Klaster Obliczeniowy IMGW Cosmo

Dokumentacja Powykonawcza

Karta dokumentu:

Zatwierdzone przez:

Wersja	*1. zatwierdzający*	*2. zatwierdzający*
1.0		

Historia wersji dokumentu:

Wersja	*Data*	*Autor*	*Komentarze*
1.0	10-02-2014	Marek Marcola	
		Zygmunt Krawczyk	

Spis treści

Wstęp

Realizacja umowy nr 1696/WU/NT/2013 w ramach projektu "Dostawa, instalacja i wdrożenie systemu komputerowego o wysokiej mocy obliczeniowej , jako elementu tworzonej przez Zamawiającego platformy METEO-RISK, oraz wsparcie Zamawiającego w implementacji numerycznych modeli pogodowych COSMO.

Rozdział 1: Sprzęt i oprogramowanie

1.1 System kasetowy HP c7000

System kasetowy *c7000* składa się z obudowy *c7000* do której można włożyć do 16 serwerów *BL460c Gen8* (sloty serwerowe z przodu). Połączenia miedzy serwerami a także sieciami zewnętrznymi zapewniają moduły interkonekt *HP6120XG* (LAN), *BLc 4X QDR* (IB) i *Brocade 8/24* (FC) (sloty interkonekt z tyłu).

Klaster złożony jest z 10 systemów kasetowych *c7000* (1 w specyfikacji administracyjnej i 9 w specyfikacji obliczeniowej).

Specyfikacja systemu kasetowego *c7000* w wersji administracyjnej:

- 1 x interkonekt *HP6120XG*
- 1 x interkonekt *BLc 4X QDR*
- 2 x interkonekt *Brocade 8/24c*

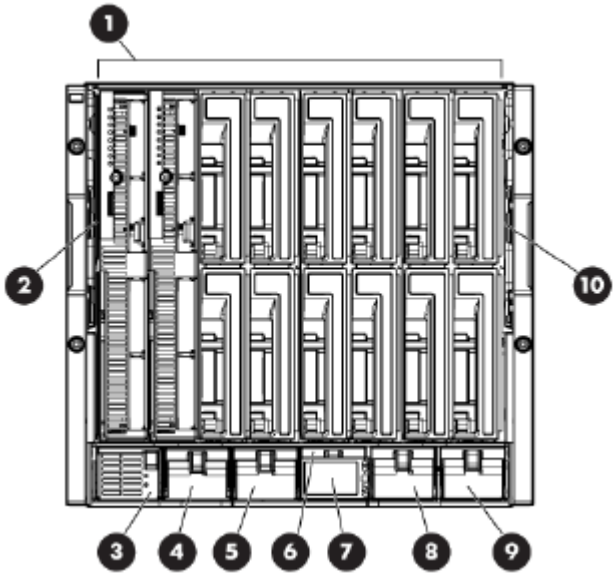
Specyfikacja systemu kasetowego *c7000* w wersji obliczeniowej:

- 1 x interkonekt *HP6120XG*
- 1 x interkonekt *BLc 4X QDR*

W systemach kasetowych znajduje się 145 serwerów *BL460c Gen8* (6 w specyfikacji węzłów administracyjnych i 139 w specyfikacji wezłów obliczeniowych).

1.1.1 Obudowa kasetowa HP c7000

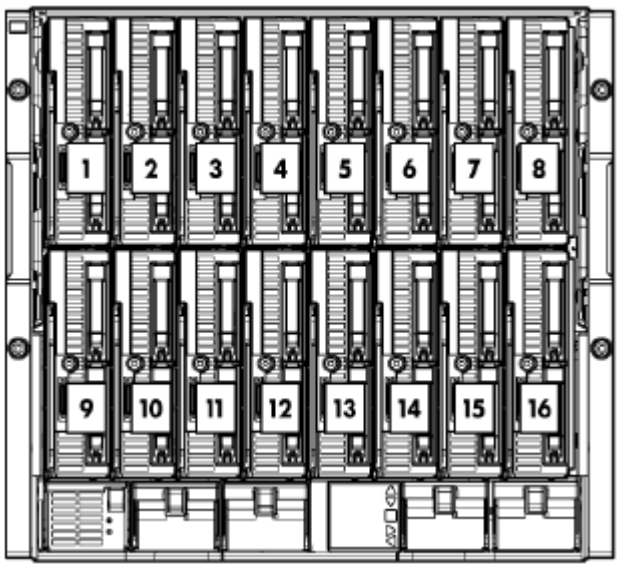
Obudowa *c7000* z przodu posiada miejsce na 16 serwerów połowy wysokości lub 8 serwerów pełnej wysokości, 6 zasilaczy i ekran *Insight Display*.



Item	Description
1	Device bays*
2	Air intake slot (Do not block.)
3	Power supply bay 1
4	Power supply bay 2
5	Power supply bay 3
6	Power supply bay 4
7	Insight Display
8	Power supply bay 5
9	Power supply bay 6
10	Air intake slot (Do not block.)

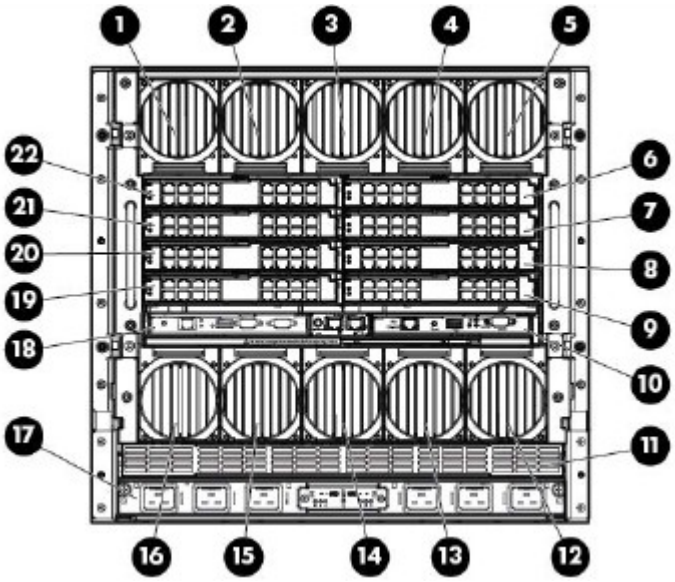
Widok obudowy c7000 z przodu

Serwery HP BL460c Gen8 są serwerami połowy wysokości.



Numeracja serwerów połowy wysokości w obudowie c7000 (przód)

Obudowa *c7000* z tyłu posiada miejsce na 8 modułów interkonekt, 10 wentylatorów i kartę OA.

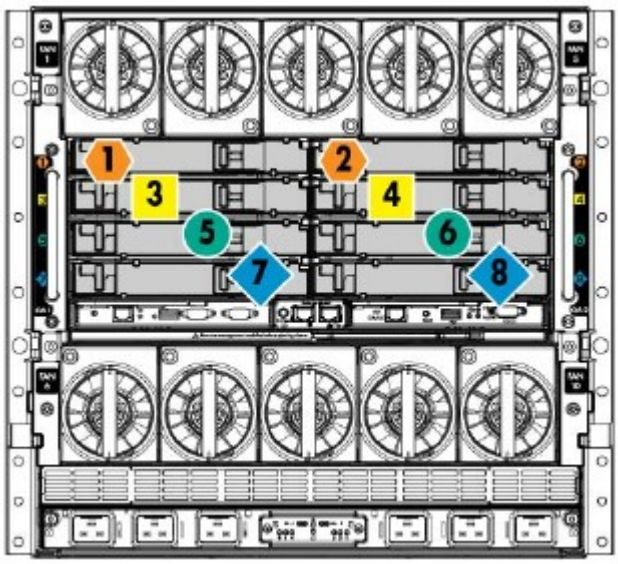


Item	Description
1	Fan bay 1
2	Fan bay 2
3	Fan bay 3
4	Fan bay 4
5	Fan bay 5
6	Interconnect bay 2
7	Interconnect bay 4
8	Interconnect bay 6
9	Interconnect bay 8

Item	Description
10	Onboard Administrator bay 2
11	Power supply exhaust vent (do not block)
12	Fan bay 10
13	Fan bay 9
14	Fan bay 8
15	Fan bay 7
16	Fan bay 6
17	AC power connectors
18	Onboard Administrator bay 1
19	Interconnect bay 7
20	Interconnect bay 5
21	Interconnect bay 3
22	Interconnect bay 1

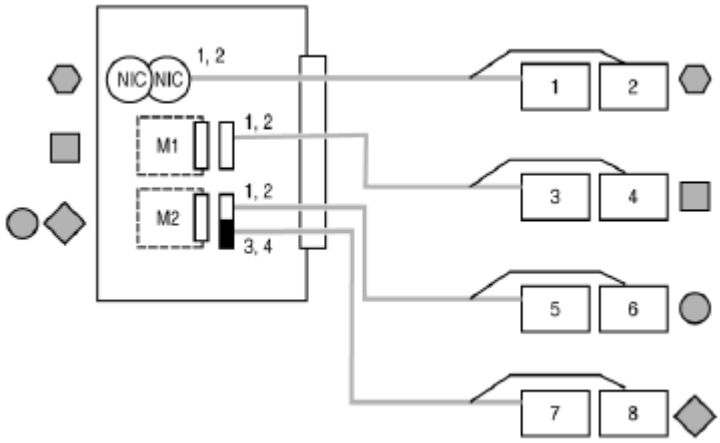
Widok obudowy c7000 z tyłu

Moduły interkonekt mogą zajmować jeden slot (HP6120XG,8/24c) lub dwa (BLc 4X QDR).



Numeracja interkonektów w obudowie c7000 (tył)

Obudowa *c7000* posiada predefiniowane wyprowadzenia sygnałów z serwerów (kart wbudowanych lub kart Mezzanine) na moduły interkonekt.



Mapowania sygnałów serwera połowy wysokości na moduły interkonekt obudowy c7000

1.1.2 Serwer kasetowy HP BL460c Gen8

W systemach kasetowych znajduje się 145 serwerów *BL460c Gen8* (6 w specyfikacji węzłów administracyjnych i 139 w specyfikacji wezłów obliczeniowych).

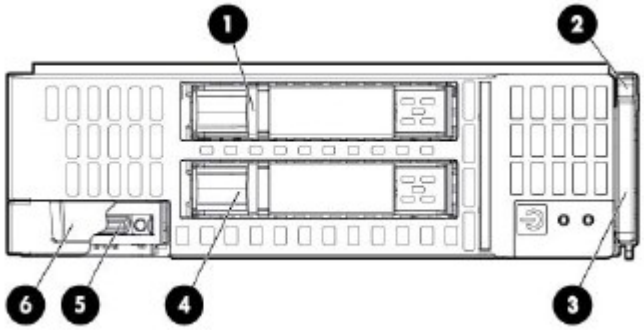
Specyfikacja serwera kasetowego w wersji administracyjnej:

- 2 x 8-Core CPU, 128GB RAM
- karta HP FlexFabric 10Gb 2P (FlexLOM)
- karta HP IB QDR 2P (Mezzanine)
- karta HP FC 8Gb HBA 2P (Mezzanine)

Specyfikacja serwera kasetowego w wersji obliczeniowej:

- 2 x 10-Core CPU, 128GB RAM
- karta HP FlexFabric 10Gb 2P (FlexLOM)
- karta HP IB QDR 2P (Mezzanine)

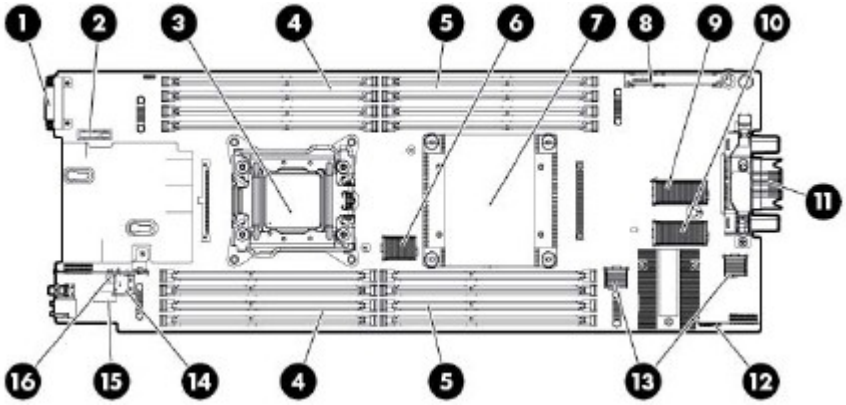
Wszystkie serwery kasetowe nie posiadają dysków wewnętrznych.



Item	Description
1	Hard drive bay 1
2	Server blade release button
3	Server blade release lever
4	Hard drive bay 2
5	HP c-Class Blade SUV connector* (behind the serial label pull tab)
6	Serial label pull tab

Front serwera BL460c Gen8

Na płycie głównej mogą byc zamontowane dwie karty FlexLOM i dwie karty Mezzanine.



Item	Description
1	HP c-Class Blade SUV Cable connector
2	System battery
3	Processor socket 2
4	Processor 2 DIMM slots (8)
5	Processor 1 DIMM slots (8)
6	SAS controller connector
7	Processor socket 1 (populated)
8	Accelerator cache connector
9	Mezzanine connector 1 (Type A mezzanine only)
10	Mezzanine connector 2 (Type A or Type B mezzanine)
11	Enclosure connector
12	MicroSD card slot
13	FlexibleLOM connectors (2)
14	Internal USB connector*
15	System maintenance switch
16	TPM connector

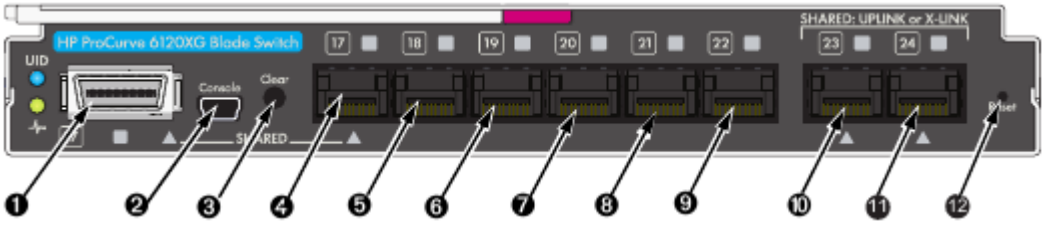
Płyta główna serwera BL460c Gen8

Symbole wyjść sygnałów serwera (FlexLOM, Mezzanine) odpowiadają symbolom umieszczonym na slotach interkonektów.

1.1.3 Przełącznik LAN HP6120XG

Interkonekt HP6120XG jest zarządzalnym przełącznikiem LAN złożonym z 24 10Gb portów.

Porty o numerach 1-16 są portami wewnętrznymi (*downlink*) przyłączonymi do serwerów Blade. Porty o numerach 17-24 są portami zewnętrznymi (*uplink*).



Item	Description
1	Port 17 (10GBASE-CX4)[1]
2	Console Port (USB 2.0 mini-AB connector)
3	Clear button
4	Port 17 SFP+ (10-GbE) slot[1,2]
5	Port 18 SFP+ (10-GbE) slot[2]
6	Port 19 SFP+ (10-GbE) slot[2]
7	Port 20 SFP+ (10-GbE) slot[2]
8	Port 21 SFP+ (10-GbE) slot[2]
9	Port 22 SFP+ (10-GbE) slot[2]
10	Port 23 SFP+ (10-GbE) slot[2]
11	Port 24 SFP+ (10-GbE) slot[2]
12	Reset button (recessed)

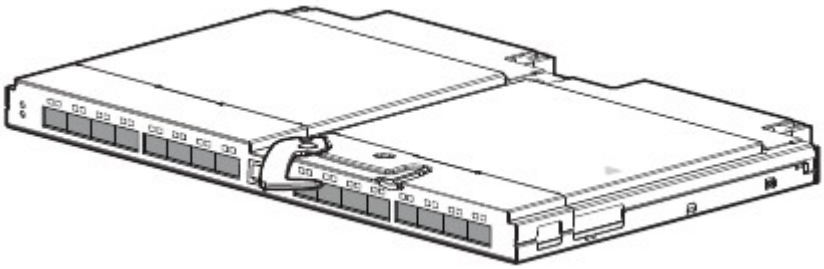
Front przełącznika HP6120XG

Porty zewnętrzne 23 i 24 są dzielone z wewnętrznymi portami inter-link do przełącznika partnera (jeżli jest zainstalowany). Porty zewnętrzne mają prirytet nad portami inter-link, jeżeli w porcie zewnętrznym znajduje się moduł SFP+ to odpowiedni port inter-link zostanie automatycznie zablokowany (w danej chwili może być aktywny tylko jeden port).

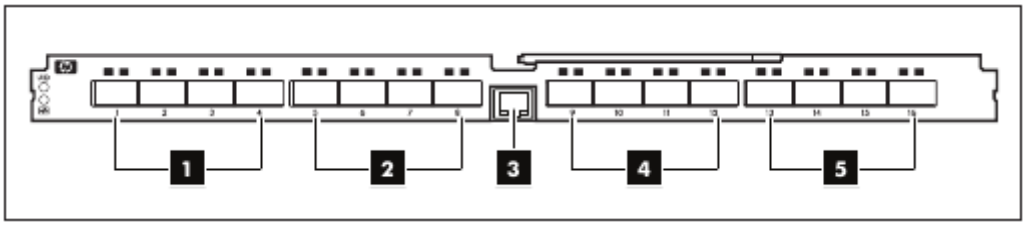
Przełącznik zajmuje jeden slot przeznaczony na moduły interkonekt.

1.1.4 Przełącznik Infiniband HP BLc 4X QDR

Przełącznik HP BLc 4X QDR jest niezarządzalnym przełącznikiem IB. Do działania wymaga zewnętrznego Subnet Managera.



Front przełącznika HP BLc 4X QDR



1. Uplink connectors, bays 1-4
2. Uplink connectors, bays 5-8
3. I²C port (inter-integrated circuit port)
4. Uplink connectors, bays 9-12
5. Uplink connectors, bays 13-16

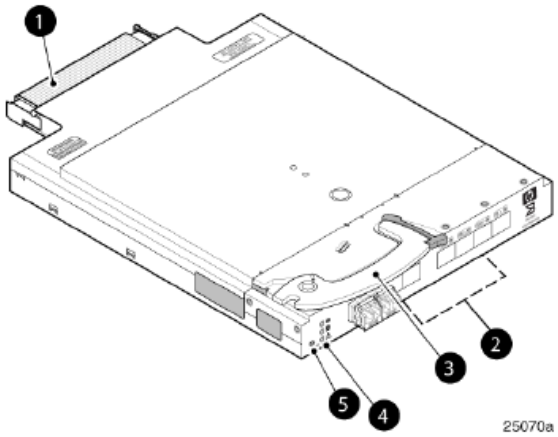
Przypisania portów przełącznika do grup serwerów blade

Przełącznik zajmuje dwa sąsiadujące sloty przeznaczone na moduły interkonekt.

1.1.5 Przełącznik FC Brocade 8/24c

Interkonekt *Brocade 8/24c* jest zarządzalnym przełącznikiem SAN złożonym z 24 8Gb portów.

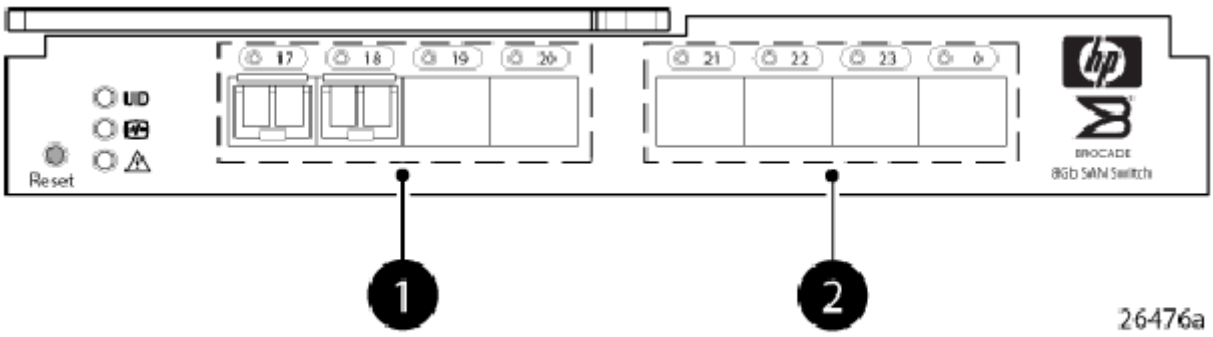
Porty o numerach 1-16 są portami wewnętrznymi (*downlink*) przyłączonymi do serwerów Blade. Porty o numerach 17-23,0 są portami zewnętrznymi (*uplink*).



1. Midplane connector
2. Eight external SFP ports
3. Installation handle
4. Unit ID (UID), Health, and Status LEDs
5. Reset button

Przełącznik Brocade 8/24c

Porty zewnętrzne połączone są w dwa banki.



1. Left bank—ports 17, 18, 19, 20
2. Right bank—ports 21, 22, 23, 0

Front przełącznika Brocade 8/24c

Przełącznik zajmuje jeden slot przeznaczony na moduły interkonekt.

