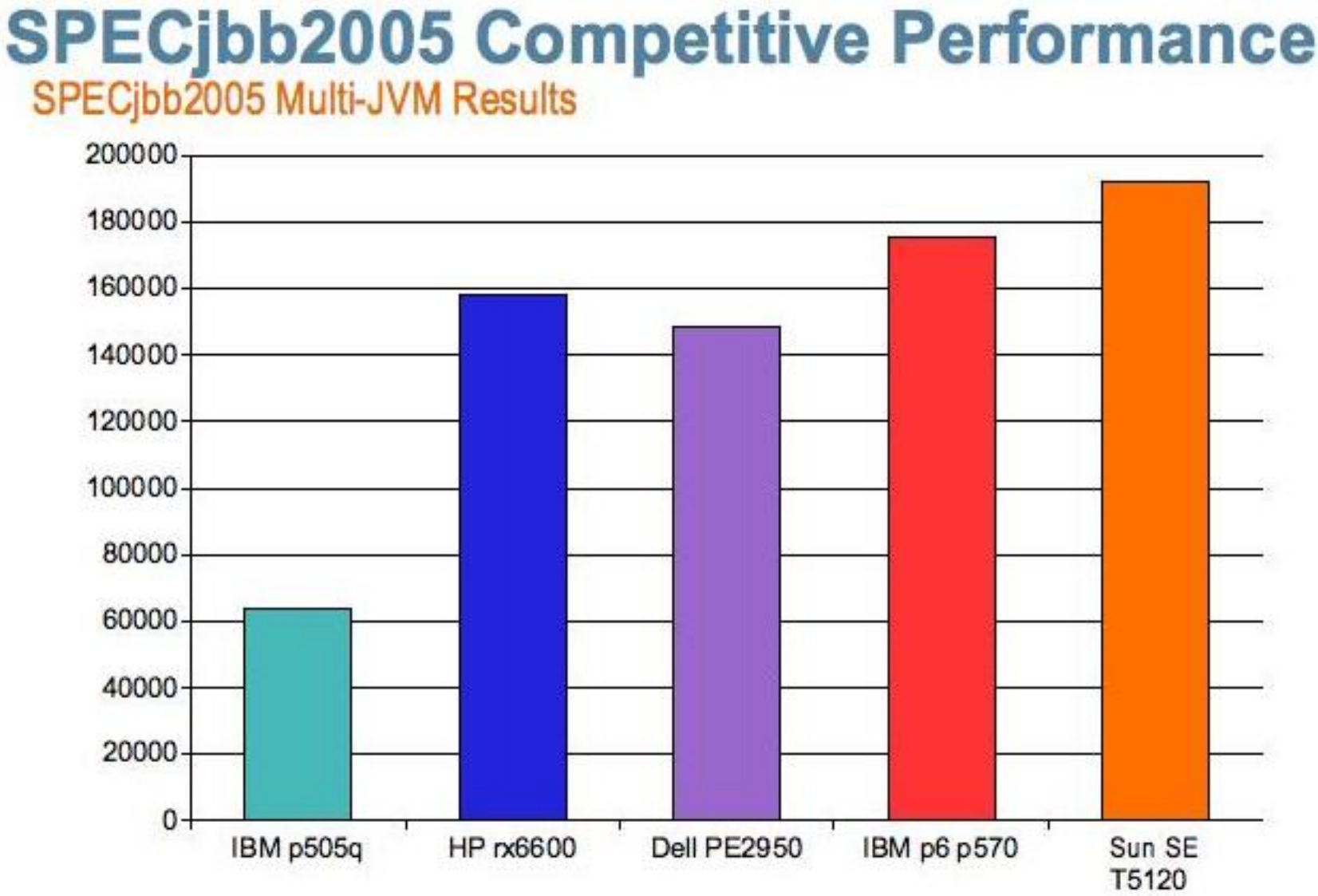
Serwery obsługujące technologię CoolThreads

- Sun SPARC Enterprise T5120 Server
- Sun SPARC Enterprise T5220 Server
- Sun Fire T1000 Server
- Sun Fire T2000 Server
- Sun SPARC Enterprise T1000 Server
- Sun SPARC Enterprise T2000 Server
- Netra T2000 Server
- Netra CP3060 ATCA Blade
- Netra CP3260 ATCA Blade
- Sun Blade T6300 Server Module
- Sun Blade T6320 Server Module

Nawet tanio...

Sun SPARC T5120 - 13.995$

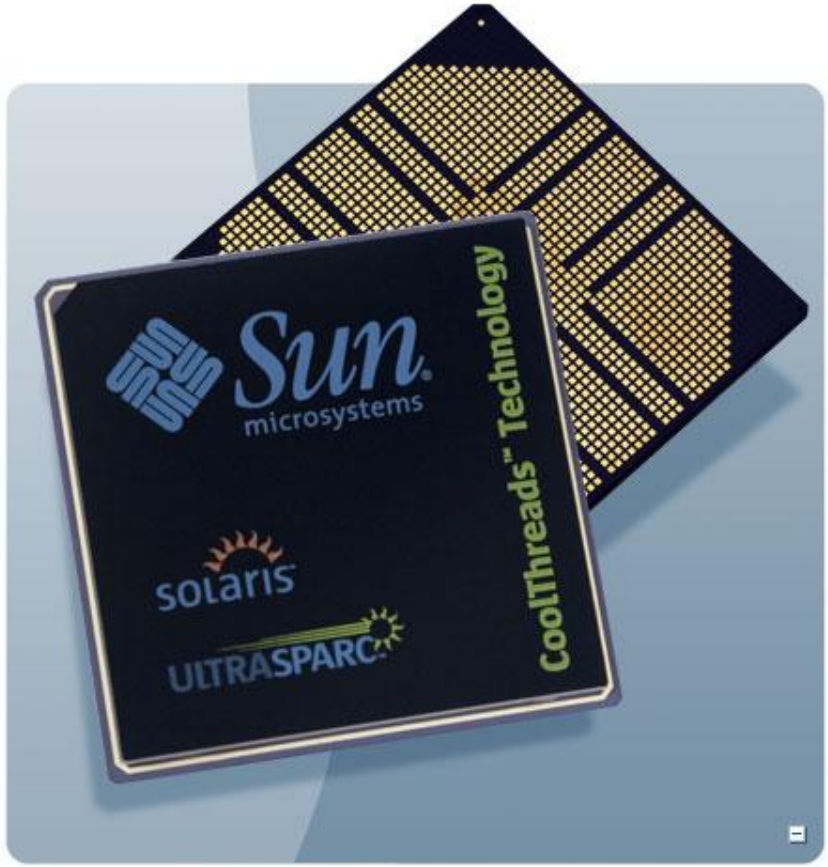
Wydajność


SPECjbb2005 Competitive Performance
SPECjbb2005 Multi-JVM Results
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
IBM p505q
HP rx6600
Dell PE2950
IBM p6 p570
Sun SE T5120

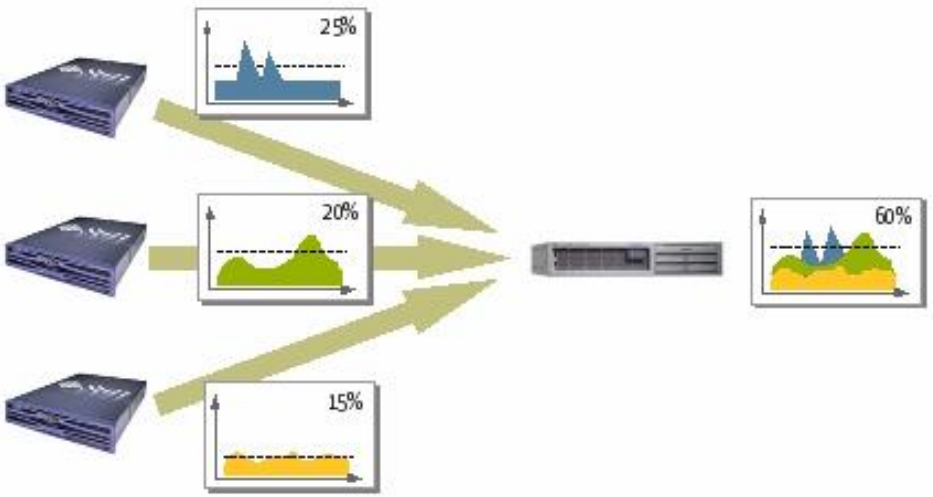
Procesory

osiem 4-wątkowych rdzeni
Moc 72W
Przekątna 2”