1.2 Macierz dyskowa HP 7400

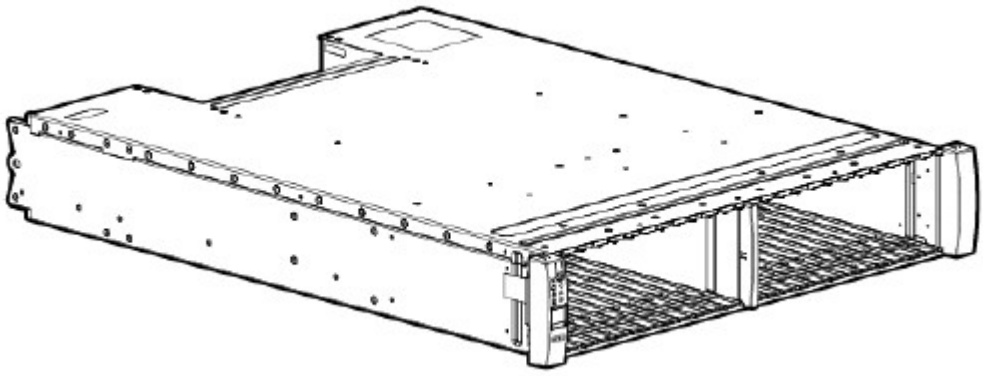
Macierz HP 3PAR StoreServ 7400 posiada następującą specyfikację:

- 4 kontrolery macierzowe w dwóch półkach DCN1, 8 portów FC 8GB (dwa na kontroler)
- 20 dysków 200GB SSD w półkach DCN1 (dziesięc dysków na półkę)
- 48 dysków 2TB SAS w dwóch półkach M6720 (24 dyski na półkę)

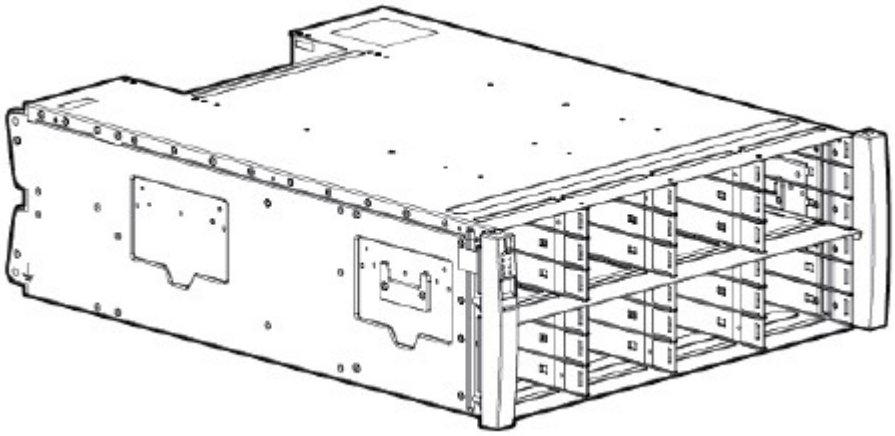
1.2.1 Półki dyskowe

System dyskowy HP 3PAR StoreServ 7000 może składać się z dwóch rodzajów dysków i dwóch rodzajów półek dyskowych:

- Półka dyskowa **HP M6710 (2U24)** może zawierać do 24 2.5” dysków SAS SFF (*small form factor*). Z tyłu półki zamontowane są dwa moduły PCM 580W i dwa moduły I/O.



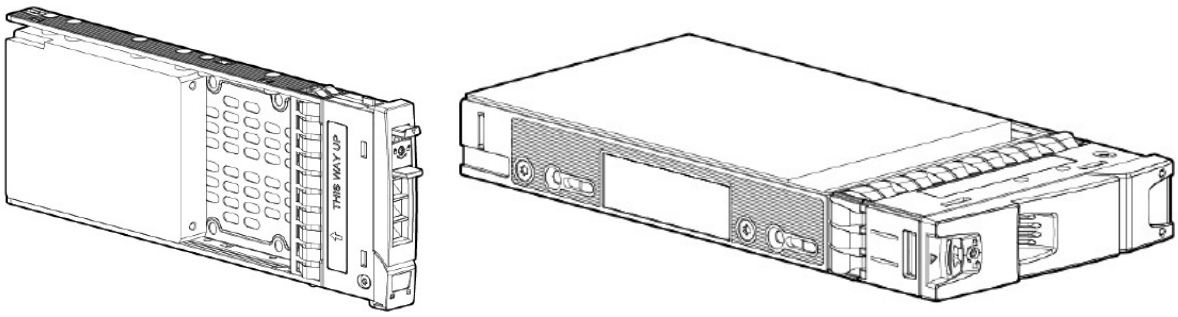
- Półka dyskowa **HP M6720 (4U24)** może zawierać do 24 3.5” dyskow SAS LFF (*large form factor*). Z tyłu półki zamontowane są dwa moduły PCM 580W i dwa moduły I/O.



- Półka dyskowa **HP 3PAR StoreServ 7200** lub **7400** (dwie półki) może zawierać do 24 2.5” dyskow SAS SFF (*small form factor*) (modyfikacja półki HP M6710).
  Z tyłu półki zamontowane są dwa moduły PCM 764W i dwa kontrolery macierzowe.

Półki dyskowe wyświetlone w interfejsie zarządzajacym (GUI lub CLI) mają następujące oznaczenia: **DCS2** (M6710/2U24), **DCS1** (M6720/4U24), **DCN1** (półka z kontrolerami).

Macierze HP 3PAR StoreServ 7000 obsługują tylko dyski SAS 2.5” SFF i SAS 3.5” LFF.



Dyski SAS 2.5” SFF i SAS 3.5” LFF

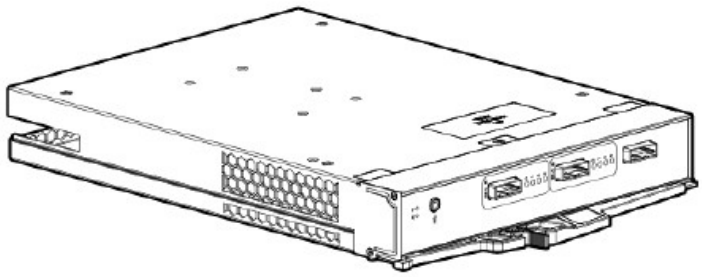
Dyski w półce **M6710** lub w półce z kontrolerami numerowane są w zakresie 0-23 i instalowane pionowo w jednym rzędzie.

Numeracja dysków SFF w półce HP M6710 (2U24) lub półce z kontrolerami (widok z przodu)

Dyski w półce **M6720** numerowane są w zakresie 0-23 i instalowane poziomo w czterech kolumnach po sześć dysków.

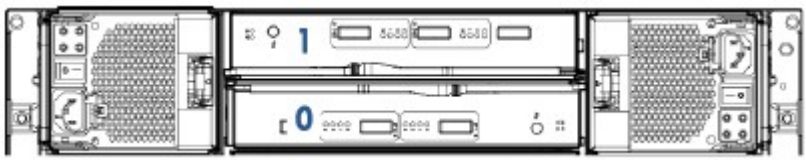
Numeracja dysków LFF w półce HP M6720 (4U24) (widok z przodu)

Moduły I/O łączą półki dyskowe z półkami kontrolerów za pomocą kabli SAS i przekazują dane pomiędzy półkami.



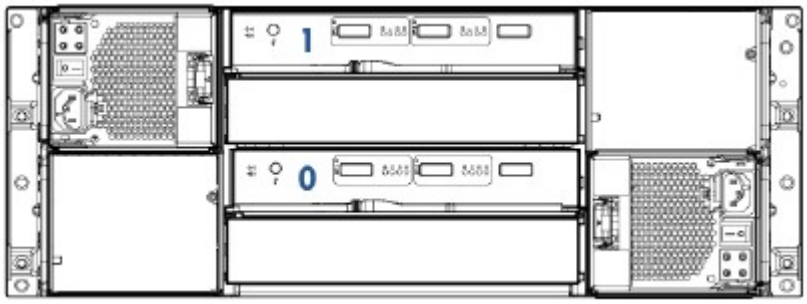
Moduł I/O

Półka **M6710** posiada dwa moduły I/O numerowane w zakresie 0-1 od dołu do góry.



Numeracja modułów I/O w półce HP M6710 (2U24) (widok z tyłu)

Półka **M6720** posiada dwa moduły I/O numerowane w zakresie 0-1 od dołu do góry.

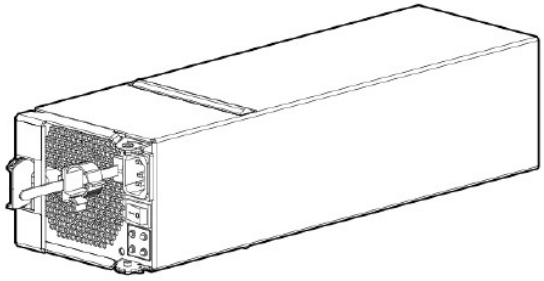


Numeracja modułów I/O w półce HP M6720 (4U24) (widok z tyłu)

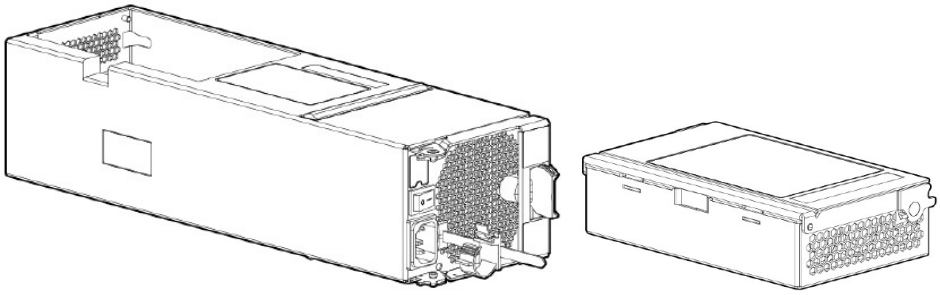
Moduł PCM (*Power Cooling Module*) to zintegrowany zasilacz, wentylator i bateria podtrzymania pamięci *cache*.

Występują dwa rodzaje modułów PCM:

- o mocy 580W wykorzystywany w półkach dyskowych M6710 i M6720 (bez baterii)
- o mocy 764W wykorzystywany w połkach z kontrolerami (z baterią)

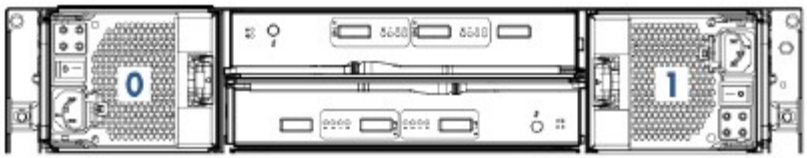


Moduł PCM 580W



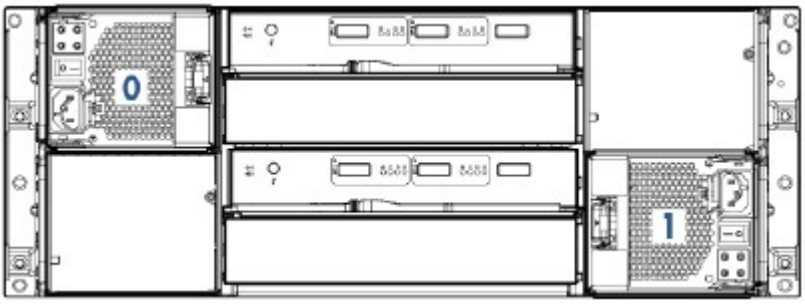
Moduł PCM 764W z baterią

Półka **M6710** posiada dwa moduły PCM numerowane w zakresie 0-1 od lewej do prawej.



Numeracja modułów PCM w półce HP M6710 (2U24) (widok z tyłu)

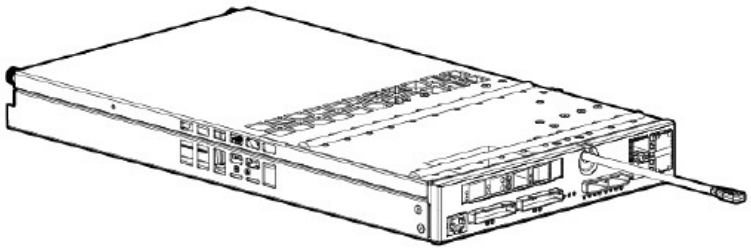
Półka **M6720** posiada dwa moduły PCM umieszczone po przekątnej (w pozostałych slotach umieszczone są maskownice) numerowane w zakresie 0-1 od lewej do prawej.



Numeracja modułów PCM w półce HP M6720 (4U24) (widok z tyłu)

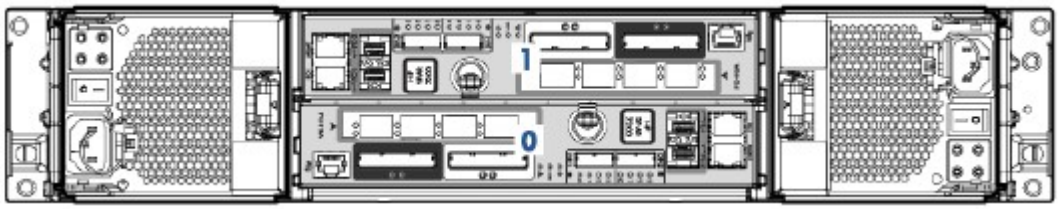
1.2.2 Węzły z kontrolerami macierzowymi

Kontrolery w systemie dyskowym zarządzają danymi systemu (w tym podsystemem *cache*) oraz udostępniają przyłączonym komputerom spójny i zwirtualizowany widok przestrzeni dyskowej.



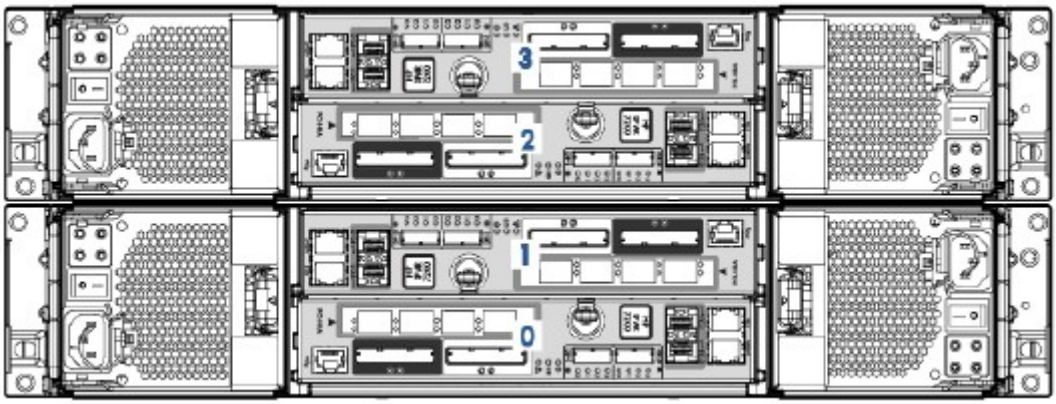
Kontroler

Kontrolery w modelu 7200 numerowane są w zakresie 0-1 od dołu do góry.



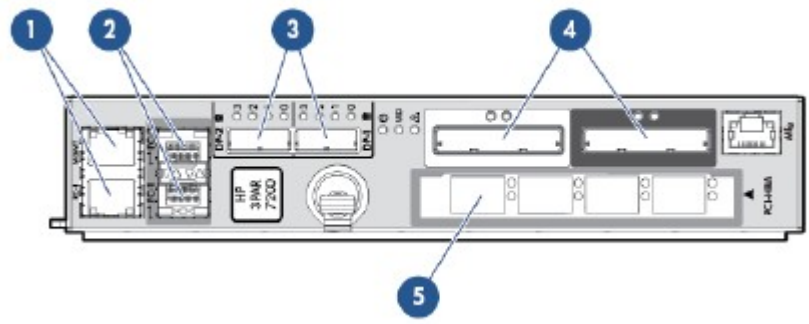
Numeracja kontrolerów w modelu 7200 (widok z tyłu)

Kontrolery w modelu 7400 numerowane są w zakresie 0-3 od dołu do góry.



Numeracja kontrolerów w modelu 7400 (widok z tyłu)

Każdy kontroler posiada porty przeznaczone do podłączenia dodatkowych skrzynek dyskowych (SAS), podłączenia komputerów (FC,Eternet), replikacji (FC,Eternet) i zarządzania (RS232).

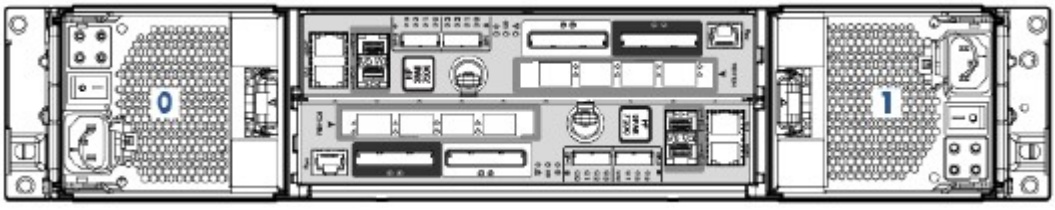


Callout	Port
1	Ethernet (2) MGMT - used to connect to the storage array management interfaces RC - connects to Remote Copy
2	Fibre channel (FC-1 and FC-2) - used to connect to Host Systems
3	SAS (DP-2 and DP-1) - connects to the drive enclosures and I/O modules using SAS cables
4	Node Interconnect - Used with four directional interconnect cables that connect the controller nodes (4-node 7400 only)
5	Fibre channel adaptor (4-ports) or CNA (2-ports).

NOTE: The Mfg port is not used.

Lokalizacja portów kontrolera (widok z tyłu)

Każda półka dyskowa z kontrolerami (*nodes*) posiada dwa moduły PCM (*Power Cooling Module*) numerowane w zakresie 0-1 od lewej do prawej.



Numeracja modułów PCM w półce kontrolera (widok z tyłu)

1.3 Przełącznik InfiniBand Mellanox QDR/FDR10 36P (SX6025)

Przełącznik SX6025 jest zewnętrznie zarządzalnym 36 portowym przełącznikiem IB o wysokości 1U. Do działania wymaga zewnętrznego Subnet Managera.

Front przełącznika SX6025

Tył przełącznika SX6025

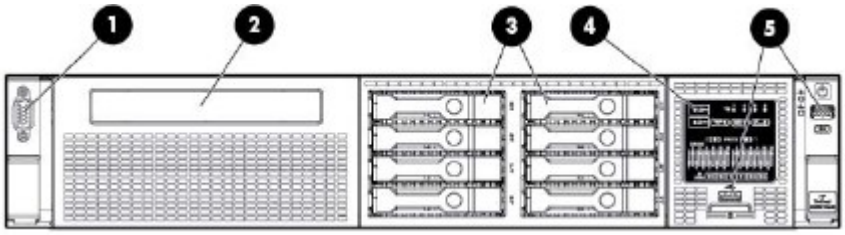
Numeracja portów przełącznika SX6025

Porty przełącznika w trybie IB mogą pracować z prędkościami FDR10/QDR(40Gb/s), DDR(20Gb/s) lub SDR(10Gb/s).