- UltraSPARC IV+
- UltraSPARC IV
- UltraSPARC III
- UltraSPARC IIIi
- UltraSPARC IIi



Minimalizacja kosztów



- Prąd
- Chłodzenie
- Miejsce
- Hardware

Oszczędnośc nawet do $5000 rocznie /serwer!

Hypervisor

Platforma pozwalająca na równoległe działanie wielu OS na jednej maszynie

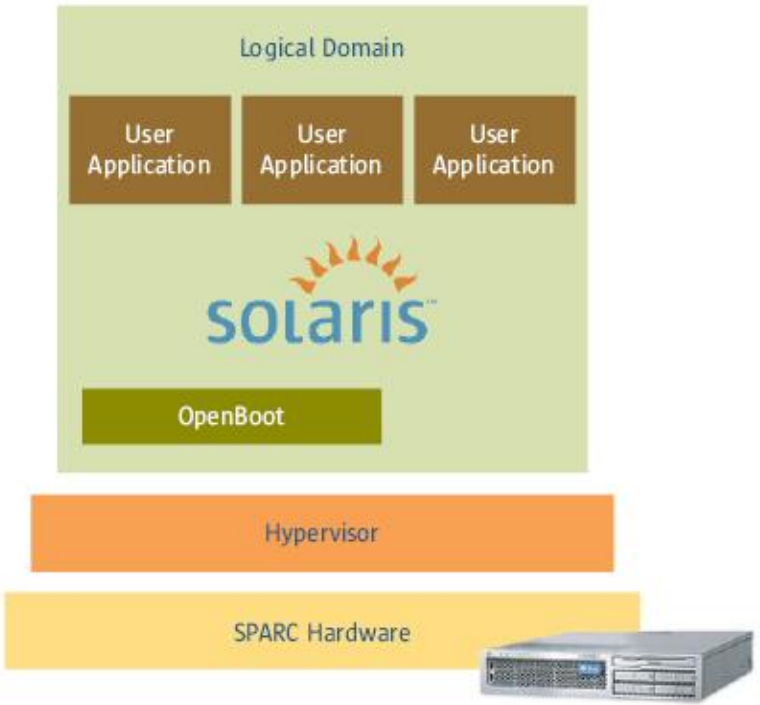


Figure 2-1. The hypervisor firmware layer virtualizes server resources and buffers logical domains from direct hardware access.

- Zarządza zasobami maszyny
- Może stworzyć wirtualną maszynę
- Nakładka na hardware dla OS
- Platforma Sun4v
- Stabilność
- Pierwsza odsłona w 2005

Hypervisor

Obsługiwane systemy operacyjne:

- Solaris 10 11/06 or later
- Ubuntu Linux Server Edition
- FreeBSD (niebawem)
- Wind River Platform for Network Equipment, Linux Edition

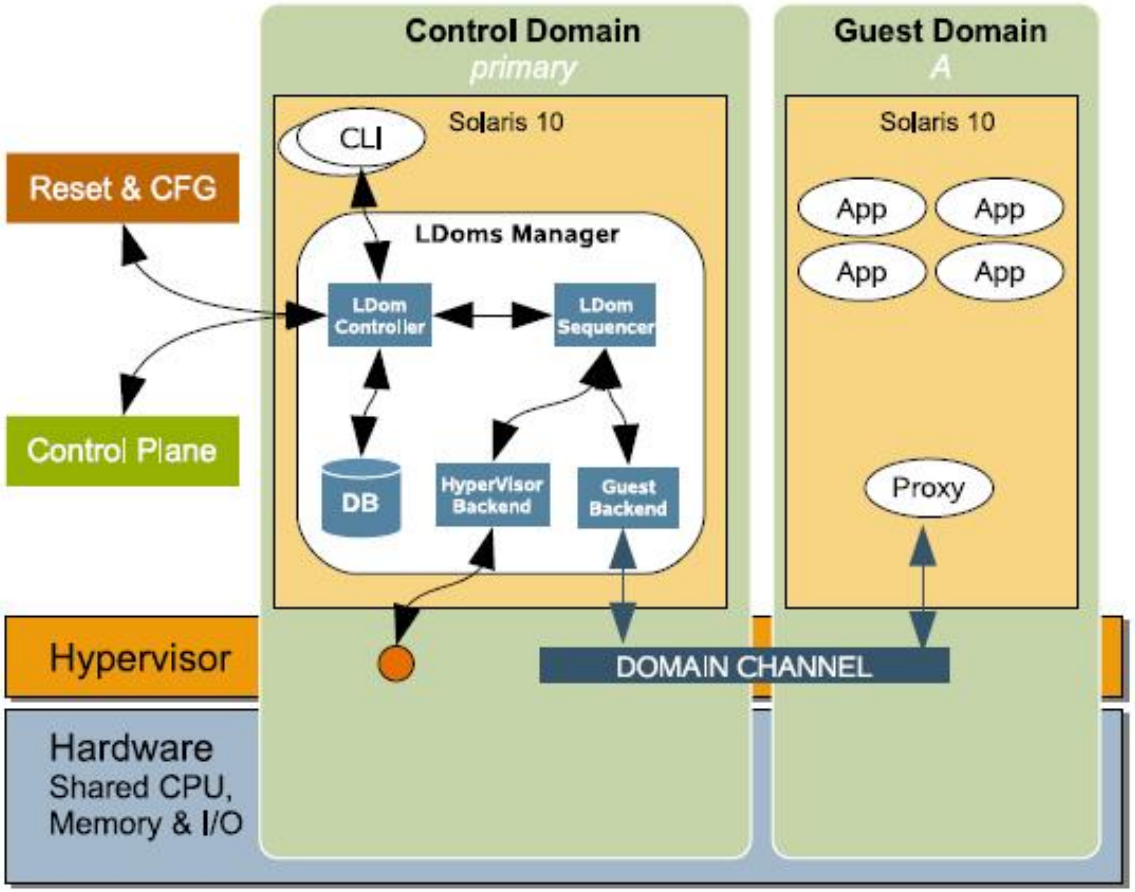
Domeny logiczne

Rodzaje:

- Control Domain
- Service Domain
- I/O Domain
- Guest Domain

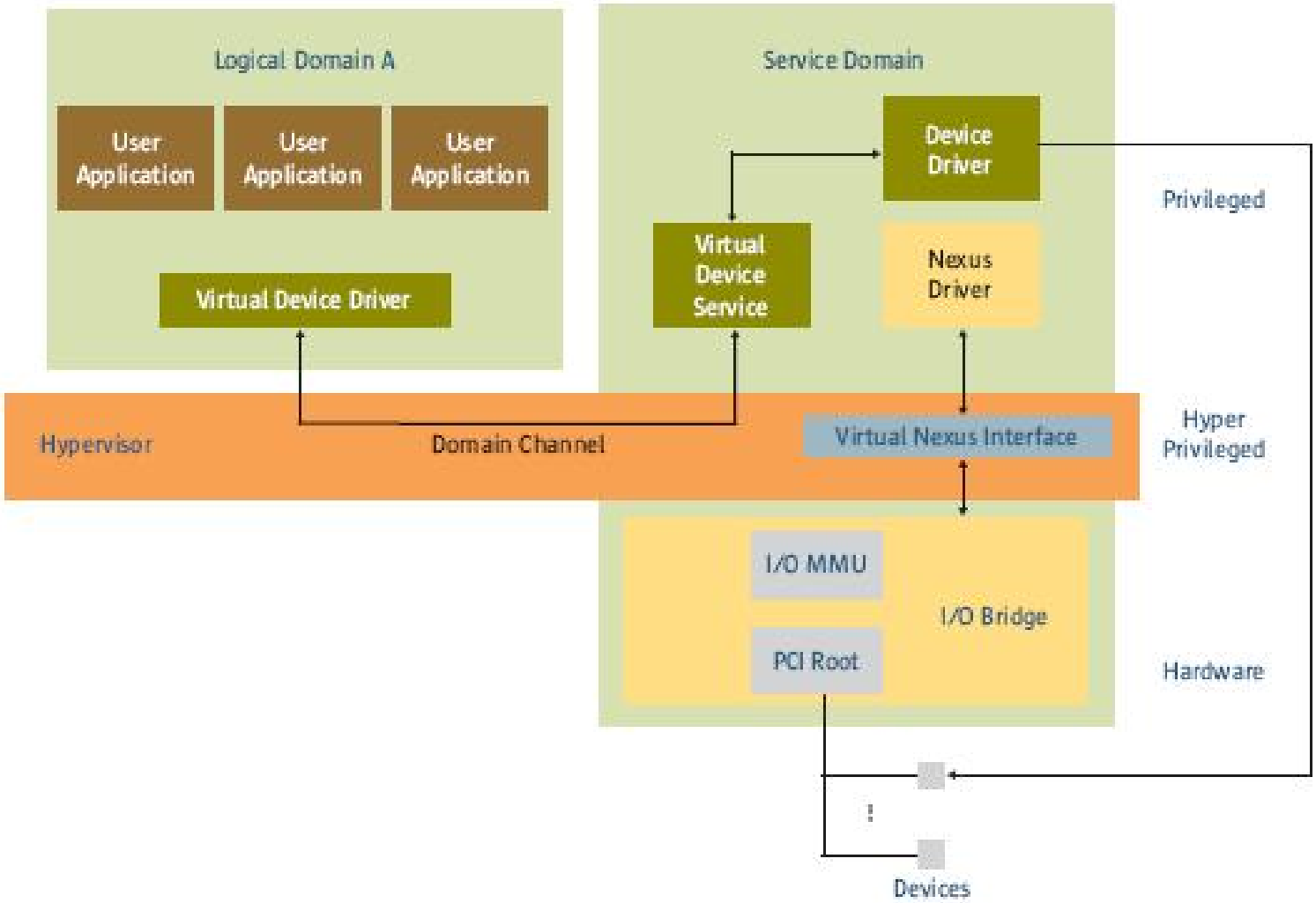
Control Domain

- Tworzy i zarządza innymi logicznymi partycjami
- Komunikuje się poprzez Hypervisor