1.4 Serwer HP Proliant DL380p Gen8

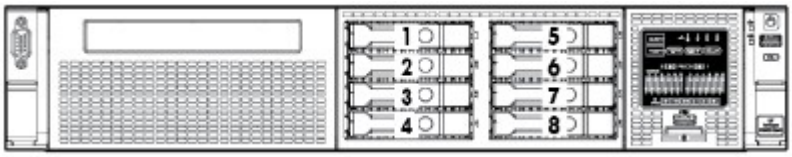
Serwer backupu HP Proliant DL380p Gen8 posiada następującą specyfikację:

- 2 x 4-Core CPU, 96GB RAM
- karta HP IB QDR/ETH10 2P (PCIe)
- karta 2 x HP FC 8Gb HBA 2P (PCIe)

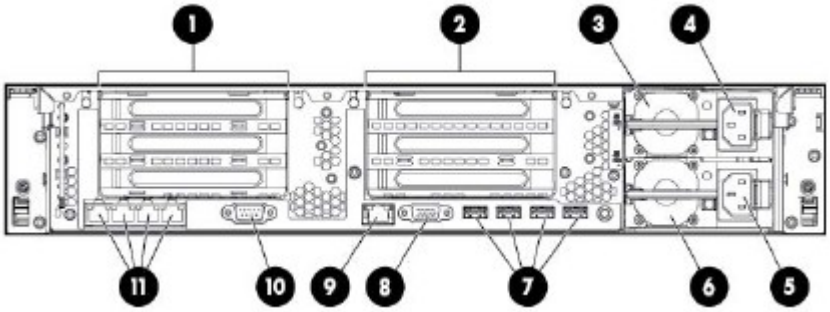


Item	Description
1	Video connector
2	SATA optical drive bay
3	Drive bays
4	Systems Insight Display
5	USB connectors (2)

Front serwera HP Proliant DL380 Gen8 (8-SFF)



Numeracja dysków serwera HP Proliant DL380 Gen8 (8-SFF)



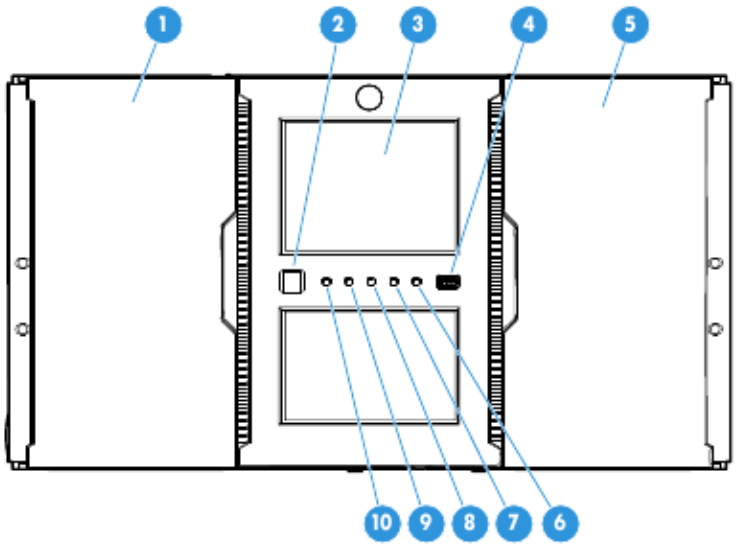
Item	Description
1	PCIe slots 1–3 (top to bottom)
2	PCIe slots 4–6 (top to bottom)
3	Power supply 1 (PS1)
4	PS1 power connector
5	PS2 power connector
6	Power supply 2 (PS2)
7	USB connectors (4)
8	Video connector
9	iLO connector
10	Serial connector
11	FlexibleLOM ports (Shown: 4x1Gb/Optional: 2x10Gb); port 1 on right side

Tył serwera HP Proliant DL380 Gen8 (8-SFF)

1.6 Biblioteka MSL 6480

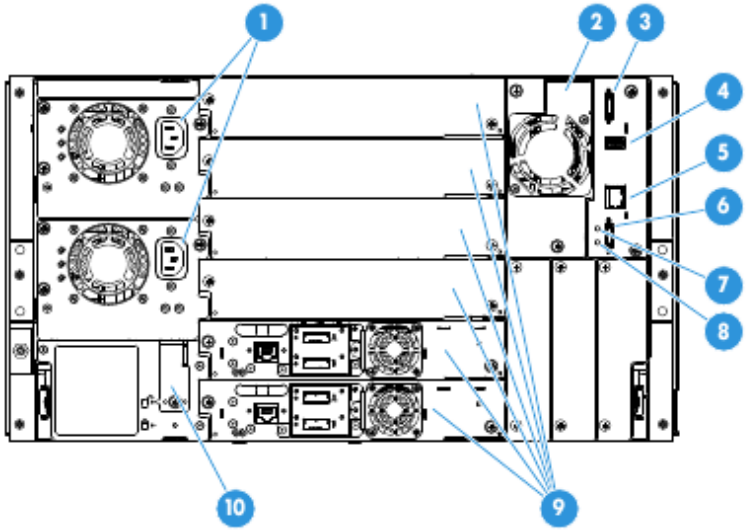
Specyfikacja biblioteki MSL6480

- HP MSL 6480 SBM, HP MSL 6480 SEM
- 4 napędy HP LTO-6



1. Magazine access door
2. Power button
3. LCD touch screen
4. USB port
5. Mailslot, magazine access door
6. Error LED, amber
7. Attention LED, amber
8. Clean LED, amber
9. Ready LED, green
10. Unit Identification (UID) LED, blue

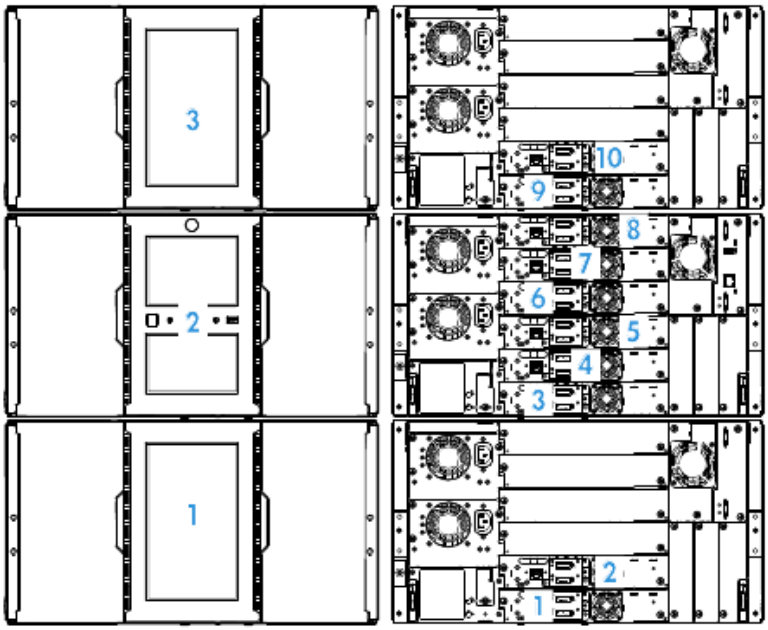
Front bilioteki MSL6480



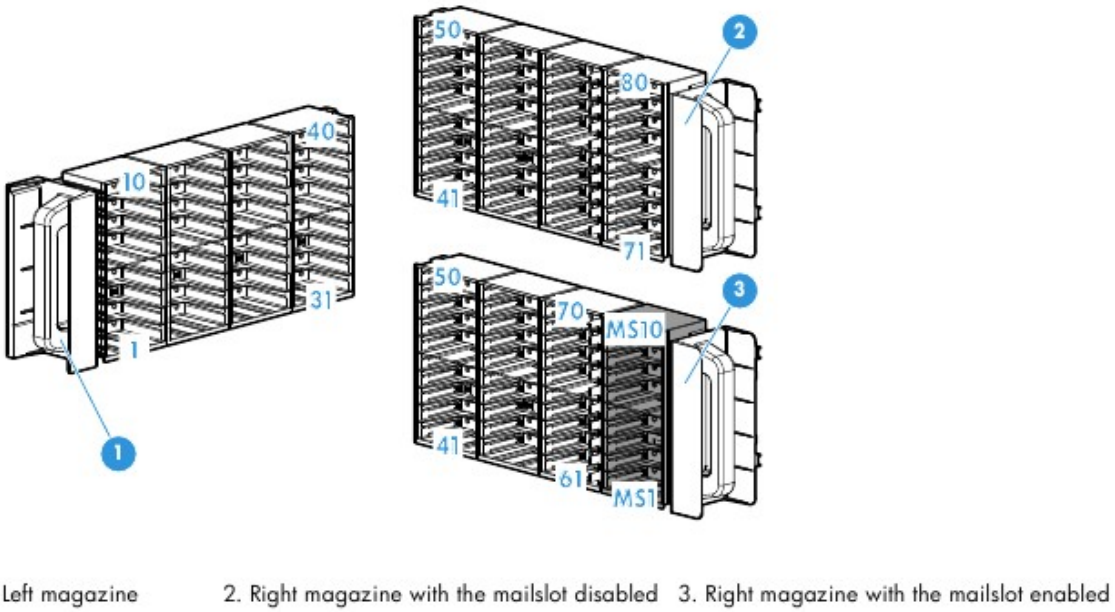
1. Power supplies
2. Chassis fan
3. Upper expansion module interconnect port
4. USB port (base module only)
5. Ethernet port (base module only)
6. Lower expansion module interconnect port
7. Controller health status LED, green
8. UID LED, blue
9. Half-height tape drive locations
10. Module alignment mechanism

Tył bilioteki MSL6480

Moduły i napędy w bibliotece MSL6480 numerowane są z dołu do góry zaczynając od 1.

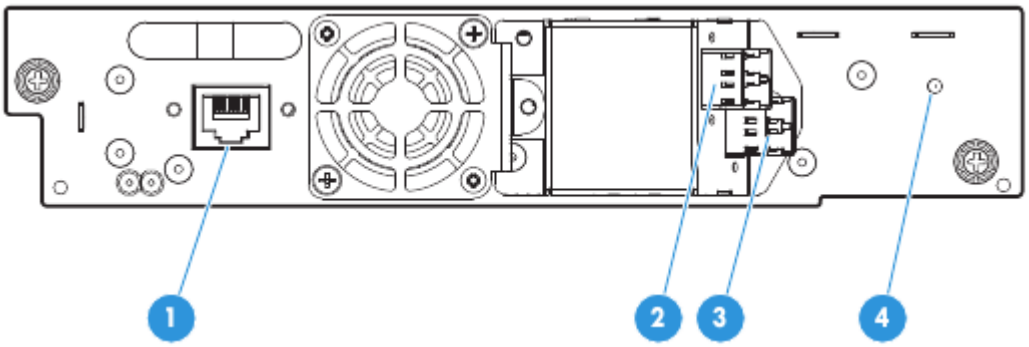


Numeracja modułów i napędów w bibliotece MSL6480



Numeracja slotów w bibliotece MSL6480

Napędy LTO-6 są napędami połowy wysokości.



1. Tape drive Ethernet port (reserved for future use)
2. FC port A
3. FC port B (LTO-6 only)
4. Tape drive power LED, green

Napęd taśmowy LTO-6 (panel tylny)

1.7 Oprogramowanie

1.7.1 Oprogramowanie komercyjne

Oprogramowanie komercyjne wykorzystane w klastrze:

- HP Data Protector 8.1
- Intel Cluster Studio XE 2013
- HP SIM 7.3

1.7.2 Oprogramowanie niekomercyjne

Oprogramowanie niekomercyjne wykorzystane w klastrze:

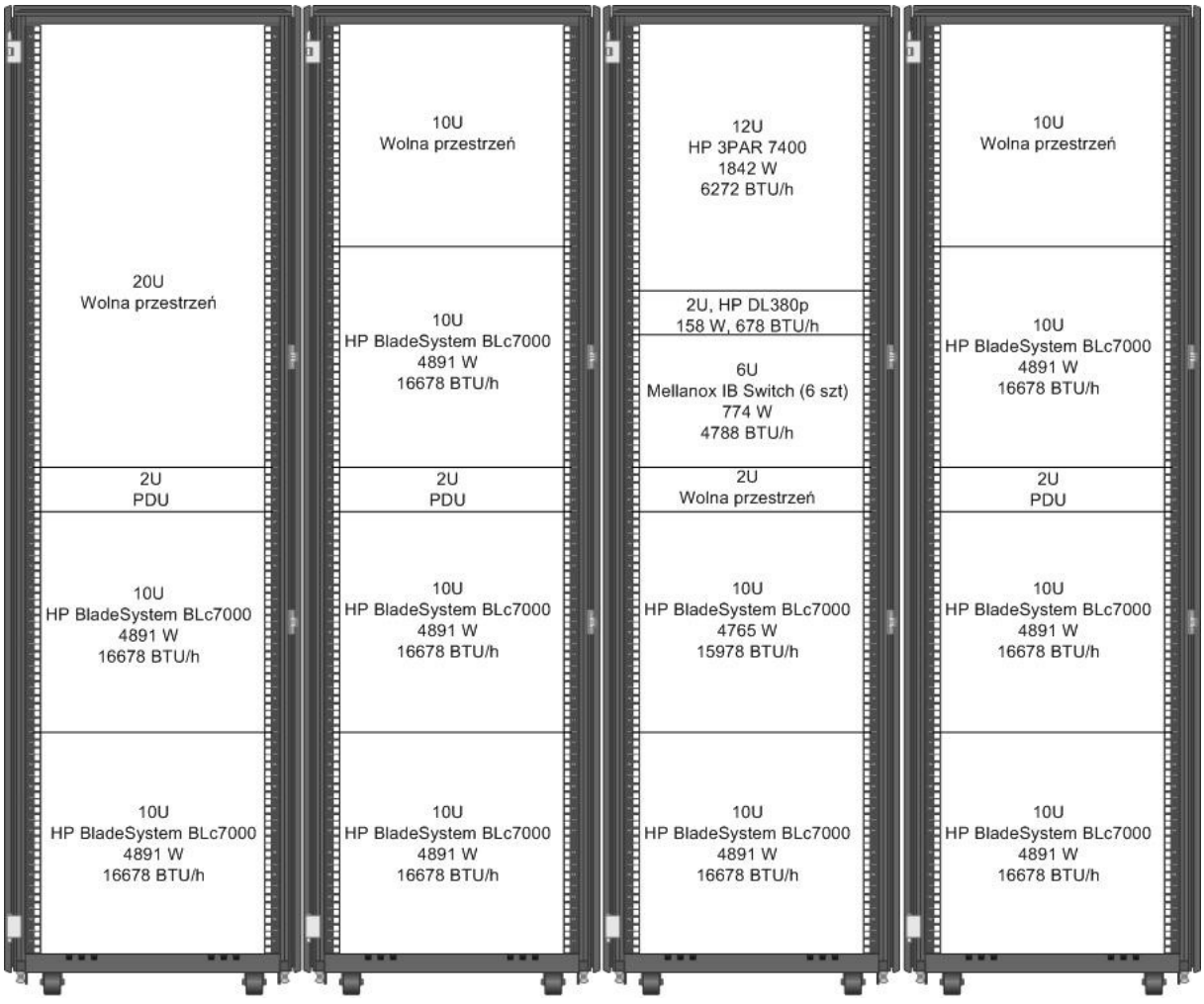
- ScientificLinux 6.4 x86_64 (z wirtualizacją KVM i klastrem HA)
- Ganglia 3.6.0
- Torque 4.2.5
- Lustre 2.4.2

Rozdział 2: Instalacja fizyczna

2.1 Szafy teleinformatyczne

Wraz ze sprzętem dostarczone zostały 4 szafy 42U. Przy wyborze rozmieszczenia sprzętu uwzględniono maksymalne rozłożenie wydzielanego ciepła pomiędzy szafy, specyfikę prowadzenia kabli IB oraz możliwości rozbudowy.

Poniższy rysunek przedstawia sposób rozmieszczenia sprzętu w szafach:

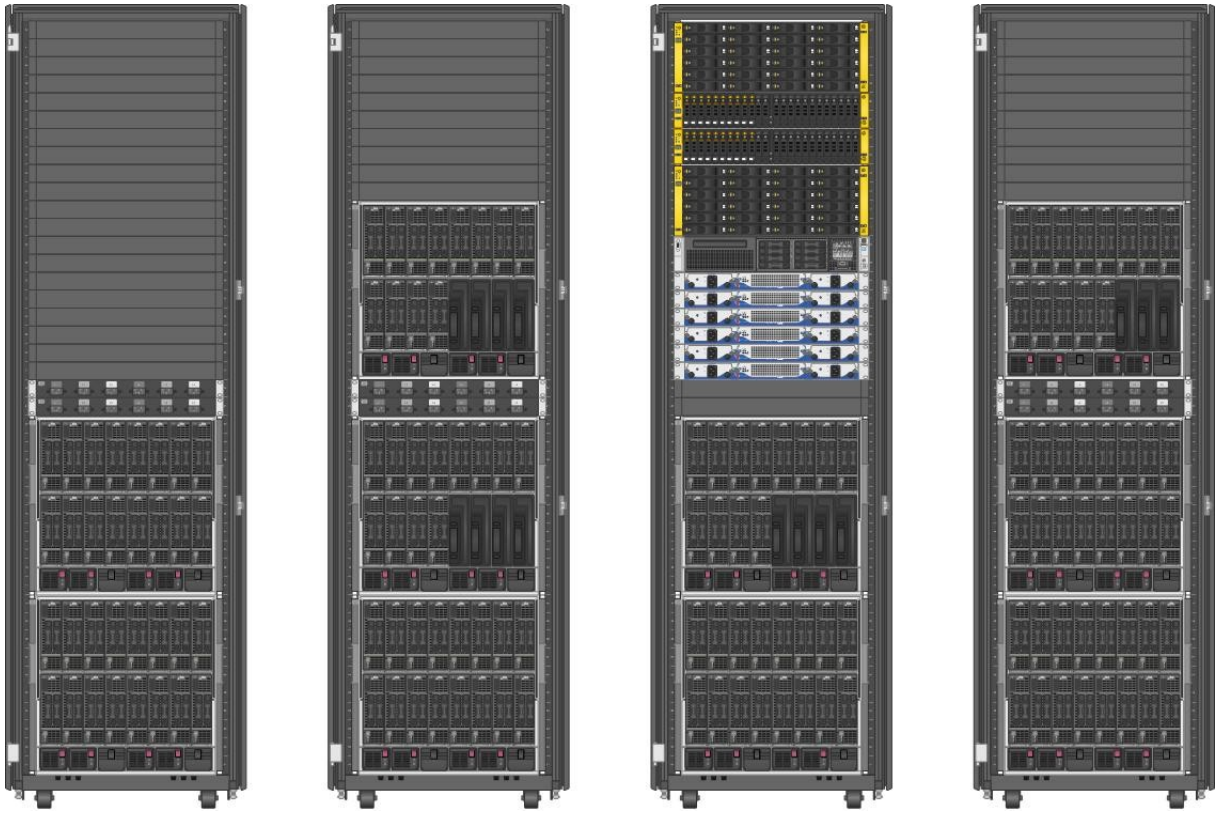


Rozmieszczenie sprzętu w szafach

Biblioteka taśmowa HP StoreEver MSL6480 zajmuje wysokość 12U i została umieszczona w oddzielnej szafie.

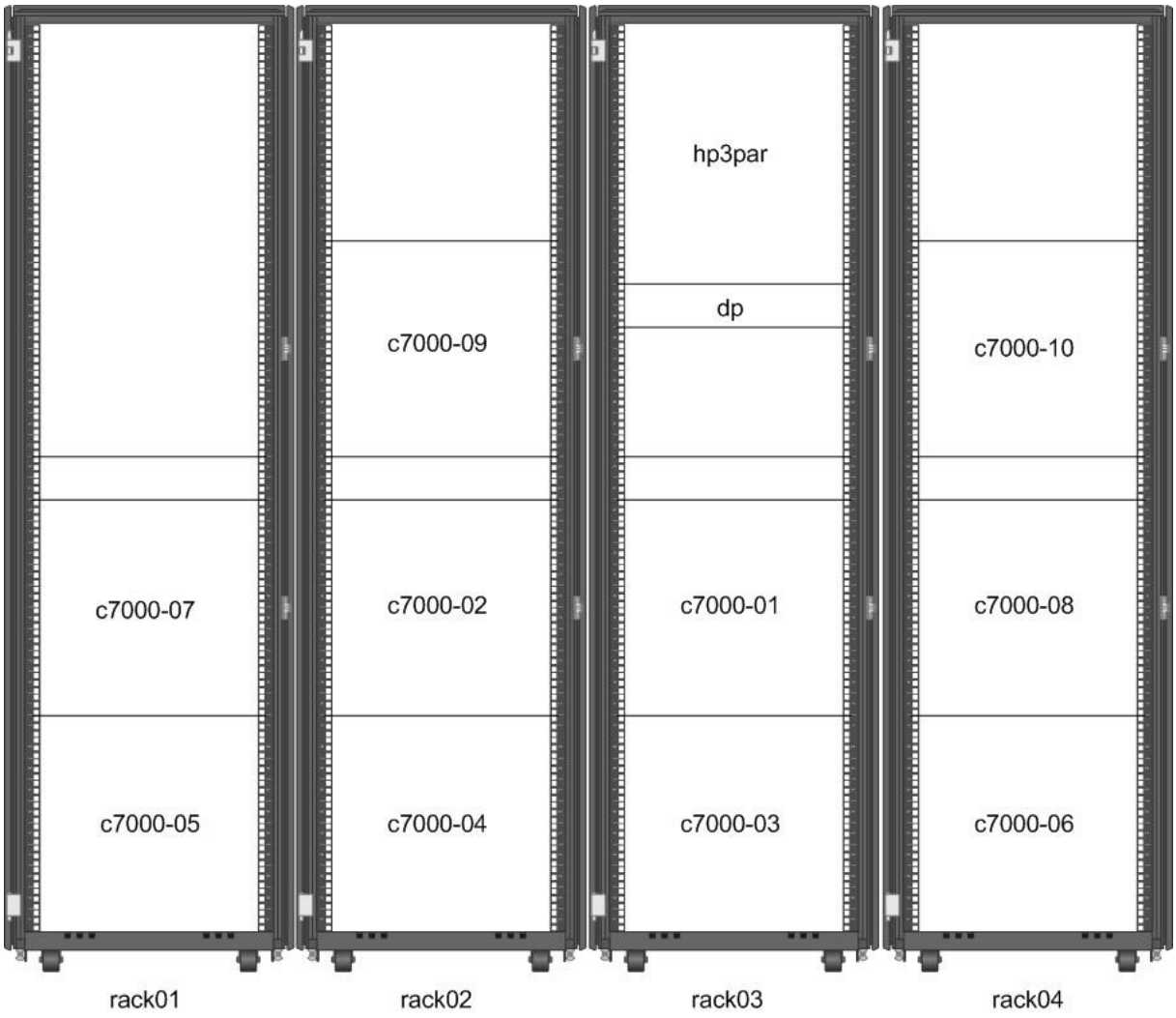
2.2 Sprzęt w szafach teleinformatycznych

Poniższy rysunek przedstawia wizualizację frontów szaf z uwzględnieniem zainstalowanych urządzeń:



Wizualizacja frontów szaf

Poniższy rysunek przedstawia wizualizację frontów szaf z uwzględnieniem logicznych nazw zainstalowanych urządzeń:



Wizualizacja frontów szaf – lokalizacja urządzeń

2.3 Parametry środowiskowe

Pobór mocy elektrycznej i waga:

Nazwa	*Pobór mocy [W]*	*Waga [kg]*
Szafa nr 1 (*rack01*)	9782	544
Szafa nr 2 (*rack02*)	14673	748
Szafa nr 3 (*rack03*)	12430	715
Szafa nr 4 (*rack04*)	14673	748
Razem:	**51558**	**2755**

Parametry środowiskowe:

Nazwa parametru	*Zakres wartości*
Temperatura pracy	10 ÷ 35 ˚C
Maksymalna zmiana temperatury	10 ˚C/h
Wilgotność względna	20 ÷ 80 %

Rozdział 3: Sieć LAN

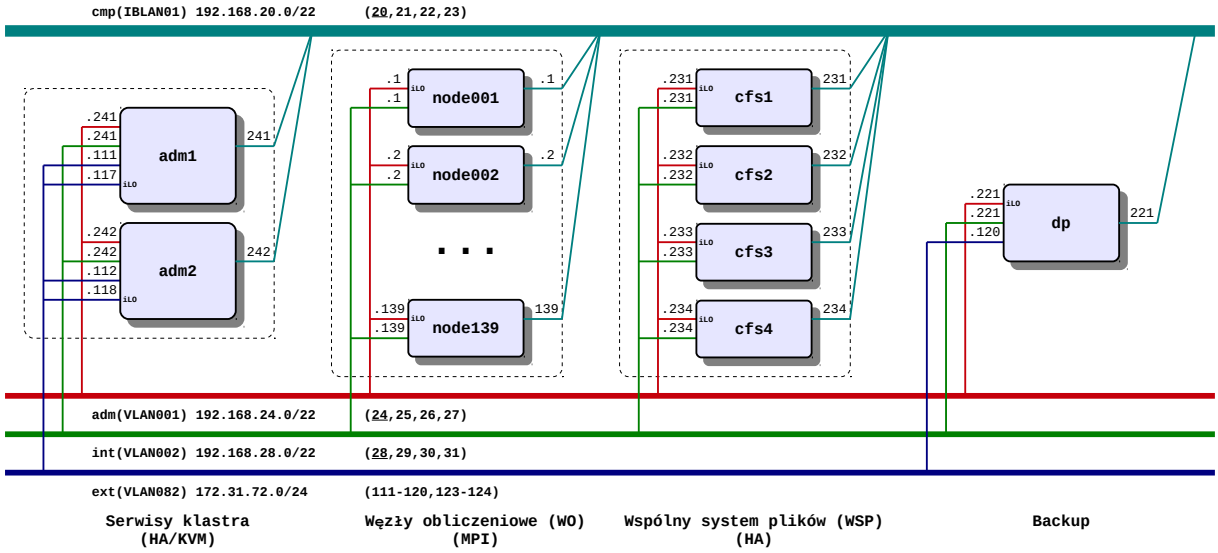
3.1 Architektura

Na potrzeby klastra zdefiniowano następujące klasy adresów IP:

Sieć IP	*VLAN*	*Opis*	*Uwagi*
192.168.20.0/22	IBLAN01(cmp)	Sieć obliczeniowa	Aktywność WO.
192.168.24.0/22	VLAN001(adm)	Sieć administracyjna	Adresacja iLO, OA, przełączników.
192.168.28.0/22	VLAN002(int)	Sieć wewnętrzna	Uruchamianie WO.
172.31.72.0/24	VLAN082(ext)	Sieć zewnętrzna	Punkt styku z siecią IMGW.

Sieć InfiniBand klastra zrealizowana jest na bazie wewnętrznych przełączników HP BLc 4X QDR (interkonekty) i przełączników zewnętrznych SX6025 (topologia ***Fat Tree***).

Siec Eternet klastra zrealizowana jest na bazie wewnętrznych przełączników HP6120XG z wykorzystaniem VLAN'ów. Poprzez sieć Eternet realizowany jest styk z siecią zewnętrzną.

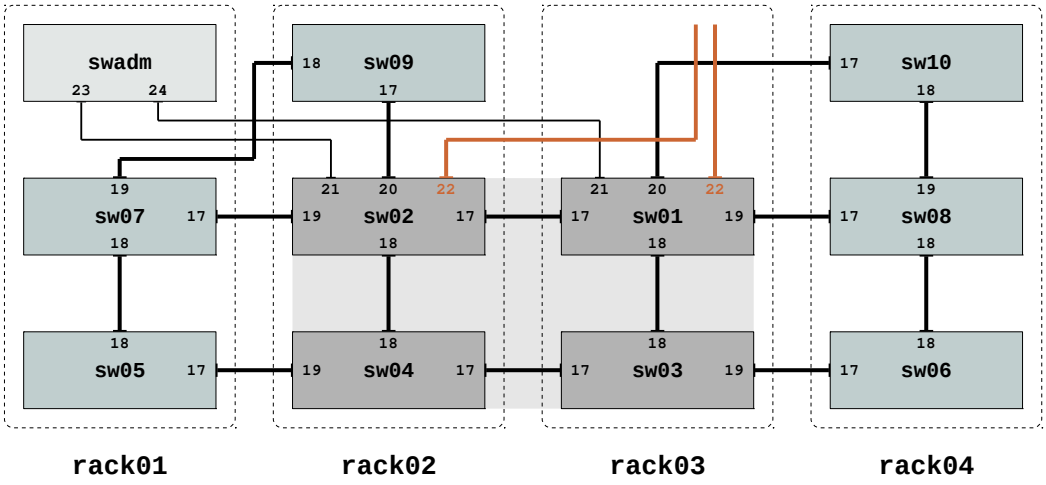


Architektura LAN klastra obliczeniowego (warstwa IP/VLAN)

Sieci *cmp*, *int* i *adm* są sieciami wewnętrznymi klastra utworzonymi z maską 255.255.252.0 (/22) zapewniające adresację dla 1022 komputerów (1022=4*256-2).

Dla każdego węzła obliczeniowego końcówka numeru IP interfejsu InfiniBand, interfejsu Eternet i iLO jest taka sama. Np dla *node001*: ib0=192.168.20.1, eth0=192.168.24.1, iLO=192.168.28.1.

Sieć Eternet oparta jest na dziesięciu wewnętrznych przełącznikach HP6120XG 24P (16 portów downlink 10Gb, 8 portów uplink 10Gb) umieszczonych po jednym w każdej skrzyni c7000.



Architektura połączeń Eternet klastra obliczeniowego (front)

Przełączniki *sw1*, *sw2*, *sw3* i *sw4* pełnią rolę przełączników centralnych i są połączone czterema linkami 10Gb. Z przełączników *sw1* i *sw2* realizowane jest wyjście do sieci zewnętrznej (porty 22 realizują VLAN tagging dla VLAN082).

Przełącznik *swadm* posiada 24 porty o prędkości 1Gb do których podłączone są interfejsy zarządzające skrzyniami C7000 *oa[01-10]*, macierzą HP 3PAR 7400 oraz *mpdp*.

Przed przyłączeniem do sieci zewnętrznej wszystkie przełączniki muszą mieć aktywny protokół ***Spanning Tree*** (domyślnie aktywny na ***sw[01-10]***, nieaktywny na ***swadm***).

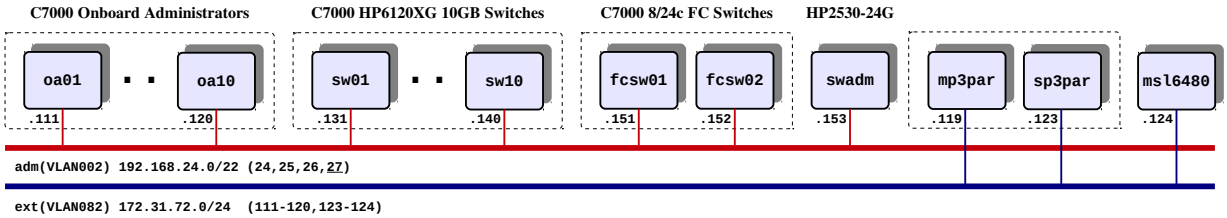
Sieć eternet wykorzystywana jest do:

- bezdyskowego uruchamiania WO
- administracji
- monitoringu
- wykonywania kopii zapasowych

Autentykacja i autoryzacja użytkowników na węzle *hn* (head node) realizowana jest z użyciem serwera ***sssd*** w połączeniu z serwerem AD IMGW.

Autentykacja użytkowników na węzłach obliczeniowych (WO) relizowana jest z użyciem mechanizmów ***SSH*** a autoryzacja w oparciu o serwer ***sssd*** w połączeniu z serwerem AD IMGW.

Do sieci administracyjnej oprócz HP iLO4 Węzłów Obliczeniowych podłączone są interfejsy administracyjne pozostałych urządzeń:



Adresacja interfejsów administracyjnych pozostałych urządzeń

3.2 Konfiguracja przełączników sw[01-10]

Konfiguracja przełącznika ***sw01***:

```
hostname "sw01"
interface 23
  disable
  lacp Active
exit
interface 24
  disable
  lacp Active
exit
ip routing
vlan 1
  name "adm"
  untagged 17-21,23-24
  ip address 192.168.24.254 255.255.252.0
  tagged 1-2
  no untagged 3-16,22
  exit
vlan 2
  name "int"
  untagged 3-16
  ip address 192.168.28.254 255.255.252.0
  tagged 1-2,17-21,23-24
  exit
vlan 82
  name "ext"
  untagged 1-2
  tagged 17-24
  exit
snmp-server community "public" unrestricted
spanning-tree
spanning-tree priority 15
oobm
  ip address dhcp-bootp
  exit
password manager
password operator
```

Konfiguracja przełącznika *sw02*:

```
hostname "sw02"
interface 23
  disable
  lacp Active
exit
interface 24
  disable
  lacp Active
exit
vlan 1
  name "adm"
  untagged 17-21,23-24
  no untagged 1-16,22
  no ip address
  exit
vlan 2
  name "int"
  untagged 1-16
  tagged 17-21,23-24
  no ip address
  exit
vlan 82
  name "ext"
  tagged 17-24
  no ip address
  exit
snmp-server community "public" unrestricted
spanning-tree
spanning-tree 22 path-cost 20000
spanning-tree priority 15
oobm
  ip address dhcp-bootp
  exit
password manager
password operator
```

Konfiguracja przełączników *sw[03-10]* jest taka sama (z dokładnością do nazwy przełącznika) i jest pokazana na podstawie konfiguracji przełącznika *sw03*:

```
hostname "sw03"
interface 23
  disable
  lacp Active
exit
interface 24
  disable
  lacp Active
exit
vlan 1
  name "adm"
  untagged 17-24
  no untagged 1-16
  no ip address
  exit
vlan 2
  name "int"
  untagged 1-16
  tagged 17-24
  no ip address
  exit
vlan 82
  name "ext"
  tagged 17-24
  no ip address
  exit
snmp-server community "public" unrestricted
spanning-tree
spanning-tree priority 15
oobm
  ip address dhcp-bootp
  exit
password manager
password operator
```

Do portu 20 przełącznika ***sw08*** podłączony jest serwer ***dp*** i w związku z tym domyślny VLAN na tym porcie został zmieniony na 2 (int).

3.3 Konfiguracja przełącznika swadm

Do przełącznika *swadm* podłączone są interfejsy zarządzające skrzyniami C7000 *oa[01-10]* (odpowiednio porty 1-10), macierzą HP 3PAR 7400 (porty 19-22) oraz *mpdp* (port 16).

Konfiguracja przełącznika ***swadm***:

```
hostname "swadm"
snmp-server community "public" unrestricted
vlan 1
  name "adm"
  no untagged 19-22
  untagged 1-18,23-28
  ip address 192.168.27.153 255.255.252.0
  exit
vlan 2
  name "int"
  tagged 23-24
  no ip address
  exit
vlan 82
  name "ext"
  untagged 19-22
  tagged 1,23-24
  no ip address
  exit
spanning-tree
spanning-tree priority 15
no tftp server
no dhcp config-file-update
password manager
password operator
```

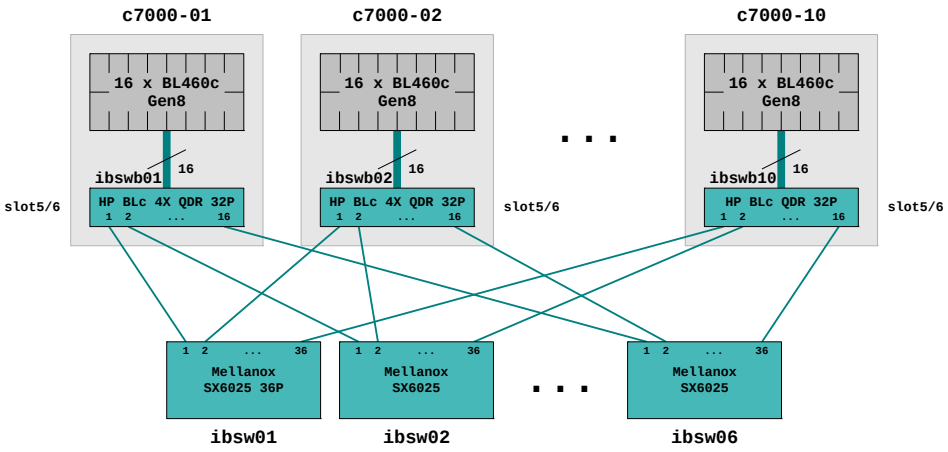
Rozdział 4: Sieć IB

4.1 Architektura

Sieć IB oparta jest na dziesięciu wewnętrznych przełącznikach HP BLc 4X QDR 32P (16 portów downlink, 16 portów uplink) umieszczonych po jednym w każdej skrzyni c7000 oraz sześciu zewnętrznych przełącznikach SX6025 36P.

Na potrzeby niezarządzalnych przełączników InfiniBand, bezpośrednio na serwerach *adm1* i *adm2* uruchomiono oprogramowanie *OpenSM* (będące elementem pakietu *OFED* - OpenFabrics Enterprise Distribution) zapewniające funkcjonalność Subnet Managera.

Sieć IB zrealizowano w topologii *Fat Tree*:



Architektura połączeń InfiniBand klastra obliczeniowego

Sieć InfiniBand wykorzystywana jest do:

- montowania głównego systemu plików (*root filesystem*) przez WO z WSP
- montowania dodatkowych systemów plików przez WO z WSP
- wymiany ruchu obliczeniowego (np. MPI)
- wykonywania kopii zapasowych

4.2 Procedury administracyjne IB

W celu wyświetlenia statusu karty IB na węźle należy wykonać:

```
adm1# ibstatus
Infiniband device 'mlx4_0' port 1 status:
     default gid:       fe80:0000:0000:0000:0002:c903:0041:6d11
     base lid:          0x1
     sm lid:            0x8
     state:             4: ACTIVE
     phys state:        5: LinkUp
     rate:              40 Gb/sec (4X QDR)
     link_layer:        InfiniBand

Infiniband device 'mlx4_0' port 2 status:
     default gid:       fe80:0000:0000:0000:0002:c903:0041:6d12
     base lid:          0x0
     sm lid:            0x0
     state:             1: DOWN
     phys state:        2: Polling
     rate:              10 Gb/sec (4X)
     link_layer:        InfiniBand
```

W celu wyświetlenia informacji o Subnet Managerze należy wykonać:

```
adm1# sminfo
sminfo: sm lid 8 sm guid 0x2c90300416cf1, activity count 11890408 \
     priority 0 state 3 SMINFO_MASTER
adm1# smpquery nd 8
Node Description:.....................adm2 HCA-1
```

W celu identyfikacji wszystkich przełączników IB należy wykonać:

```
adm1# ibnetdiscover -S
adm1# ibswitches
Switch: 0xf4521403007e4e10 ports 36 "SwitchX -  Mellanox ..." base port 0 lid 17 lmc 0
Switch: 0x0002c902004a6e88 ports 32 "Infiniscale-IV Mellanox ..." base port 0 lid 96 lmc 0
Switch: 0x0002c902004a8938 ports 32 "Infiniscale-IV Mellanox ..." base port 0 lid 43 lmc 0
Switch: 0x0002c902004a8940 ports 32 "Infiniscale-IV Mellanox ..." base port 0 lid 97 lmc 0
Switch: 0x0002c902004a8968 ports 32 "Infiniscale-IV Mellanox ..." base port 0 lid 44 lmc 0
Switch: 0x0002c902004a4df8 ports 32 "Infiniscale-IV Mellanox ..." base port 0 lid 95 lmc 0
Switch: 0x0002c902004a8948 ports 32 "Infiniscale-IV Mellanox ..." base port 0 lid 18 lmc 0
Switch: 0x0002c902004a6e28 ports 32 "Infiniscale-IV Mellanox ..." base port 0 lid 42 lmc 0
Switch: 0x0002c902004a8978 ports 32 "Infiniscale-IV Mellanox ..." base port 0 lid 4 lmc 0
Switch: 0x0002c902004a6e90 ports 32 "Infiniscale-IV Mellanox ..." base port 0 lid 41 lmc 0
Switch: 0xf4521403007e4d90 ports 36 "SwitchX -  Mellanox ..." base port 0 lid 19 lmc 0
Switch: 0xf4521403007e4600 ports 36 "SwitchX -  Mellanox ..." base port 0 lid 2 lmc 0
Switch: 0xf4521403007e4680 ports 36 "SwitchX -  Mellanox ..." base port 0 lid 5 lmc 0
Switch: 0xf4521403007e4c90 ports 36 "SwitchX -  Mellanox ..." base port 0 lid 7 lmc 0
Switch: 0xf4521403007e4a80 ports 36 "SwitchX -  Mellanox ..." base port 0 lid 6 lmc 0
Switch: 0x0002c902004a6e48 ports 32 "Infiniscale-IV Mellanox ..." base port 0 lid 3 lmc 0
```

W celu identyfikacji wszystkich HCA należy wykonać:

```
adm1# ibnetdiscover -H
adm1# ibhosts
Ca: 0x24be05ffff8c6e40 ports 2 "node139 mlx4_0"
Ca: 0x24be05ffff8c9e60 ports 2 "node138 mlx4_0"
Ca: 0x24be05ffff8cce90 ports 2 "node137 mlx4_0"
...
Ca: 0x24be05ffff8cfe20 ports 2 "node003 mlx4_0"
Ca: 0x0002c90300417290 ports 2 "node002 mlx4_0"
Ca: 0x24be05ffff8cfe20 ports 2 "node001 mlx4_0"
Ca: 0x0002c90300416d60 ports 2 "cfs4 mlx4_0"
Ca: 0x0002c90300416eb0 ports 2 "cfs3 mlx4_0"
Ca: 0x0002c90300416d00 ports 2 "cfs2 mlx4_0"
Ca: 0x0002c90300416f10 ports 2 "cfs1 mlx4_0"
Ca: 0x0002c90300416cf0 ports 2 "adm2 HCA-1"
Ca: 0x0002c90300416d10 ports 2 "adm1 HCA-1"
Ca: 0x24be05ffff9a1800 ports 2 "dp mlx4_0"
```

W celu wyświetlenia topologi IB należy wykonać:

```
adm1# iblinkinfo
adm1# ibnetdiscover
...
vendid=0x2c9
devid=0xbd36
sysimgguid=0x2c902004a6e48
switchguid=0x2c902004a6e48(2c902004a6e48) #
Switch 32 "S-0002c902004a6e48"  # "Infiniscale-IV Mellanox" base port 0 lid 3 lmc 0
[1]  "S-f4521403007e4a80"[1]  # "SwitchX -  Mellanox Technologies" lid 6 4xQDR
[2]  "S-f4521403007e4a80"[2]  # "SwitchX -  Mellanox Technologies" lid 6 4xQDR
[3]  "S-f4521403007e4a80"[3]  # "SwitchX -  Mellanox Technologies" lid 6 4xQDR
[4]  "S-f4521403007e4c90"[1]  # "SwitchX -  Mellanox Technologies" lid 7 4xQDR
[5]  "S-f4521403007e4c90"[2]  # "SwitchX -  Mellanox Technologies" lid 7 4xQDR
[6]  "S-f4521403007e4c90"[3]  # "SwitchX -  Mellanox Technologies" lid 7 4xQDR
[7]  "S-f4521403007e4680"[1]  # "SwitchX -  Mellanox Technologies" lid 5 4xQDR
[8]  "S-f4521403007e4680"[2]  # "SwitchX -  Mellanox Technologies" lid 5 4xQDR
[9]  "S-f4521403007e4680"[3]  # "SwitchX -  Mellanox Technologies" lid 5 4xQDR
[10] "S-f4521403007e4600"[1]  # "SwitchX -  Mellanox Technologies" lid 2 4xQDR
[11] "S-f4521403007e4600"[2]  # "SwitchX -  Mellanox Technologies" lid 2 4xQDR
[12] "S-f4521403007e4600"[3]  # "SwitchX -  Mellanox Technologies" lid 2 4xQDR
[15] "H-24be05ffff9a1800"[1](24be05ffff9a1801) # "dp mlx4_0" lid 163 4xQDR
[16] "S-f4521403007e4d90"[35] # "SwitchX -  Mellanox Technologies" lid 19 4xQDR
[17] "H-0002c90300416d10"[1](2c90300416d11)    # "adm1 HCA-1" lid 1 4xQDR
[18] "H-0002c90300416cf0"[1](2c90300416cf1)    # "adm2 HCA-1" lid 8 4xQDR
[19] "H-0002c90300416f10"[1](2c90300416f11)    # "cfs1 mlx4_0" lid 9 4xQDR
[20] "H-0002c90300416d00"[1](2c90300416d01)    # "cfs2 mlx4_0" lid 10 4xQDR
[21] "H-0002c90300416eb0"[1](2c90300416eb1)    # "cfs3 mlx4_0" lid 11 4xQDR
[22] "H-0002c90300416d60"[1](2c90300416d61)    # "cfs4 mlx4_0" lid 12 4xQDR
[23] "H-24be05ffff8c4f50"[1](24be05ffff8c4f51) # "node001 mlx4_0" lid 13 4xQDR
[24] "H-0002c90300417290"[1](2c90300417291)    # "node002 mlx4_0" lid 161 4xQDR
[25] "H-24be05ffff8cfe20"[1](24be05ffff8cfe21) # "node003 mlx4_0" lid 159 4xQDR
[26] "H-0002c90300417710"[1](2c90300417711)    # "node004 mlx4_0" lid 48 4xQDR
...
vendid=0x2c9
devid=0x1003
sysimgguid=0x2c90300416cf3
caguid=0x2c90300416cf0
Ca 2 "H-0002c90300416cf0" # "adm2 HCA-1"
[1](2c90300416cf1) "S-0002c902004a6e48"[18] \
      # lid 8 lmc 0 "Infiniscale-IV Mellanox Technologies" lid 3 4xQDR
```