Service Domain

- Pośredni dostęp do I/O dla Guest Domain



I/O Domain

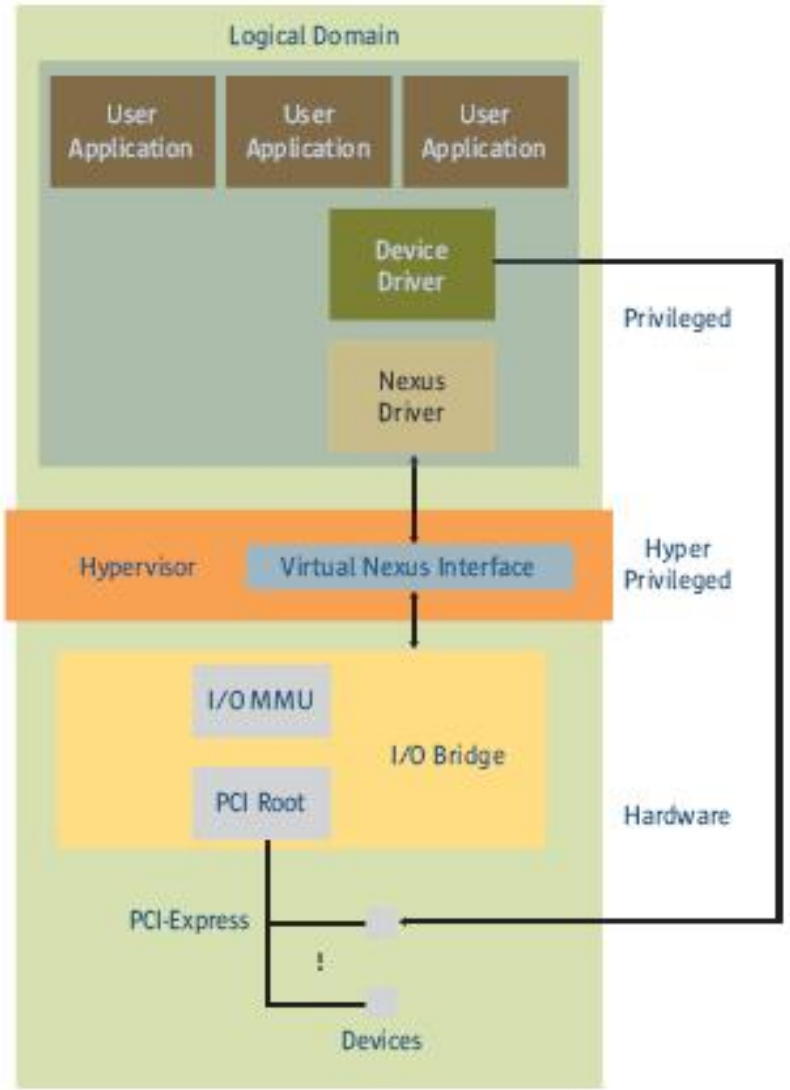


Figure 2-2. I/O domain with direct device ownership

- Bezpośredni dostęp do urządzeń wejścia / wyjścia
- Może udostępniać urządzenia dla innych (Service Domain)
- Max 2 I/O Domain

Bezpośredni dostęp do I/O

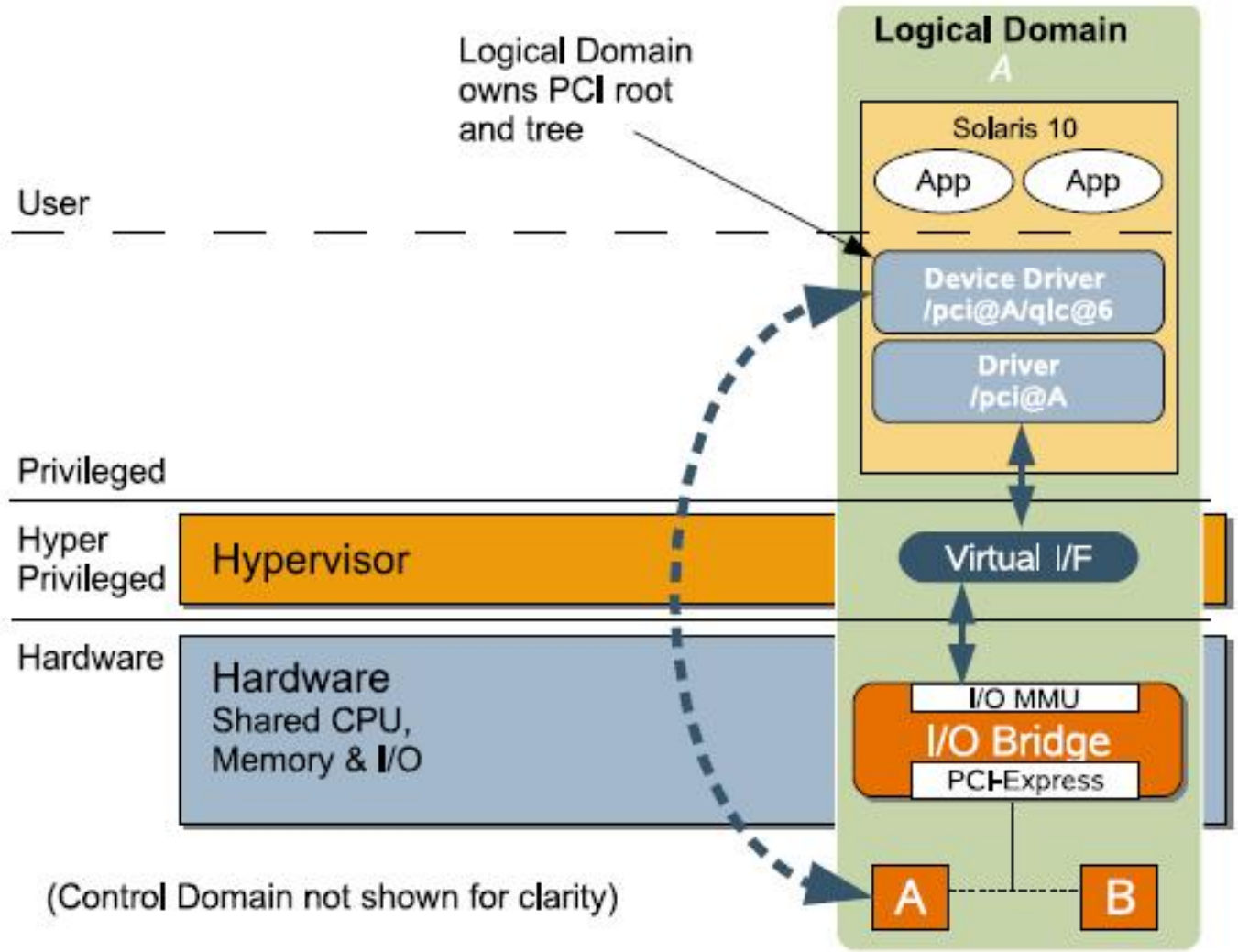


Figure 4. Direct I/O Model, Detailing Ownership at a PCI Root Level

Wirtualny dostęp do I/O

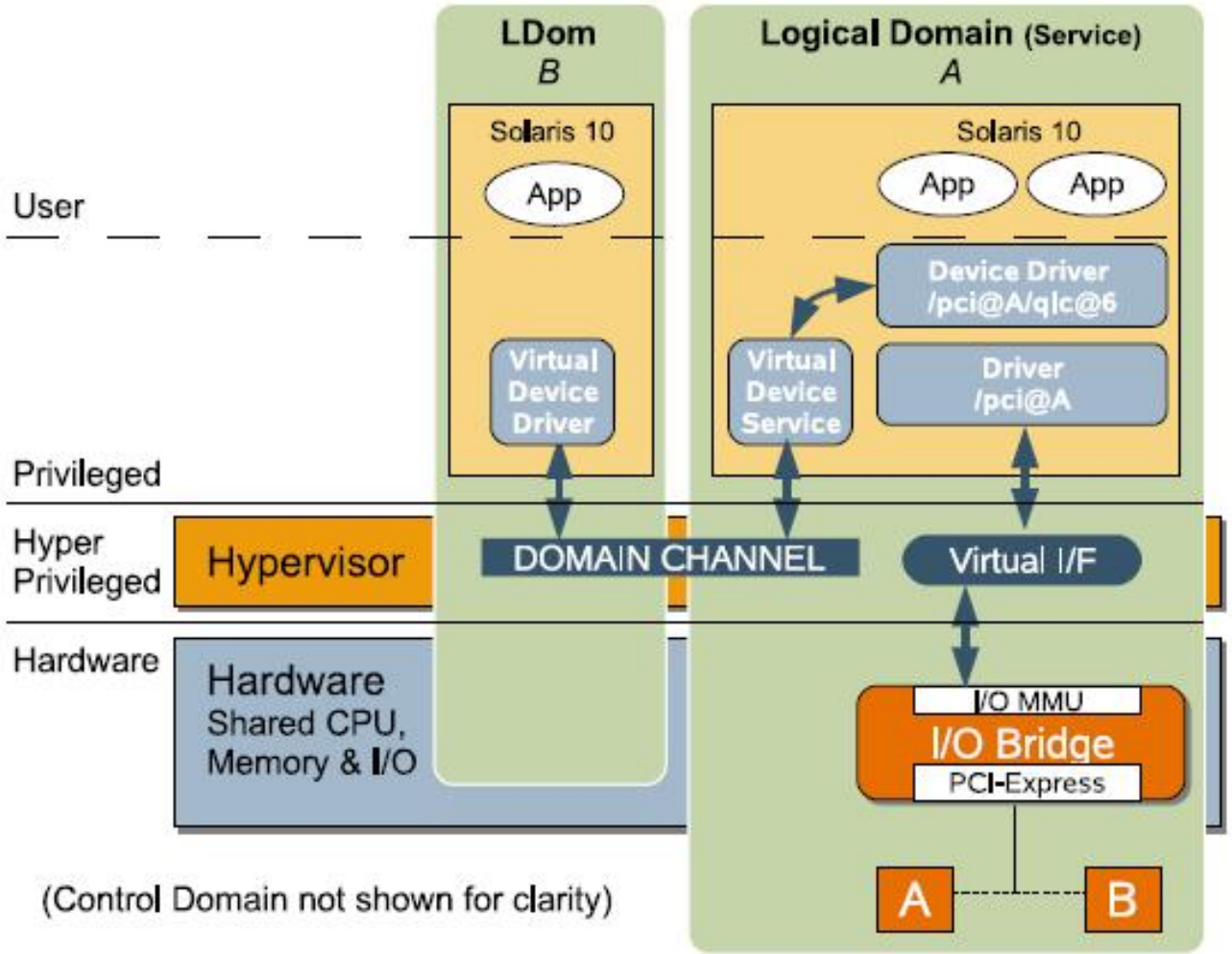


Figure 5. Virtualized I/O Model, Showing Devices Shared From a I/O Service Domain Through a Logical Domain Channel (LDC) to a Guest Domain

Bezpieczeństwo

3 tryby pracy

- użytkownika (user)
- uprzywilejowany (jądro)
- hiper-uprzywilejowany (hipervisor)

Rekonfiguracja

Rozdzielanie zasobów pomiędzy różne partycje logiczne

- CPU
- Pamięć
- Urządzenia I/O

Rekonfiguracja

Rozdzielanie zasobów pomiędzy różne partycje logiczne

- CPU
- Pamięć
- Urządzenia I/O
- Dynamiczna
- Opóźniona

Rekonfiguracja

Rozdzielanie zasobów pomiędzy różne partycje logiczne

- CPU
- Pamięć
- Urządzenia I/O
- Dynamiczna
- Opóźniona
- Configuration Mode

Rekonfiguracja dynamiczna

- Przydzielanie zasobów do włączonej, działającej partycji
- Hypervisor i SO muszą umieć to obsłużyć
- Obecnie tylko CPU można przydzielać dynamicznie (dla zaawansowanych OS)

Rekonfiguracja opóźniona

- Zmiany zachodzą dopiero po restarcie
- Naraz można rekonfigurować tylko jedna partycje
- Możliwość wielu zmian
- Możliwość anulowania zmian (przed reset)

Guest Domain:

```
myldom1# psrinfo -vp
The physical processor has 12 virtual processors (0-11)
  UltraSPARC-T1 (cpuid 0 clock 1000 MHz)
myldom1# prtconf |grep Mem
Memory size: 4096 Megabytes
```

```
primary# /opt/SUNWldm/bin/ldm list-bindings myldom1

Name:   myldom1
State:  active
Flags:  transition
OS:
Util:   100%
Uptime: 10m
Vcpu:   12
        vid     pid     util strand
        0       4       100%   100%
        1       5       100%   100%
        2       6       100%   100%
        3       7       100%   100%
        4       8       100%   100%
        5       9       100%   100%
        6       10      100%   100%
        7       11      100%   100%
        8       12      100%   100%
        9       13      100%   100%
        10      14      100%   100%
        11      15      100%   100%
Memory: 4G
        real-addr          phys-addr          size
        0x4000000          0x44000000         4G
......
```

Control Domain:

Guest Domain:

```
primary# ldm set-vcpu 16 myldom1
```

Control Domain:

```
myldom1# psrinfo -vp
The physical processor has 16 virtual processors (0-15)
  UltraSPARC-T1 (cpuid 0 clock 1000 MHz)
```