W celu wyświetlenia trasy pomiędzy LID'ami (tu: ***adm2*** -> ***node062***) należy wykonać:

```
adm1# ibtracert src_lid dst_lid
adm1# ibtracert 8 20
From ca {0x0002c90300416cf0} portnum 1 lid 8-8 "adm2 HCA-1"
[1]  -> switch port {0x0002c902004a6e48}[18] lid 3-3 \
     "Infiniscale-IV Mellanox Technologies"
[5]  -> switch port {0xf4521403007e4c90}[2]  lid 7-7 \
     "SwitchX -  Mellanox Technologies"
[15] -> switch port {0x0002c902004a8948}[6]  lid 18-18 \
     "Infiniscale-IV Mellanox Technologies"
[28] -> ca port {0x0002c90300416ca1}[1]      lid 20-20 \
     "node062 mlx4_0"
To ca {0x0002c90300416ca0} portnum 1 lid 20-20 "node062 mlx4_0"
```

W celu wyświetlenia statusu portu węzła IB (tu: ***adm2 port 1***) należy wykonać:

```
adm1# ibportstate lid port
adm1# ibportstate 8 1
CA PortInfo:
# Port info: Lid 8 port 1
LinkState:.....................Active
PhysLinkState:.................LinkUp
Lid:...........................8
SMLid:.........................8
LMC:...........................0
LinkWidthSupported:............4X (IBA extension)
LinkWidthEnabled:..............4X
LinkWidthActive:...............4X
LinkSpeedSupported:............2.5 Gbps or 5.0 Gbps or 10.0 Gbps
LinkSpeedEnabled:..............2.5 Gbps or 5.0 Gbps or 10.0 Gbps
LinkSpeedActive:...............10.0 Gbps
LinkSpeedExtSupported:.........0
LinkSpeedExtEnabled:...........0
LinkSpeedExtActive:............No Extended Speed
Mkey:..........................<not displayed>
MkeyLeasePeriod:...............0
ProtectBits:...................0
...
```

W celu wyświetlenia liczników węzła należy wykonać:

```
adm1# perfquery lid port
adm1# perfquery 8 1
# Port counters: Lid 8 port 1 (CapMask: 0x1400)
PortSelect:......................1
CounterSelect:...................0x0000
SymbolErrorCounter:..............0
LinkErrorRecoveryCounter:........0
LinkDownedCounter:...............0
PortRcvErrors:...................0
PortRcvRemotePhysicalErrors:.....0
PortRcvSwitchRelayErrors:........0
PortXmitDiscards:................0
PortXmitConstraintErrors:........0
PortRcvConstraintErrors:.........0
CounterSelect2:..................0x00
LocalLinkIntegrityErrors:........0
ExcessiveBufferOverrunErrors:....0
VL15Dropped:.....................0
PortXmitData:....................2312702309
PortRcvData:.....................4294967295
PortXmitPkts:....................86487115
PortRcvPkts:.....................112944135
PortXmitWait:....................1
```

W celu wyświetlenia niskopoziomowych informacji o przełączniku należy wykonać:

```
adm1# smpquery switchinfo lid
adm1# smpquery switchinfo 2
# Switch info: Lid 2
LinearFdbCap:.....................49151
RandomFdbCap:.....................0
McastFdbCap:......................15872
LinearFdbTop:.....................163
DefPort:..........................0
DefMcastPrimPort:.................255
DefMcastNotPrimPort:..............255
LifeTime:.........................19
StateChange:......................0
OptSLtoVLMapping:.................1
LidsPerPort:......................0
PartEnforceCap:...................0
InboundPartEnf:...................0
OutboundPartEnf:..................0
FilterRawInbound:.................0
FilterRawOutbound:................0
EnhancedPort0:....................0
MulticastFDBTop:..................0xc09f
```

W celu wyświetlenia niskopoziomowych informacji o porcie należy wykonać:

```
adm1# smpquery portinfo lid port
adm1# smpquery portinfo 8 1
# Port info: Lid 8 port 1
Mkey:.............................<not displayed>
GidPrefix:........................0xfe80000000000000
Lid:..............................8
SMLid:............................8
CapMask:..........................0x2514868
                      IsTrapSupported
                      IsAutomaticMigrationSupported
                      IsSLMappingSupported
                      IsSystemImageGUIDsupported
                      IsExtendedSpeedsSupported
                      IsCommunicatonManagementSupported
                      IsVendorClassSupported
                      IsCapabilityMaskNoticeSupported
                      IsClientRegistrationSupported
DiagCode:.........................0x0000
MkeyLeasePeriod:..................0
LocalPort:........................1
LinkWidthEnabled:.................4X
LinkWidthSupported:...............4X (IBA extension)
LinkWidthActive:..................4X
LinkSpeedSupported:...............2.5 Gbps or 5.0 Gbps or 10.0 Gbps
LinkState:........................Active
PhysLinkState:....................LinkUp
LinkDownDefState:.................Polling
ProtectBits:......................0
LMC:..............................0
LinkSpeedActive:..................10.0 Gbps
LinkSpeedEnabled:.................2.5 Gbps or 5.0 Gbps or 10.0 Gbps
NeighborMTU:......................4096
SMSL:.............................0
VLCap:............................VL0-7
InitType:.........................0x00
VLHighLimit:......................4
VLArbHighCap:.....................8
VLArbLowCap:......................8
```

```
InitReply:.....................0x00
MtuCap:........................4096
VLStallCount:..................0
HoqLife:.......................31
OperVLs:.......................VL0-7
PartEnforceInb:................0
PartEnforceOutb:...............0
FilterRawInb:..................0
FilterRawOutb:.................0
MkeyViolations:................0
PkeyViolations:................0
QkeyViolations:................0
GuidCap:.......................128
ClientReregister:..............0
McastPkeyTrapSuppressionEnabled:.0
SubnetTimeout:.................18
RespTimeVal:...................16
LocalPhysErr:..................8
OverrunErr:....................8
MaxCreditHint:.................0
RoundTrip:.....................0
CapabilityMask2:...............0x0000
LinkSpeedExtActive:............No Extended Speed
LinkSpeedExtSupported:.........0
LinkSpeedExtEnabled:...........0
```

W celu wykonania pełnej diagnostyki sieci IB należy wykonać:

```
adm1# ibdiagnet -pc       (skasowanie liczników)
adm1# ibdiagnet -r        (generacja raportu)
adm1# more /var/tmp/ibdiagnet2/ibdiagnet2.log
```

W celu wypisania błędów sieci IB należy wykonać:

```
adm1# ibqueryerrors -k    (skasowanie aktualnych błędów)
adm1# ibqueryerrors       (wypisanie nowych błędów)
```

W celu wykonania szybkiej diagnostyki portów IB należy wykonać:

```
adm1# ibcheckstate -v
```

W celu wykonania resetu portu IB (tu: ***adm2*** port 1) należy wykonać:

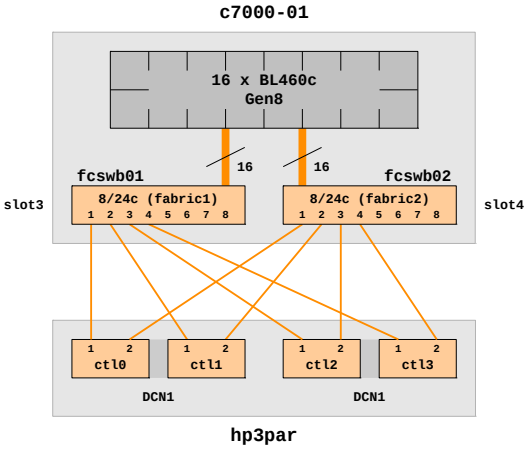
```
adm1# smpquery nd 8
Node Description:.....................adm2 HCA-1
adm1# ibportstate 8 1 reset
Initial CA PortInfo:
# Port info: Lid 8 port 1
LinkState:.......................Active
PhysLinkState:...................LinkUp
Lid:.............................8
SMLid:...........................8
LMC:.............................0
LinkWidthSupported:..............4X (IBA extension)
LinkWidthEnabled:................4X
LinkWidthActive:.................4X
LinkSpeedSupported:..............2.5 Gbps or 5.0 Gbps or 10.0 Gbps
LinkSpeedEnabled:................2.5 Gbps or 5.0 Gbps or 10.0 Gbps
LinkSpeedActive:.................10.0 Gbps
...

After PortInfo set:
# Port info: Lid 8 port 1
LinkState:.......................Active
PhysLinkState:...................LinkUp
Lid:.............................8
SMLid:...........................8
LMC:.............................0
LinkWidthSupported:..............4X (IBA extension)
LinkWidthEnabled:................4X
LinkWidthActive:.................4X
LinkSpeedSupported:..............2.5 Gbps or 5.0 Gbps or 10.0 Gbps
LinkSpeedEnabled:................2.5 Gbps or 5.0 Gbps or 10.0 Gbps
LinkSpeedActive:.................10.0 Gbps
...
adm2# dmesg
mlx4_core 0000:21:00.0: mlx4_ib: Port 1 logical link is down
...
mlx4_core 0000:21:00.0: mlx4_ib: Port 1 logical link is up
...
```

Rozdział 5: Sieć SAN

5.1 Architektura

Sieć SAN oparta jest na czterech wewnętrznych przełącznikach Brocade 8/24c (16 portów downlink, 8 portów uplink) umieszczonych w skrzyni *c7000-01*. Przełączniki tworzą dwa niezależne SANy *fabric1* i *fabric2*.



Architektura połączeń SAN klastra obliczeniowego

Dostęp do sieci SAN posiadają serwery kasetowe BL460c Gen8 w specyfikacji administracyjnej (dodatkowa karta FC) umieszczone w skrzyni *c7000-01*.

Macierz HP 3PAR 7400 przyłączona jest do sieci FC czterema kontrolerami (na każdym kontrolerze wykorzystano 4 porty FC).

Zonning na przełącznikach zdefiniowano w oparciu o adresy WWPN.

Sieć SAN wykorzystywana jest do:

- udostępniania zasobów macierzy HP 3PAR 7400 serwerom WSP
- wykonywania kopii zapasowych

5.2 Konfiguracja przełączników FC

Przełącznik ***fcswb01*** skonfigurowano następująco:

(konfiguracja nazwy przełącznika)

```
swd77:admin> switchname "fcswb01"
swd77:admin> reboot
```

(wstępna konfiguracja zoningu)

```
fcswb01:admin> zonecreate "zone1", "1,1; 1,2; 1,3; 1,4; 1,5; 1,6; \
    1,17; 1,18; 1,19; 1,20"
fcswb01:admin> zonecreate "zoneb", "1,21; 1,22; 1,23; 1,0"
fcswb01:admin> cfgcreate "fabric1", "zone1; zoneb"
fcswb01:admin> cfgenable "fabric1"
```

(modyfikacja konfiguracji zoningu)

```
fcswb01:admin> alicreate "dp", "50:01:43:80:24:2a:fc:98; \
    50:01:43:80:24:2a:fe:6c"
fcswb01:admin> zoneadd "zone1", "dp"
fcswb01:admin> cfgenable "fabric1"
```

Przełącznik ***fcswb02*** skonfigurowano następująco:

(konfiguracja nazwy przełącznika)

```
swd77:admin> switchname "fcswb02"
swd77:admin> reboot
```

(wstępna konfiguracja zoningu)

```
fcswb02:admin> zonecreate "zone1", "1,1; 1,2; 1,3; 1,4; 1,5; 1,6; \
    1,17; 1,18; 1,19; 1,20"
fcswb02:admin> zonecreate "zoneb", "1,21; 1,22; 1,23; 1,0"
fcswb02:admin> cfgcreate "fabric2", "zone1; zoneb"
fcswb02:admin> cfgenable "fabric2"
```

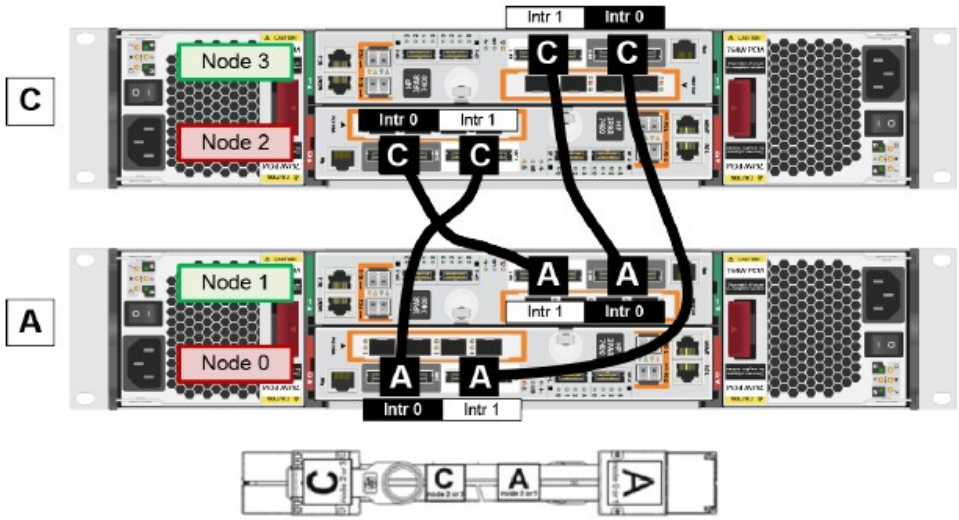
(modyfikacja konfiguracji zoningu)

```
fcswb02:admin> alicreate "dp", "50:01:43:80:24:2a:fc:9a; \
    50:01:43:80:24:2a:fe:6e"
fcswb02:admin> zoneadd "zone1", "dp"
fcswb02:admin> cfgenable "fabric2"
```

5.3 Konfiguracja macierzy HP 7400

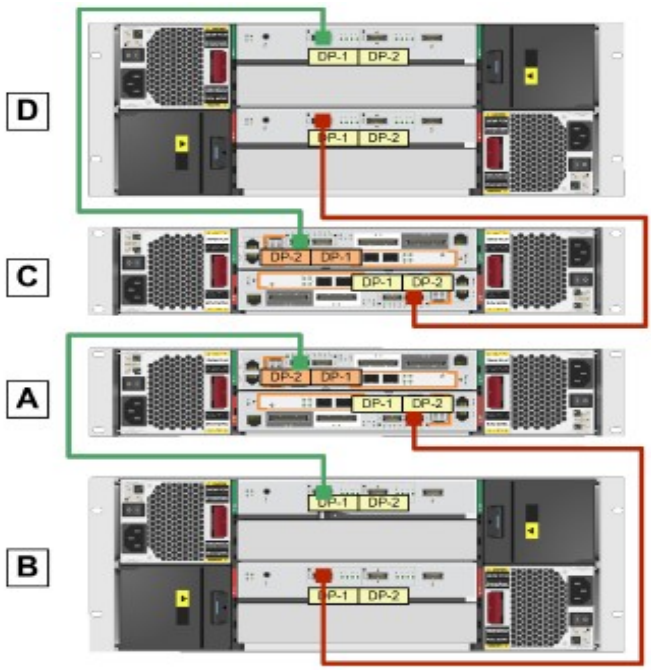
5.3.1 Połączenia fizyczne

Węzły macierzy połączone są czterema **kierunkowymi** kablami typu *interconnect*:



Architektura połączeń węzłów macierzy HP 7400

Półki macierzy połączone są z węzłami czterema kablami typu *SAS:*



Architektura połączeń półek dyskowych macierzy HP 7400

5.3.2 Licencjonowanie

Wgranie docelowej licencji:

```
hp3par cli% setlicense
...
Please enter the new license key below.  When finished, press enter on an
empty line.  If the key is entered by hand, note that characters other
than letters and numbers are ignored, and the key is not case sensitive.
60R3-0C1G-60R3-2C1G-60R3-0C9G-70R3-0C1G
60RK-0C07-KFY0-QGT0-C417-VWB1-WYD1-F2KS
KA7N-K0NB-M1S7-MR61-JHE9-KLCY-5BDW-TQ3X
W1LW-QVHY-N6LE-DE9S-WQSL-10GL-S1K8-Q32W
B26C-Z3WJ-DNLA-G1F3-A8BN-88HL-6407-6CGA
2WNB-VWN5-ZFQ4-F32D-D32H-BF5E-55WB-HGSA
500B-XFF1-ZQ0J-CR8H-1BAK-9RD8-YCMQ-CFT2
JXXW-MCMG-2ZNW-3J9H-JZM9-MRL1-JCKK-96AL
4MWE-75GS-NCND-FXM6-HWHJ-2T48-ZFAY-0R5R
Q0XZ-ARKL-1JTQ-TNNW-4B42-7LQX-674C-6LZ4
EMCF-LQ0R-VYWD-NH02-3MJM-0692-98EX-SLTG
WKM5-A6Q7-ZWQT-RRQ7-3YAH-Q7B4-QKVE-W899
ZR4A-YQ4J-9DZ8-NW3E-RF42-JHVB-6YNW-1QRV
05Q8-8HGY-KKTL-AZHL-GR9H-V9QR-Z0XH-7R17
<ENTER>
The system will be licensed for 68 disks instead of 8 disks.
The following features will be enabled:
Peer Motion                              (Expires August 03, 2014)
System Tuner                             (No expiration date)
Thin Conversion                          (No expiration date)
Thin Persistence                         (No expiration date)
Thin Provisioning (10240000G)            (No expiration date)
Virtual Copy                             (No expiration date)
VSS Provider for Microsoft Windows       (No expiration date)

Are these the expected changes? (yes/no) yes
License key successfully set.
```

5.3.3 Konfiguracja portów FC

Domyślnie porty FC kontrolerów macierzy HP 3PAR 7400 są skonfigurowane w trybie *loop*, aby podłączyć je do przełącznika FC należy zmienić ich konfigurację na *fabric*:

```
hp3par cli% controlport offline -f \
    0:1:1 0:1:2 1:1:1 1:1:2 2:1:1 2:1:2 3:1:1 3:1:2
hp3par cli% controlport config host -ct point -f \
    0:1:1 0:1:2 1:1:1 1:1:2 2:1:1 2:1:2 3:1:1 3:1:2
hp3par cli% controlport rst -f \
    0:1:1 0:1:2 1:1:1 1:1:2 2:1:1 2:1:2 3:1:1 3:1:2


hp3par cli% showport -c
N:S:P      Mode Device Pos Config       Topology  Rate Cls Mode_change
0:0:1 initiator  cage0   0  valid point_to_point 6Gbps n/a  prohibited
0:0:2 initiator  cage1   0  valid point_to_point 6Gbps n/a  prohibited
0:1:1    target    ---   -    ---         fabric 8Gbps   3     allowed
0:1:2    target    ---   -    ---         fabric 8Gbps   3     allowed
...
1:0:1 initiator  cage0   0  valid point_to_point 6Gbps n/a  prohibited
1:0:2 initiator  cage1   0  valid point_to_point 6Gbps n/a  prohibited
1:1:1    target    ---   -    ---         fabric 8Gbps   3     allowed
1:1:2    target    ---   -    ---         fabric 8Gbps   3     allowed
...
2:0:1 initiator  cage2   0  valid point_to_point 6Gbps n/a  prohibited
2:0:2 initiator  cage3   0  valid point_to_point 6Gbps n/a  prohibited
2:1:1    target    ---   -    ---         fabric 8Gbps   3     allowed
2:1:2    target    ---   -    ---         fabric 8Gbps   3     allowed
...
3:0:1 initiator  cage2   0  valid point_to_point 6Gbps n/a  prohibited
3:0:2 initiator  cage3   0  valid point_to_point 6Gbps n/a  prohibited
3:1:1    target    ---   -    ---         fabric 8Gbps   3     allowed
3:1:2    target    ---   -    ---         fabric 8Gbps   3     allowed
...
----------------------------------------------------------------------
   32
```

5.3.3 Konfiguracja hostów

Wszystkie serwery korzystające z macierzy pracują pod kontrolą systemu operacyjnego Linux i dlatego podczas tworzenia hosta należy wybrać typ systemu 1 (*-persona 1*).

Zdefiniowano następujące komputery:

(utworzenie definicji komputerów adm1-adm2 posiadających po jednej dwuportowej karcie FC)

```
hp3par cli% createhost -persona 1 adm1 \
     5001438026E8C340 5001438026E8C342
hp3par cli% createhost -persona 1 adm2 \
     5001438026E8A4F0 5001438026E8A4F2
```

(utworzenie definicji komputerów cfs1-cfs4 posiadających po jednej dwuportowej karcie FC)

```
hp3par cli% createhost -persona 1 cfs1 \
     5001438026E8B18C 5001438026E8B18E
hp3par cli% createhost -persona 1 cfs2 \
     5001438026E8ADE8 5001438026E8ADEA
hp3par cli% createhost -persona 1 cfs3 \
     5001438026E8AE78 5001438026E8AE7A
hp3par cli% createhost -persona 1 cfs4 \
     5001438026E8A504 5001438026E8A506
```

(utworzenie definicji komputera dp posiadającego dwie dwuportowe karty FC)

```
hp3par cli% createhost -persona 1 dp \
     50014380242AFC98 50014380242AFC9A \
     50014380242AFE6C 50014380242AFE6E
```

5.3.4 Konfiguracja wirtualnych woluminów

Grupa CPG *cpg01* została zdefiniowana na dyskach SSD, grupy *cpg02* i *cpg03* zostały zdefiniowane na dyskach NL.

Dla potrzeb uruchamiania serwerów fizycznych lub maszyn wirtualnych zostały zdefiniowane następujące wirtualne woluminy:

(utworzenie vv boot dla serwerów adm1-adm2)

```
hp3par cli% createvv -tpvv cpg01 adm1_disk01 100g
hp3par cli% createvv -tpvv cpg01 adm2_disk01 100g
```

(utworzenie vv boot dla serwerów cfs1-cfs4)

```
hp3par cli% createvv -tpvv cpg01 cfs1_disk01 20g
hp3par cli% createvv -tpvv cpg01 cfs2_disk01 20g
hp3par cli% createvv -tpvv cpg01 cfs3_disk01 20g
hp3par cli% createvv -tpvv cpg01 cfs4_disk01 20g
```

(utworzenie vv boot dla maszyn wirtualnych adm,mon,hn,wo1)

```
hp3par cli% createvv -tpvv cpg01 adm_disk01 100g
hp3par cli% createvv -tpvv cpg01 mon_disk01 100g
hp3par cli% createvv -tpvv cpg01  hn_disk01 100g
hp3par cli% createvv -tpvv cpg01 wo1_disk01  20g
```

(utworzenie vv boot dla serwera dp)

```
hp3par cli% createvv -tpvv cpg01  dp_disk01 100g
```

*(utworzenie vv quorum-disk dla klastrów **cl01-cl02**)*

```
hp3par cli% createvv cpg01 cl01_qdisk 1g
hp3par cli% createvv cpg01 cl02_qdisk 1g
```

Dla potrzeb systemu plików Lustre zostały zdefiniowane następujące woluminy:

(utworzenie vv dla serwera MGS systemu plików Lustre)

```
hp3par cli% createvv cpg01 mgs-mgt0000 1g
```

*(utworzenie vv dla systemu plików Lustre **fs1**)*

```
hp3par cli% createvv -tpvv cpg01 fs1-mdt0000 2g
hp3par cli% createvv -tpvv cpg01 fs1-ost0000 50g
hp3par cli% createvv -tpvv cpg01 fs1-ost0001 50g
hp3par cli% createvv -tpvv cpg01 fs1-ost0002 50g
hp3par cli% createvv -tpvv cpg01 fs1-ost0003 50g
```

*(utworzenie vv dla systemu plików Lustre **fs2**)*

```
hp3par cli% createvv -tpvv cpg01 fs2-mdt0000 400g
hp3par cli% createvv -tpvv cpg02 fs2-ost0000 8192g
hp3par cli% createvv -tpvv cpg02 fs2-ost0001 8192g
hp3par cli% createvv -tpvv cpg02 fs2-ost0002 8192g
hp3par cli% createvv -tpvv cpg02 fs2-ost0003 8192g
```

*(utworzenie vv dla systemu plików Lustre **fs3**)*

```
hp3par cli% createvv -tpvv cpg01 fs3-mdt0000 400g
hp3par cli% createvv -tpvv cpg03 fs3-ost0000 8192g
hp3par cli% createvv -tpvv cpg03 fs3-ost0001 8192g
hp3par cli% createvv -tpvv cpg03 fs3-ost0002 8192g
hp3par cli% createvv -tpvv cpg03 fs3-ost0003 8192g
```

5.3.5 Konfiguracja wirtualnych LUNów

Wirtualne LUNy to sposób prezentacji w macierzy HP 3PAR wirtualnych woluminów do zdefiniowanych hostów:

(zaprezentowanie vv boot serwerów adm1-adm2)

```
hp3par cli% createvlun adm1_disk01 1 adm1
hp3par cli% createvlun adm2_disk01 1 adm2
```

(zaprezentowanie vv boot serwerów cfs1-cfs4)

```
hp3par cli% createvlun cfs1_disk01 1 cfs1
hp3par cli% createvlun cfs2_disk01 1 cfs2
hp3par cli% createvlun cfs3_disk01 1 cfs3
hp3par cli% createvlun cfs4_disk01 1 cfs4
```

*(zaprezentowanie vv boot maszyn wirtualnych adm,mon,hn,wo1 do klastra **cl01**)*

```
hp3par cli% createvlun adm_disk01 11 adm1
hp3par cli% createvlun adm_disk01 11 adm2
hp3par cli% createvlun mon_disk01 12 adm1
hp3par cli% createvlun mon_disk01 12 adm2
hp3par cli% createvlun  hn_disk01 13 adm1
hp3par cli% createvlun  hn_disk01 13 adm2
hp3par cli% createvlun wo1_disk01 14 adm1
hp3par cli% createvlun wo1_disk01 14 adm2
```

(zaprezentowanie dysku boot serwera dp)

```
hp3par cli% createvlun dp_disk01 1 dp
```

*(zaprezentowanie vv quorum-disk klastrów **cl01**-**cl02**)*

```
hp3par cli% createvlun cl01_qdisk 2 adm1
hp3par cli% createvlun cl01_qdisk 2 adm2
hp3par cli% createvlun cl02_qdisk 2 cfs1
hp3par cli% createvlun cl02_qdisk 2 cfs2
hp3par cli% createvlun cl02_qdisk 2 cfs3
hp3par cli% createvlun cl02_qdisk 2 cfs4
```

*(zaprezentowanie vv dla serwera MGS systemu plików Lustre do klastra **cl02**)*

```
hp3par cli% createvlun mgs-mgt0000 10 cfs1
hp3par cli% createvlun mgs-mgt0000 10 cfs2
hp3par cli% createvlun mgs-mgt0000 10 cfs3
hp3par cli% createvlun mgs-mgt0000 10 cfs4
```

*(zaprezentowanie vv dla systemu plików Lustre **fs1** do klastra **cl02**)*

```
hp3par cli% createvlun fs1-mdt0000 11 cfs1
hp3par cli% createvlun fs1-mdt0000 11 cfs2
hp3par cli% createvlun fs1-mdt0000 11 cfs3
hp3par cli% createvlun fs1-mdt0000 11 cfs4
hp3par cli% createvlun fs1-ost0000 12 cfs1
hp3par cli% createvlun fs1-ost0000 12 cfs2
hp3par cli% createvlun fs1-ost0000 12 cfs3
hp3par cli% createvlun fs1-ost0000 12 cfs4
hp3par cli% createvlun fs1-ost0001 13 cfs1
hp3par cli% createvlun fs1-ost0001 13 cfs2
hp3par cli% createvlun fs1-ost0001 13 cfs3
hp3par cli% createvlun fs1-ost0001 13 cfs4
hp3par cli% createvlun fs1-ost0002 14 cfs1
hp3par cli% createvlun fs1-ost0002 14 cfs2
hp3par cli% createvlun fs1-ost0002 14 cfs3
hp3par cli% createvlun fs1-ost0002 14 cfs4
hp3par cli% createvlun fs1-ost0003 15 cfs1
hp3par cli% createvlun fs1-ost0003 15 cfs2
hp3par cli% createvlun fs1-ost0003 15 cfs3
hp3par cli% createvlun fs1-ost0003 15 cfs4
```

*(zaprezentowanie vv dla systemu plików Lustre **fs2** do klastra **cl02**)*

```
hp3par cli% createvlun fs2-mdt0000 21 cfs1
hp3par cli% createvlun fs2-mdt0000 21 cfs2
hp3par cli% createvlun fs2-mdt0000 21 cfs3
hp3par cli% createvlun fs2-mdt0000 21 cfs4
hp3par cli% createvlun fs2-ost0000 22 cfs1
hp3par cli% createvlun fs2-ost0000 22 cfs2
hp3par cli% createvlun fs2-ost0000 22 cfs3
hp3par cli% createvlun fs2-ost0000 22 cfs4
```

```
hp3par cli% createvlun fs2-ost0001 23 cfs1
hp3par cli% createvlun fs2-ost0001 23 cfs2
hp3par cli% createvlun fs2-ost0001 23 cfs3
hp3par cli% createvlun fs2-ost0001 23 cfs4
hp3par cli% createvlun fs2-ost0002 24 cfs1
hp3par cli% createvlun fs2-ost0002 24 cfs2
hp3par cli% createvlun fs2-ost0002 24 cfs3
hp3par cli% createvlun fs2-ost0002 24 cfs4
hp3par cli% createvlun fs2-ost0003 25 cfs1
hp3par cli% createvlun fs2-ost0003 25 cfs2
hp3par cli% createvlun fs2-ost0003 25 cfs3
hp3par cli% createvlun fs2-ost0003 25 cfs4
```

*(zaprezentowanie vv dla systemu plików Lustre **fs3** do klastra **cl02**)*

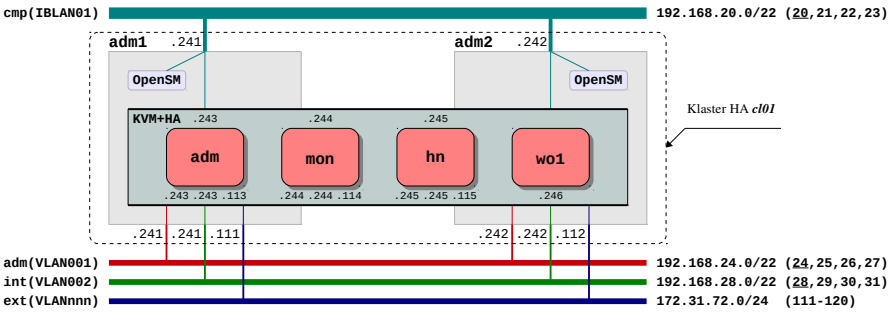
```
hp3par cli% createvlun fs3-mdt0000 31 cfs1
hp3par cli% createvlun fs3-mdt0000 31 cfs2
hp3par cli% createvlun fs3-mdt0000 31 cfs3
hp3par cli% createvlun fs3-mdt0000 31 cfs4
hp3par cli% createvlun fs3-ost0000 32 cfs1
hp3par cli% createvlun fs3-ost0000 32 cfs2
hp3par cli% createvlun fs3-ost0000 32 cfs3
hp3par cli% createvlun fs3-ost0000 32 cfs4
hp3par cli% createvlun fs3-ost0001 33 cfs1
hp3par cli% createvlun fs3-ost0001 33 cfs2
hp3par cli% createvlun fs3-ost0001 33 cfs3
hp3par cli% createvlun fs3-ost0001 33 cfs4
hp3par cli% createvlun fs3-ost0002 34 cfs1
hp3par cli% createvlun fs3-ost0002 34 cfs2
hp3par cli% createvlun fs3-ost0002 34 cfs3
hp3par cli% createvlun fs3-ost0002 34 cfs4
hp3par cli% createvlun fs3-ost0003 35 cfs1
hp3par cli% createvlun fs3-ost0003 35 cfs2
hp3par cli% createvlun fs3-ost0003 35 cfs3
hp3par cli% createvlun fs3-ost0003 35 cfs4
```

Rozdział 6: Serwisy klastra

6.1 Architektura

Serwisy klastra uruchamiane są w zwirtualizowanym środowisku klastra HA na komputerach *adm1* i *adm2* – ADM (wykorzystano wbudowaną w system ScientificLinux 6.4 x86_64 wirtualizację KVM i oprogramowanie klastra HA).

Serwery ADM nie posiadają dysków wewnętrznych i uruchamiają się z LUNów wystawionych z macierzy HP 3PAR 7400. Po uruchomieniu formują dwuwęzłowy klastr HA *cl01*.



Architektura serwisów klastra obliczeniowego

Bezpośrednio na komputerach *adm1* i *adm2* uruchomiane jest oprogramowanie OpenSM świadczące usługę Subnet Manager dla przełączników InfiniBand.

Pozostałe serwisy uruchomiane są na maszynach wirtualnych (VM):

VM	*Węzeł domyślny*	*Funkcja*	*Usługi/oprogramowanie*
adm	adm1	Serwer administracyjny	PDSH, DHCP, TFTP, DNS
mon	adm2	Monitoring klastra	PDSH, HP SIM, Ganglia
hn	adm1	Węzeł dostępowy, manager zadań	PDSH, Torque
wo1	adm2	Pierwszy wzorzec *rootfs* dla WO	-

6.2 Konfiguracja klastrów HA

W środowisku klastra Cosmo skonfigurowano dwa klastry HA:

- **cl01** – klaster administracyjny (ADM) oparty na serwerach ***adm[1-2]***
- **cl02** – klaster Lustre (CFS) oparty na serwerach ***cfs[1-4]***

6.2.1 Konfiguracja klastra cl01

Instalacja oprogramowania HA (na każdym węźle)

```
# yum grouplist -v | grep -i "high availability"
   High Availability (ha)
   High Availability Management (ha-management)
# yum grouplist -v | grep -i resilient
   Resilient Storage (resilient-storage)

# yum groupinstall ha
# yum groupinstall ha-management
# yum groupinstall resilient-storage
```

Wstępna konfiguracja klastra

(zablokowanie ACPI Soft-Off (dla fencingu))

```
# service acpid stop
# chkconfig acpid off
```

(konfiguracja ricci (Remote Cluster and Storage Management System))

```
# chkconfig ricci on
# service ricci start
# passwd ricci
```

(stworzenie nowej konfiguracji klastra)

```
adm1# ccs -h adm1 --createcluster cl01
adm1.cosmo.local password: <xxxxx>
```

(dodanie węzłów do konfiguracji klastra)

```
adm1# ccs -h adm1 --addnode adm1.cosmo.local --nodeid 1
adm1# ccs -h adm1 --addnode adm2.cosmo.local --nodeid 2
```

(synchronizacja konfiguracji)

```
 adm1# ccs -h adm1 --checkconf
 adm2.cosmo.local password: <xxxxx>
 Node: adm2.cosmo.local does not match
adm1# ccs -h adm1 --sync
adm1# ccs -h adm1 --checkconf
 All nodes in sync.
```

(uruchomienie i aktywacja serwisów klastra na wszystkich węzłach)

```
 adm1# ccs -h adm1 --startall
 Started adm1.cosmo.local
 Started adm2.cosmo.local
```

(wyspisanie statusu klastra)

```
 adm1# clustat
 Cluster Status for cl01 @ Sat Feb  8 23:07:25 2014
 Member Status: Quorate

 Member Name                                ID   Status
  ------ ----                                ---- ------
  adm1.cosmo.local                             1 Online, Local
  adm2.cosmo.local                             2 Online
```

Konfiguracja fencingu

*(utworzenie użytkownika fencingu na HP iLO **mpadm[1-2]**)*

```
</>hpiLO-> cd /map1/accounts1
</map1/accounts1>hpiLO-> create username=admin password=xxxxx
</map1/accounts1>hpiLO-> set admin \
     group=admin,config,oemhp_rc,oemhp_power,oemhp_vm
```

(konfiguracja fence devices)

```
adm1# ccs -h adm1 --addfencedev mpadm1 \
     agent=fence_ilo3 ipaddr=172.31.72.117 \
     login=admin passwd=xxxxx verbose=1
adm1# ccs -h adm1 --addfencedev mpadm2 \
     agent=fence_ilo3 ipaddr=172.31.72.118 \
     login=admin passwd=xxxxx verbose=1
```

(konfiguracja fence method dla każdego węzła)

```
adm1# ccs -h adm1 --addmethod ipmi adm1.cosmo.local
adm1# ccs -h adm1 --addmethod ipmi adm2.cosmo.local
```

(konfiguracja fence instance dla fence method dla każdego węzła)

```
adm1# ccs -h adm1 --addfenceinst mpadm1 adm1.cosmo.local ipmi
adm1# ccs -h adm1 --addfenceinst mpadm2 adm2.cosmo.local ipmi
```

(synchronizacja i aktywacja konfiguracji)

```
adm1# ccs -h adm1 --checkconf
adm1# ccs -h adm1 --sync --activate
adm1# ccs -h adm1 --checkconf
```

Konfiguracja quorum disk

(inicjalizacja i wylistowanie informacji o quorum disk)

```
adm1# mkqdisk -c /dev/mapper/cl01_qdisk -l cl01qdisk
adm1# mkqdisk -L -d
mkqdisk v3.0.12.1
/dev/block/253:5:
/dev/disk/by-id/dm-name-cl01_qdisk:
/dev/disk/by-id/dm-uuid-mpath-360002ac0000000000000014000053c6:
/dev/dm-5:
/dev/mapper/cl01_qdisk:
        Magic:                eb7a62c2
        Label:                cl01qdisk
        Created:              Sat Feb  8 23:17:11 2014
        Host:                 adm1
        Kernel Sector Size:   512
        Recorded Sector Size: 512
```

(konfiguracja quorum disk w klastrze)

```
adm1# ccs -h adm1 --setquorumd interval=2 label=cl01qdisk \
      tko=5 votes=1
adm1# ccs -h adm1 --settotem token=33000
adm1# ccs -h adm1 --sync -activate
```

(aktywacja quorum disk)

```
adm1# ccs -h adm1 --stopall
adm1# ccs -h adm1 --startall
adm1# ccs -h adm1 --setcman expected_votes="3" two_node="0"
adm1# ccs -h adm1 --sync --activate
adm1# clustat
Member Name                             ID   Status
 ------ ----                             ---- ------
 adm1.cosmo.local                           1 Online, Local
 adm2.cosmo.local                           2 Online
 /dev/block/253:5                           0 Online, Quorum Disk
```

Konfiguracja failover domains

*(utworzenie domeny **dom1**)*

```
adm1# ccs -h adm1 --addfailoverdomain dom1 \
      ordered nofailback restricted
adm1# ccs -h adm1 --addfailoverdomainnode dom1 adm1.cosmo.local 1
adm1# ccs -h adm1 --addfailoverdomainnode dom1 adm2.cosmo.local 2
```

*(utworzenie domeny **dom2**)*

```
adm1# ccs -h adm1 --addfailoverdomain dom2 \
      ordered nofailback restricted
adm1# ccs -h adm1 --addfailoverdomainnode dom2 adm2.cosmo.local 1
adm1# ccs -h adm1 --addfailoverdomainnode dom2 adm1.cosmo.local 2
```

(synchronizacja i aktywacja konfiguracji)

```
adm1# ccs -h adm1 --checkconf
adm1# ccs -h adm1 --sync --activate
adm1# ccs -h adm1 --checkconf
```

6.2.2 Konfiguracja klastra cl02

Instalacja oprogramowania HA (na każdym węźle)

```
# yum grouplist -v | grep -i "high availability"
   High Availability (ha)
   High Availability Management (ha-management)
# yum grouplist -v | grep -i resilient
   Resilient Storage (resilient-storage)

# yum groupinstall ha
# yum groupinstall ha-management
# yum groupinstall resilient-storage
```

Wstępna konfiguracja klastra

(zablokowanie ACPI Soft-Off (dla fencingu))

```
# service acpid stop
# chkconfig acpid off
```

(konfiguracja ricci (Remote Cluster and Storage Management System))

```
# chkconfig ricci on
# service ricci start
# passwd ricci
```

(stworzenie nowej konfiguracji klastra)

```
cfs1# ccs -h cfs1 --createcluster cl02
cfs1.cosmo.local password: <xxxxx>
```

(dodanie węzłów do konfiguracji klastra)

```
cfs1# ccs -h cfs1 --addnode cfs1.cosmo.local --nodeid 1
cfs1# ccs -h cfs1 --addnode cfs2.cosmo.local --nodeid 2
cfs1# ccs -h cfs1 --addnode cfs3.cosmo.local --nodeid 3
cfs1# ccs -h cfs1 --addnode cfs4.cosmo.local --nodeid 4
```

(synchronizacja konfiguracji)

```
cfs1# ccs -h cfs1 --checkconf
cfs2.cosmo.local password: <xxxxx>
Node: cfs2.cosmo.local does not match
Node: cfs3.cosmo.local does not match
Node: cfs4.cosmo.local does not match
cfs1# ccs -h cfs1 --sync
cfs1# ccs -h cfs1 --checkconf
All nodes in sync.
```

(uruchomienie i aktywacja serwisów klastra na wszystkich węzłach)

```
cfs1# ccs -h cfs1 --startall
Started cfs4.cosmo.local
Started cfs2.cosmo.local
Started cfs3.cosmo.local
Started cfs1.cosmo.local
```

(wyspisanie statusu klastra)

```
cfs1# clustat
Cluster Status for cl02 @ Sat Feb  8 16:07:27 2014
Member Status: Quorate

 Member Name                                ID   Status
  ------ ----                                ---- ------
  cfs1.cosmo.local                              1 Online, Local
  cfs2.cosmo.local                              2 Online
  cfs3.cosmo.local                              3 Online
  cfs4.cosmo.local                              4 Online
```

Konfiguracja fencingu

*(utworzenie użytkownika fencingu na HP iLO **mpcfs[1-4]**)*

```
</>hpiLO-> cd /map1/accounts1
</map1/accounts1>hpiLO-> create username=admin password=xxxxx
</map1/accounts1>hpiLO-> set admin \
     group=admin,config,oemhp_rc,oemhp_power,oemhp_vm
```

(konfiguracja fence devices)

```
cfs1# ccs -h cfs1 --addfencedev mpcfs1 agent=fence_ilo3 \
     ipaddr=192.168.24.231 login=admin passwd=xxxxx verbose=1
cfs1# ccs -h cfs1 --addfencedev mpcfs2 agent=fence_ilo3 \
     ipaddr=192.168.24.232 login=admin passwd=xxxxx verbose=1
cfs1# ccs -h cfs1 --addfencedev mpcfs3 agent=fence_ilo3 \
     ipaddr=192.168.24.233 login=admin passwd=xxxxx verbose=1
cfs1# ccs -h cfs1 --addfencedev mpcfs4 agent=fence_ilo3 \
     ipaddr=192.168.24.234 login=admin passwd=xxxxx verbose=1
```

(konfiguracja fence method dla każdego węzła)

```
cfs1# ccs -h cfs1 --addmethod ipmi cfs1.cosmo.local
cfs1# ccs -h cfs1 --addmethod ipmi cfs2.cosmo.local
cfs1# ccs -h cfs1 --addmethod ipmi cfs3.cosmo.local
cfs1# ccs -h cfs1 --addmethod ipmi cfs4.cosmo.local
```

(konfiguracja fence instance dla fence method dla każdego węzła)

```
cfs1# ccs -h cfs1 --addfenceinst mpcfs1 cfs1.cosmo.local ipmi
cfs1# ccs -h cfs1 --addfenceinst mpcfs2 cfs2.cosmo.local ipmi
cfs1# ccs -h cfs1 --addfenceinst mpcfs3 cfs3.cosmo.local ipmi
cfs1# ccs -h cfs1 --addfenceinst mpcfs4 cfs4.cosmo.local ipmi
```

(synchronizacja i aktywacja konfiguracji)

```
cfs1# ccs -h cfs1 --checkconf
Node: cfs2.cosmo.local does not match
Node: cfs3.cosmo.local does not match
Node: cfs4.cosmo.local does not match
cfs1# ccs -h cfs1 --sync --activate
cfs1# ccs -h cfs1 --checkconf
All nodes in sync.
```

*(test fencingu węzła – (**hard reset węzła**))*

```
cfs1# clustat
Cluster Status for cl02 @ Sat Feb  8 18:24:58 2014
Member Status: Quorate

 Member Name                              ID   Status
 ------ ----                              ---- ------
 cfs1.cosmo.local                            1 Online, Local
 cfs2.cosmo.local                            2 Online
 cfs3.cosmo.local                            3 Online
 cfs4.cosmo.local                            4 Online

cfs1# fence_node cfs4.cosmo.local
fence cfs4.cosmo.local success

cfs1# clustat
Cluster Status for cl02 @ Sat Feb  8 18:26:07 2014
Member Status: Quorate

 Member Name                              ID   Status
 ------ ----                              ---- ------
 cfs1.cosmo.local                            1 Online, Local
 cfs2.cosmo.local                            2 Online
 cfs3.cosmo.local                            3 Online
 cfs4.cosmo.local                            4 Offline
```

Konfiguracja failover domains

*(utworzenie domeny **dom1**)*

```
cfs1# ccs -h cfs1 -addfailoverdomain dom1 \
      ordered nofailback restricted
cfs1# ccs -h cfs1 --addfailoverdomainnode dom1 cfs1.cosmo.local 1
cfs1# ccs -h cfs1 --addfailoverdomainnode dom1 cfs2.cosmo.local 2
cfs1# ccs -h cfs1 --addfailoverdomainnode dom1 cfs3.cosmo.local 3
cfs1# ccs -h cfs1 --addfailoverdomainnode dom1 cfs4.cosmo.local 4
```

*(utworzenie domeny **dom2**)*

```
cfs1# ccs -h cfs1 -addfailoverdomain dom2 \
      ordered nofailback restricted
cfs1# ccs -h cfs1 --addfailoverdomainnode dom2 cfs2.cosmo.local 1
cfs1# ccs -h cfs1 --addfailoverdomainnode dom2 cfs3.cosmo.local 2
cfs1# ccs -h cfs1 --addfailoverdomainnode dom2 cfs4.cosmo.local 3
cfs1# ccs -h cfs1 --addfailoverdomainnode dom2 cfs1.cosmo.local 4
```

*(utworzenie domeny **dom3**)*

```
cfs1# ccs -h cfs1 -addfailoverdomain dom3 \
      ordered nofailback restricted
cfs1# ccs -h cfs1 --addfailoverdomainnode dom3 cfs3.cosmo.local 1
cfs1# ccs -h cfs1 --addfailoverdomainnode dom3 cfs4.cosmo.local 2
cfs1# ccs -h cfs1 --addfailoverdomainnode dom3 cfs1.cosmo.local 3
cfs1# ccs -h cfs1 --addfailoverdomainnode dom3 cfs2.cosmo.local 4
```

*(utworzenie domeny **dom4**)*

```
cfs1# ccs -h cfs1 -addfailoverdomain dom4 \
      ordered nofailback restricted
cfs1# ccs -h cfs1 --addfailoverdomainnode dom4 cfs4.cosmo.local 1
cfs1# ccs -h cfs1 --addfailoverdomainnode dom4 cfs1.cosmo.local 2
cfs1# ccs -h cfs1 --addfailoverdomainnode dom4 cfs2.cosmo.local 3
cfs1# ccs -h cfs1 --addfailoverdomainnode dom4 cfs3.cosmo.local 4
```

(synchronizacja i aktywacja konfiguracji)

```
cfs1# ccs -h cfs1 --checkconf
cfs1# ccs -h cfs1 --sync --activate
cfs1# ccs -h cfs1 --checkconf
```

(klastrowa konfiguracja serwera Lustre MGS)

```
cfs1# ccs -h cfs1 --addservice lfs-mgs-MGT0000 domain=dom1 \
    recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-mgs-MGT0000 lustrefs \
    name=mgs-MGT0000 device=/dev/mapper/mgs-mgt0000 \
    mountpoint=/lustre/fs/mgs/MGT0000
cfs1# ccs -h cfs1 --sync --activate
```

*(klastrowa konfiguracja systemu plików Lustre **fs1**)*

```
cfs1# ccs -h cfs1 --addservice lfs-fs1-MDT0000 domain=dom2 \
    recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs1-MDT0000 lustrefs \
    name=fs1-MDT0000 device=/dev/mapper/fs1-mdt0000 \
    mountpoint=/lustre/fs/fs1/MDT0000

cfs1# ccs -h cfs1 --addservice lfs-fs1-OST0000 domain=dom1 \
    recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs1-OST0000 lustrefs \
    name=fs1-OST0000 device=/dev/mapper/fs1-ost0000 \
    mountpoint=/lustre/fs/fs1/OST0000
cfs1# ccs -h cfs1 --addservice lfs-fs1-OST0001 domain=dom2 \
    recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs1-OST0001 lustrefs \
    name=fs1-OST0001 device=/dev/mapper/fs1-ost0001 \
    mountpoint=/lustre/fs/fs1/OST0001
cfs1# ccs -h cfs1 --addservice lfs-fs1-OST0002 domain=dom3 \
    recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs1-OST0002 lustrefs \
    name=fs1-OST0002 device=/dev/mapper/fs1-ost0002 \
    mountpoint=/lustre/fs/fs1/OST0002
cfs1# ccs -h cfs1 --addservice lfs-fs1-OST0003 domain=dom4 \
    recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs1-OST0003 lustrefs \
    name=fs1-OST0003 device=/dev/mapper/fs1-ost0003 \
    mountpoint=/lustre/fs/fs1/OST0003
```

*(klastrowa konfiguracja systemu plików Lustre **fs2**)*

```
cfs1# ccs -h cfs1 --addservice lfs-fs2-MDT0000 domain=dom3 \
    recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs2-MDT0000 lustrefs \
    name=fs2-MDT0000 device=/dev/mapper/fs2-mdt0000 \
    mountpoint=/lustre/fs/fs2/MDT0000

cfs1# ccs -h cfs1 --addservice lfs-fs2-OST0000 domain=dom1 \
    recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs2-OST0000 lustrefs \
    name=fs2-OST0000 device=/dev/mapper/fs2-ost0000 \
    mountpoint=/lustre/fs/fs2/OST0000

cfs1# ccs -h cfs1 --addservice lfs-fs2-OST0001 domain=dom2 \
    recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs2-OST0001 lustrefs \
    name=fs2-OST0001 device=/dev/mapper/fs2-ost0001 \
    mountpoint=/lustre/fs/fs2/OST0001

cfs1# ccs -h cfs1 --addservice lfs-fs2-OST0002 domain=dom3 \
    recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs2-OST0002 lustrefs \
    name=fs2-OST0002 device=/dev/mapper/fs2-ost0002 \
    mountpoint=/lustre/fs/fs2/OST0002

cfs1# ccs -h cfs1 --addservice lfs-fs2-OST0003 domain=dom4 \
    recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs2-OST0003 lustrefs \
    name=fs2-OST0003 device=/dev/mapper/fs2-ost0003 \
    mountpoint=/lustre/fs/fs2/OST0003
```

*(klastrowa konfiguracja systemu plików Lustre **fs3**)*

```
cfs1# ccs -h cfs1 --addservice lfs-fs3-MDT0000 domain=dom4 \
      recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs3-MDT0000 lustrefs \
      name=fs3-MDT0000 device=/dev/mapper/fs3-mdt0000 \
      mountpoint=/lustre/fs/fs3/MDT0000

cfs1# ccs -h cfs1 --addservice lfs-fs3-OST0000 domain=dom1 \
      recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs3-OST0000 lustrefs \
      name=fs3-OST0000 device=/dev/mapper/fs3-ost0000 \
      mountpoint=/lustre/fs/fs3/OST0000

cfs1# ccs -h cfs1 --addservice lfs-fs3-OST0001 domain=dom2 \
      recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs3-OST0001 lustrefs \
      name=fs3-OST0001 device=/dev/mapper/fs3-ost0001 \
      mountpoint=/lustre/fs/fs3/OST0001

cfs1# ccs -h cfs1 --addservice lfs-fs3-OST0002 domain=dom3 \
      recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs3-OST0002 lustrefs \
      name=fs3-OST0002 device=/dev/mapper/fs3-ost0002 \
      mountpoint=/lustre/fs/fs3/OST0002

cfs1# ccs -h cfs1 --addservice lfs-fs3-OST0003 domain=dom4 \
      recovery=relocate
cfs1# ccs -h cfs1 --addsubservice lfs-fs3-OST0003 lustrefs \
      name=fs3-OST0003 device=/dev/mapper/fs3-ost0003 \
      mountpoint=/lustre/fs/fs3/OST0003
```

(synchronizacja i aktywacja konfiguracji)

```
cfs1# ccs -h cfs1 --checkconf
cfs1# ccs -h cfs1 --sync --activate
cfs1# ccs -h cfs1 --checkconf
```

(wypisanie statusu klastra)

```
cfs1# clustat
Cluster Status for cl02 @ Thu Apr 10 15:05:43 2014
Member Status: Quorate

 Member Name                          ID   Status
 ------ ----                          ---- ------
 cfs1.cosmo.local                         1 Online, Local, rgmanager
 cfs2.cosmo.local                         2 Online, rgmanager
 cfs3.cosmo.local                         3 Online, rgmanager
 cfs4.cosmo.local                         4 Online, rgmanager

 Service Name              Owner (Last)              State
 ------- ----              ----- ------              -----
 service:lfs-fs1-MDT0000   cfs2.cosmo.local          started
 service:lfs-fs1-OST0000   cfs1.cosmo.local          started
 service:lfs-fs1-OST0001   cfs2.cosmo.local          started
 service:lfs-fs1-OST0002   cfs3.cosmo.local          started
 service:lfs-fs1-OST0003   cfs4.cosmo.local          started
 service:lfs-fs2-MDT0000   cfs3.cosmo.local          started
 service:lfs-fs2-OST0000   cfs1.cosmo.local          started
 service:lfs-fs2-OST0001   cfs2.cosmo.local          started
 service:lfs-fs2-OST0002   cfs3.cosmo.local          started
 service:lfs-fs2-OST0003   cfs4.cosmo.local          started
 service:lfs-fs3-MDT0000   cfs4.cosmo.local          started
 service:lfs-fs3-OST0000   cfs1.cosmo.local          started
 service:lfs-fs3-OST0001   cfs2.cosmo.local          started
 service:lfs-fs3-OST0002   cfs3.cosmo.local          started
 service:lfs-fs3-OST0003   cfs4.cosmo.local          started
 service:lfs-mgs-MGT0000   cfs1.cosmo.local          started
```

6.2.3 Procedury administracyjne klastra HA

W celu uruchomienia serwisów klastra na wszystkich węzłach (i automatycznej aktywacji serwisów przy bootowaniu) należy wykonać:

```
# ccs -h adm1 --startall
```

W celu zatrzymania serwisów klastra na wszystkich węzłach (i automatycznej deaktywacji serwisów przy bootowaniu) należy wykonać:

```
# ccs -h adm1 --stopall
```

W celu wyświetlenia statusu klastra należy wykonać:

```
# clustat
```

6.3 Konfiguracja wirtualizacji

W środowisku klastra Cosmo, na serwerach ***adm[1-2]***, skonfigurowano oprogramowanie wirtualizacyjne systemu ScientificLinux 6.4.

6.3.1 Konfiguracja maszyn wirtualnych na klastrze cl01

Instalacja oprogramowania wirtualizacyjnego (hypervisor) (na każdym węźle)

```
# yum grouplist -v | grep -i virtualization
   Virtualization (virtualization)
   Virtualization Client (virtualization-client)
   Virtualization Platform (virtualization-platform)
   Virtualization Tools (virtualization-tools)

# yum groupinstall virtualization
# yum groupinstall virtualization-client
# yum groupinstall virtualization-platform
# yum groupinstall virtualization-tools

# shutdown -r -y now
```

(konfiguracja aliasów hypervisora)

```
# cp /etc/libvirt/libvirt.conf /etc/libvirt/libvirt.conf.ORG
# vi /etc/libvirt/libvirt.conf
...
uri_aliases = [
   "adm1=qemu+ssh://root@adm1.cosmo.local/system",
   "adm2=qemu+ssh://root@adm2.cosmo.local/system",
]

uri_default = "qemu:///system"
```

Konfiguracja maszyn wirtualnych

*(wstępna konfiguracja maszyny wirtualnej **adm** w programie **virt-manager**)*

```
= Step 1 of 5
   Name: adm
   Connection: localhost (QEMU/KVM)
   Network Boor (PXE)
= Step 2 of 5
   OS type: Linux
   Version: Red Hat Enterprise Linux 6
= Step 3 of 5
   Memory (RAM): 16384 MB
   CPUs: 4
= Step 4 of 5
   Select managed of other existing storage: /dev/mapper/adm_disk01
= Step 5 of 5
   Host device eth0.2 (Bridge 'brvlan002')   <-- VLAN int
   Set a fixed MAC address: 52:54:00:d7:6b:7e
   Virt Type: kvm
   Architecture: x86_64
```

*(wstępna konfiguracja maszyny wirtualnej **mon** w programie **virt-manager**)*

```
= Step 1 of 5
   Name: mon
   Connection: localhost (QEMU/KVM)
   Network Boor (PXE)
= Step 2 of 5
   OS type: Linux
   Version: Red Hat Enterprise Linux 6
= Step 3 of 5
   Memory (RAM): 16384 MB
   CPUs: 4
= Step 4 of 5
   Select managed of other existing storage: /dev/mapper/mon_disk01
= Step 5 of 5
   Host device eth0.2 (Bridge 'brvlan002')   <-- VLAN int
   Set a fixed MAC address: 52:54:00:ea:8c:c6
   Virt Type: kvm
   Architecture: x86_64
```

*(wstępna konfiguracja maszyny wirtualnej **hn** w programie **virt-manager**)*

```
= Step 1 of 5
   Name: hn
   Connection: localhost (QEMU/KVM)
   Network Boor (PXE)
= Step 2 of 5
   OS type: Linux
   Version: Red Hat Enterprise Linux 6
= Step 3 of 5
   Memory (RAM): 16384 MB
   CPUs: 4
= Step 4 of 5
   Select managed of other existing storage: /dev/mapper/hn_disk01
= Step 5 of 5
   Host device eth0.2 (Bridge 'brvlan002')   <-- VLAN int
   Set a fixed MAC address: 52:54:00:99:8f:ab
   Virt Type: kvm
   Architecture: x86_64
```

*(wstępna konfiguracja maszyny wirtualnej **wo1** w programie **virt-manager**)*

```
= Step 1 of 5
   Name: wo1
   Connection: localhost (QEMU/KVM)
   Network Boor (PXE)
= Step 2 of 5
   OS type: Linux
   Version: Red Hat Enterprise Linux 6
= Step 3 of 5
   Memory (RAM): 1024 MB
   CPUs: 1
= Step 4 of 5
   Select managed of other existing storage: /dev/mapper/wo1_disk01
= Step 5 of 5
   Host device eth0.2 (Bridge 'brvlan002')   <-- VLAN int
   Set a fixed MAC address: 52:54:00:81:93:61
   Virt Type: kvm
   Architecture: x86_64
```

Klastrowa konfiguracja maszyn wirtualnych (na przykładzie VM ***wo1***)

(zablokowanie startu maszyn wirtualnych przy uruchamianiu systemu)

```
adm1# chkconfig libvirt-guests off
adm2# chkconfig libvirt-guests off
```

*(zatrzymanie **wo1**)*

```
adm2# virsh shutdown wo1
```

*(przeniesienie **wo1.xml** do do **/usr/local/etc/libvirt/qemu** i replikacja na drugi węzeł)*

```
adm1# mkdir -p /usr/local/etc/libvirt/qemu
adm2# mkdir -p /usr/local/etc/libvirt/qemu

adm2# mv /etc/libvirt/qemu/wo1.xml /usr/local/etc/libvirt/qemu
adm2# scp /usr/local/etc/libvirt/qemu/wo1.xml \
    adm1:/usr/local/etc/libvirt/qemu
adm1# rm /etc/libvirt/qemu/wo1.xml

adm1# service libvirtd restart
adm2# service libvirtd restart
```

(dodanie maszyny wirtualnej do klastra HA)

```
adm1# ccs -h adm1 --addvm wo1 \
    path=/usr/local/etc/libvirt/qemu use_virsh=1 domain=dom2
adm1# ccs -h adm1 --sync --activate

adm1# clustat
 Member Name                             ID   Status
 ------ ----                             ---- ------
 adm1.cosmo.local                            1 Online, Local, rgmanager
 adm2.cosmo.local                            2 Online, rgmanager
 /dev/block/253:2                            0 Online, Quorum Disk

 Service Name              Owner (Last)              State
 ------- ----              ----- ------              -----
 vm:wo1                    adm2.cosmo.local          started
```

6.3.2 Procedury administracyjne wirtualizacji

W celu zatrzymania klastrowej maszyny wirtualnej należy wykonać:

```
adm1# clusvcadm -d vm:wo1
```

W celu uruchomienia klastrowej maszyny wirtualnej na węźle ***adm1*** należy wykonać:

```
adm1# clusvcadm -e vm:wo1 -m adm1
```

W celu uruchomienia klastrowej maszyny wirtualnej na węźle ***adm2*** należy wykonać:

```
adm1# clusvcadm -e vm:wo1 -m adm2
```

W celu przełączenia *online* klastrowej maszyny wirtualnej na węzeł ***adm1*** należy wykonać:

```
adm1# clusvcadm -M vm:wo1 -m adm1
```

W celu przełączenia *online* klastrowej maszyny wirtualnej na węzeł ***adm2*** należy wykonać:

```
adm1# clusvcadm -M vm:wo1 -m adm2
```

W celu przełączenia *offline* klastrowej maszyny wirtualnej na węzeł ***adm1*** należy wykonać:

```
adm1# clusvcadm -d vm:wo1
adm1# clusvcadm -e vm:wo1 -m adm1
```

W celu przełączenia *offline* klastrowej maszyny wirtualnej na węzeł ***adm2*** należy wykonać:

```
adm1# clusvcadm -d vm:wo1
adm1# clusvcadm -e vm:wo1 -m adm2
```

W celu wyświetlenia statusu klastra należy wykonać:

```
adm1# clustat
```

W celu wyświetlenia statusu klastrowej maszyny wirtualnej należy wykonać:

```
adm1# clustat -s vm:wo1
```

6.4 AAA

Serwis AAA (*Authentication, Authorization and Accounting*) klastra Cosmo zrealizowany jest w oparciu o serwer OpenLDAP działający na maszynie wirtualnej ***adm***.

Na serwerze AAA zdefiniowano następujące domeny:

- *cosmo.local* – domena LDAP obsługiwana lokalnie (użytkownicy domeny są administrowani przez administrtora klastra Cosmo)
- *imgw.ad* – domena LDAP będąca proxy do domeny IMGW Active Directory (użytkownicy domeny są administrowani przez administratora IMGW Active Directory)

Klientami serwera AAA są serwisy ***sssd*** (*System Security Services Daemon*) aktywne na wszystkich serwerach Linux.

Serwis ***sssd*** swiadczy dla systemów Linux klastra Cosmo następujące usługi:

- *pam* – autentykacja użytkownika
- *nss* – dystrybucja informacji o użytkownikach i grupach (*uid*, *gid*, *home directory*, ...)

Podczas logowania użytkownika następuje próba weryfikacji w domenie *cosmo.local*. W przypadku niepowodzenia przeprowadzana jest próba weryfikacji w domenie *imgw.ad*. Jeżeli żadna z prób nie zakończy się sukcesem logowanie kończy się niepowodzeniem.

6.4.1 Konfiguracja serwera OpenLDAP

Instalacja serwera OpenLDAP:

```
adm# yum install openldap-servers openldap-clients migrationtools
```

Inicjalizacja serwera OpenLDAP:

(utworzenie pliku konfiguracyjnego /etc/openldap/slapd.conf)

```
adm# more /etc/openldap/slapd.conf
```

```
include /etc/openldap/schema/core.schema
include /etc/openldap/schema/cosine.schema
include /etc/openldap/schema/nis.schema

pidfile /var/run/openldap/slapd.pid
argsfile /var/run/openldap/slapd.args

TLSCACertificatePath /etc/openldap/certs
TLSCertificateFile "OpenLDAP Server"
TLSCertificateKeyFile /etc/openldap/certs/password
TLSVerifyClient never

access to attrs=userPassword by self write by * auth
access to * by * read

sizelimit 2000

database  bdb
suffix    dc=cosmo,dc=local
rootdn    cn=root,dc=cosmo,dc=local
rootpw    xxxx
directory /var/openldap/database-cosmo.local

database  ldap
readonly  yes
protocol-version 3
rebind-as-user
uri       ldaps://172.31.40.20,ldaps://172.31.40.26
suffix    dc=imgw,dc=ad
```

```
adm# chown ldap:ldap slapd.conf
adm# chmod o-rwx slapd.conf
```

(weryfikacja poprawności pliku konfiguracyjnego)

```
adm# slaptest -u -d config
```

(utworzenie bazy kluczy/certyfikatów)

```
adm# echo 1234567890 > /etc/openldap/certs/password
adm# chmod go-rwx /etc/openldap/certs/password
adm# certutil -N -d /etc/openldap/certs \
     -f /etc/openldap/certs/password
adm# chown -R ldap:ldap /etc/openldap/certs
```

(wygenerowanie certyfikatu serwera)

```
adm# /usr/libexec/openldap/generate-server-cert.sh
Creating new server certificate in '/etc/openldap/certs'.
```

(wyłączenie weryfikacji serwera przez klientów LDAP)

```
adm# vi ldap.conf
TLS_REQCERT     never
```

*(konfiguracja parametrów uruchomieniowych **slapd**)*

```
adm# vi /etc/sysconfig/ldap
SLAPD_LDAP=yes
SLAPD_LDAPI=yes
SLAPD_LDAPS=yes
```

*(konfiguracja automatychnego uruchamiania **slapd**)*

```
adm# chkconfig slapd on
adm# chkconfig --list slapd
```

*(uruchomienie demona **slapd**)*

```
adm# service slapd start
```

Wstępna konfiguracja serwera OpenLDAP:

(utworzenie plików LDIF dla domeny cosmo.local, OU Users i OU Groups)

```
 adm# cat cosmo.local.ldif
 dn: dc=cosmo,dc=local
 dc: cosmo
 description: Root LDAP entry for cosmo.local
 objectClass: dcObject
 objectClass: organizationalUnit
 ou: rootobject
adm# cat users.cosmo.local.ldif
 dn: ou=Users, dc=cosmo,dc=local
 ou: Users
 description: All users in organization
 objectClass: organizationalUnit
adm# cat groups.cosmo.local.ldif
 dn: ou=Groups, dc=cosmo,dc=local
 ou: Groups
 description: All groups in organization
 objectClass: organizationalUnit
```

(dodanie obiektów do LDAP)

```
 adm# ldapadd -x -D "cn=root,dc=cosmo,dc=local" -W \
      -f /root/ldap/cosmo.local.ldif
 Enter LDAP Password: <xxxx>
 adding new entry "dc=cosmo,dc=local"
 adm# ldapadd -x -D "cn=root,dc=cosmo,dc=local" -W \
      -f /root/ldap/users.cosmo.local.ldif
 Enter LDAP Password: <xxxx>
 adding new entry "ou=Users, dc=cosmo,dc=local"
 adm# ldapadd -x -D "cn=root,dc=cosmo,dc=local" -W \
      -f /root/ldap/groups.cosmo.local.ldif
 Enter LDAP Password: <xxxx>
 adding new entry "ou=Groups, dc=cosmo,dc=local"
```

6.4.2 Konfiguracja serwera sssd

Instalacja serwera ***sssd***:

```
# yum install openldap-clients
# yum install sssd sssd-client
```

Utworzenie użytkownika CBU (*Cosmo Bind User*) w domenie *cosmo.local*:

(utworzenie pliku LDIF)

```
adm# cat cbu.users.cosmo.local.ldif
dn: uid=cbu,ou=Users,dc=cosmo,dc=local
uid: cbu
cn: Cosmo Bind User
objectClass: account
objectClass: posixAccount
objectClass: top
loginShell: /bin/bash
uidNumber: 333
gidNumber: 100
homeDirectory: /home/cbu
gecos: Cosmo Bind User
```

(utworzenie użytkownika)

```
adm# ldapadd -x -D "cn=root,dc=cosmo,dc=local" -W \
     -f /root/ldap/cbu.users.cosmo.local.ldif
Enter LDAP Password: <xxxx>
adding new entry "uid=cbu,ou=Users,dc=cosmo,dc=local"
```

(zmiana hasla użytkownika)

```
adm# ldappasswd -s zzzz -D cn=root,dc=cosmo,dc=local -W -x \
     uid=cbu,ou=Users,dc=cosmo,dc=local
Enter LDAP Password: <xxxx>
```

(test bindowania jako CBU)

```
adm# ldapsearch -x -LLL -H ldaps://127.0.0.1 \
     -D "uid=cbu,ou=Users,dc=cosmo,dc=local" -W \
     -b "dc=cosmo,dc=local" "(objectc lass=posixAccount)"
Enter LDAP Password: <zzzz>
...
```

Utworzenie pliku konfiguracyjnego *sssd*:

```
adm# more /etc/sssd/sssd.conf
[sssd]
config_file_version = 2
services = nss, pam
domains = COSMO.LOCAL,IMGW.AD
[nss]
[pam]
##
## domain: COSMO.LOCAL
##
[domain/COSMO.LOCAL]
id_provider = ldap
auth_provider = none
chpass_provider = none
access_provider = deny

ldap_uri = ldaps://192.168.20.243,ldaps://192.168.28.243
ldap_tls_reqcert = never
ldap_default_bind_dn = uid=cbu,ou=Users,dc=cosmo,dc=local
ldap_default_authtok_type = password
ldap_default_authtok = zzzz

ldap_schema = rfc2307

ldap_user_search_base = ou=Users,dc=cosmo,dc=local
ldap_user_object_class = posixAccount
ldap_user_name = uid
ldap_user_uid_number = uidNumber
ldap_user_gid_number = gidNumber
ldap_user_gecos = gecos
ldap_user_home_directory = homeDirectory
ldap_user_shell = loginShell

ldap_group_search_base = ou=Groups,dc=cosmo,dc=local
ldap_group_object_class = posixGroup
ldap_group_name = cn
ldap_group_gid_number = gidNumber
ldap_group_member = memberUid
```

```
##
## domain: IMGW.AD
##
[domain/IMGW.AD]
id_provider = ldap
auth_provider = none
chpass_provider = none
access_provider = deny

ldap_uri = ldaps://192.168.20.243,ldaps://192.168.28.243
ldap_tls_reqcert = never
ldap_default_bind_dn = CN=ldap,CN=Users,DC=imgw,DC=ad
ldap_default_authtok_type = password
ldap_default_authtok = SECRET

ldap_schema = rfc2307

ldap_user_search_base = dc=imgw,dc=ad
ldap_user_object_class = person
ldap_user_name = uid
ldap_user_uid_number = uidNumber
ldap_user_gid_number = gidNumber
ldap_user_home_directory = unixHomeDirectory
ldap_user_shell = loginShell

ldap_group_search_base = dc=imgw,dc=ad
ldap_group_object_class = group
ldap_group_name = cn
ldap_group_gid_number = gidNumber
ldap_group_member = member
```

(modyfikacja uprawnień pliku konfiguracyjnego)

```
# chmod 600 /etc/sssd/sssd.conf
# chown root:root /etc/sssd/sssd.conf
```

Dla każdej domeny LDAP w pliku *sssd.conf* zdefiniowany jest, między innymi, dostawca kontrolujący dostęp do serwera (*access_provider*).

Dla maszyny ***adm*** konfiguracja dostawcy dostępu wygląda następująco:

```
...
[domain/COSMO.LOCAL]
...
access_provider = deny
...
[domain/IMGW.AD]
...
access_provider = deny
...
```

Dostawca *deny* oznacza że nawet jeżeli użytkownik poda prawidłowe parametry autoryzacyjne i tak nie zaloguje się na danym serwerze.

Dla maszyny ***hn*** konfiguracja dostawcy dostępu wygląda następująco:

```
...
[domain/COSMO.LOCAL]
...
access_provider = permit
...
[domain/IMGW.AD]
...
access_provider = simple
simple_allow_users = mmarcola,wlazarewicz,...
...
```

Dostawca *perm* oznacza ze każdy użytkownik który poda prawidłowe parametry autoryzacyjne zaloguje się na serwer (tu jest tak dla domeny *cosmo.local*).

Dostawca *simple* oznacza że prawo zalogowania się na serwer mają użytkownicy wyspecyfikowani w dyrektywie *simple_allow_users* (tu jest tak dla domeny *imgw.ad*).

Dzięki odpowiedniej konfiguracji dostawcy dostępu można blokować możliwość logowania się zwykłego użytkownika na pewne serwery (np. ***adm1***, ***adm2***, ***adm***, ***mon***, ***cfs[1-4]***) lub zezwalać na logowanie się na inne (np. ***hn***, ***node[001-139]***). Można też precyzyjnie specyfikować którzy użytkownicy z danej domeny mają prawo do logowania się.

Konfiguracja PAM/NSS tak by używały ***sssd*** (RedHat/CentOS/SL):

```
# authconfig --enablesssd --enablesssdauth --enablemkhomedir -update
```

Aktywacja i uruchomienie serwisu ***sssd***:

```
# chkconfig sssd on
# service sssd start
```

Uruchomienie ***sssd*** w trybie maksymalnego debugu interakcyjne na terminalu:

```
# sssd -i -d 10
```

6.4.3 Procedury administracyjne AAA

Administracją użytkownikami w domenie *imgw.ad* zajmuje się administrator Active Directory.

Użytkownik Active Directory by mógł logwać się do klastra Cosmo musi mieć zdefiniowane dodatkowe atrybuty UNIX:

Atrybut LDAP	***Opis***
`objectClass=person`	Użytkownik musi należeć do klasy *person*
`uid`	Nazwa użytkownika
`uidNumber`	Identyfikator użytkownika
`gidNumber`	Identyfikator grupy użytkownika
`unixHomeDirectory`	Katalog domowy użytkownika
`loginShell`	Powłoka użytkownika

Grupa Active Directory by mogła być wykorzystana przez klaster Cosmo musi mieć zdefiniowane dodatkowe atrybuty UNIX:

Atrybut LDAP	***Opis***
`objectClass=group`	Grupa musi należeć do klasy *group*
`cn`	Nazwa grupy
`gidNumber`	Identyfikator grupy
`member`	Użytkownicy należący do grupy

Ponieważ dostęp użytkowników do klastra Cosmo odbywa się poprzez maszynę wirtualną ***hn*** (*Head Node*) administrator klastra Cosmo musi autoryzować użytkownika Aactive Directory u dostawcy dostępu *simple* na tej maszynie (dyrektywa *simple_allow_users*).

Administracją użytkownikami w domenie *cosmo.local* zajmuje się administrator klastra Cosmo.

Administracja polega na modyfikacji konfiguracji serwera LDAP działającego na maszynie wirtualnej ***adm***.

Dla ułatwienia administracji utworzono zestaw skryptów do administracji użytkownikami i grupami (skrypty znajdują się w katalogu */usr/local/hpc/bin*).

W celu wylistowania użytkowników z domeny *cosmo.local* należy wykonać:

```
# hpc_lsuser [username]
```

W celu wylistowania grup z domeny *cosmo.local* należy wykonać:

```
# hpc_lsgroup [groupname]
```

W celu utworzenia użytkownika w domenie *cosmo.local* należy wykonać:

```
# hpc_mkuser [-n] username uid gid home shell
```

W celu utworzenia grupy w domenie *cosmo.local* należy wykonać:

```
# hpc_mkgroup [-n] groupname gid
```

W celu dodania użytkownika do grupy w domenie *cosmo.local* należy wykonać:

```
# hpc_chgroup [-n] groupname username
```

W celu usunięcia użytkownika z grupy w domenie *cosmo.local* należy wykonać:

```
# hpc_chgroup [-n] -d groupname username
```

W celu usunięcia użytkownika z domeny *cosmo.local* należy wykonać:

```
# hpc_rmuser [-n] username
```

W celu usunięcia grupy z domeny *cosmo.local* należy wykonać:

```
# hpc_rmgroup [-n] groupname
```

W celu zmiany hasła użytkownika z domeny *cosmo.local* należy wykonać:

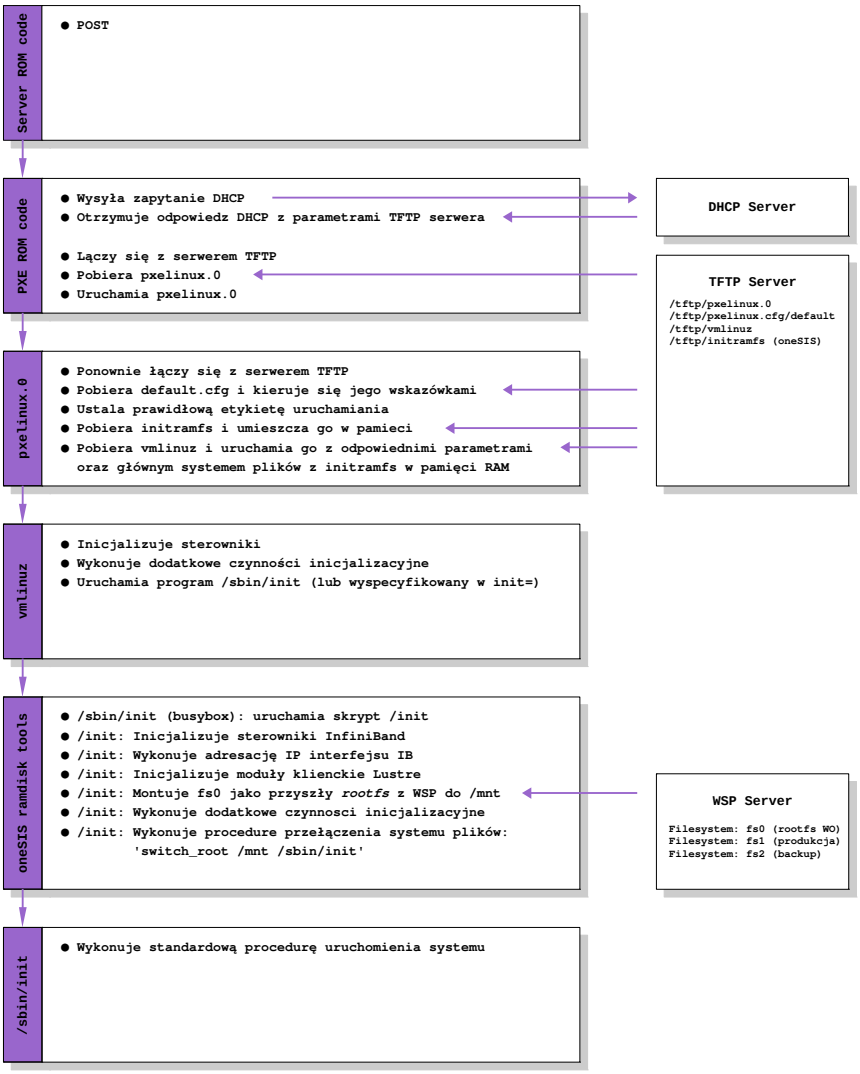
```
# hpc_passwd username
$ passwd
```

Opcja „***-n***” wypisuje informacje o wykonywanych czynnościach ale nie wykonuje ich.

Rozdział 7: Węzły obliczeniowe

7.1 Architektura

Węzły obliczeniowe uruchamiają się bezdyskowo poprzez sieć z wykorzystaniem protokołów DHCP i TFTP w środowisku PXE (*Preboot Execution Environment*) pod kontrolą systemu operacyjnego ScientificLinux 6.4 x86_64.



Procedura uruchamiania węzła obliczeniowego

Pierwszy etap uruchamiania węzłów obliczeniowych (do zamontowania głównego systemu plików i wykonania procedury *switch_root*) odbywa się z wykorzystaniem sieci Ethernet. Drugi etap uruchamiania i działanie realizowane jest w oparciu o sieć InfiniBand.

Główny system plików montowany jest przez węzły obliczeniowe w trybie *read-only* co umożliwia współdzielenie jednego obrazu przez wiele węzłow.

Ze względu na montowanie głównego systemu plików z serwerów Lustre (WSP) wykorzystano zmodyfikowane oprogramowanie oneSIS (*http://www.onesis.org*).

Oprogramowanie oneSIS służy do:

- tworzenia obrazów głównego systemu plików z istniejących instalacji Linux (możliwe jest utrzymywanie wielu obrazów głównego systemu plików dla różnych zastosowań)
- tworzenia obrazu RAM filesystemu umożliwiającego wstępną inicjalizację środowiska

Węzły obliczeniowe posiadają nazwy w formacie *nodeNNN* gdzie NNN jest kolejnym numerem węzła z zakresu 001 – 139.

7.2 Konfiguracja oneSIS

7.2.1 Instalacja oprogramowania oneSIS

Oprogramowanie oneSIS zainstalowane jest na serwerze VM *adm* (generacja rootfs i iniramfs) oraz na wzorcu WO *wo1*.

W celu instalacji oprogramowania oneSIS należy wykonać polecenia (*adm*,*wo1*):

```
adm# cd /var/tmp/sw
adm# gzip -dc oneSIS-2.0.4.1.tar.gz |  tar xvf -
adm# cd oneSIS-2.0.4.1
adm# make install
```

Jeżeli wygenerowany rootfs nie ma zainstalowanego oprogramowania oneSIS to należy je doinstalować bezpośrednio do obrazu:

```
adm# hpc_mount_lustre_fs.sh fs1
adm# cd /var/tmp/sw
adm# gzip -dc oneSIS-2.0.4.1.tar.gz | tar xvf -
adm# cd oneSIS-2.0.4.1
adm# prefix=/cfs/fs1/diskless/root-sl64-x86_64-class01 make install
adm# umount /cfs/fs1
```

Jeżeli wygenerowany rootfs nie ma zainstalowanego oprogramowania oneSIS i posiada inną wersję programu *perl* niż serwer na którym jest zamontowany filesystem Lustre z obrazem to należy wykonać instalację oneSIS w środowisku rootfs:

```
adm# hpc_mount_lustre_fs.sh fs1
adm# mkdir /cfs/fs1/diskless/root-sl64-x86_64-class01/var/tmp/sw
adm# cp /var/tmp/sw/oneSIS-2.0.4.1.tar.gz \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/tmp/sw
adm# chroot /cfs/fs1/diskless/root-sl64-x86_64-class01
adm# cd /var/tmp/sw
adm# gzip -dc oneSIS-2.0.4.1.tar.gz | tar xvf -
adm# cd oneSIS-2.0.4.1
adm# make install
adm# exit
adm# umount /cfs/fs1
```

7.2.2 Konfiguracja wzorca WO wo1

W celu podstawowej konfiguracji wzorca *wo1* wykonano:

- wstępną konfigurację systemu Linux (*Dodatek A*)
- instalację klienta Lustre
- instalację oneSIS
- instalację Torque Clients i MOM
- instalację PDSH
- konfigurację autoryzacji SSH (*Dodatek B*)
- instalację Intel MPI Runtime (*Dodatek C*)

Dodatkowo wykonano:

(zablokowanie dodatkowych serwisów)

```
wo1# chkconfig ip6tables off
wo1# chkconfig iptables off
wo1# chkconfig kdump off
wo1# chkconfig sysstat off
wo1# chkconfig mdmonitor off
wo1# chkconfig haldaemon off
wo1# chkconfig auditd off
wo1# chkconfig messagebus off
wo1# chkconfig rpcidmapd off
wo1# chkconfig rpcgssd off
wo1# chkconfig rpcbind off
wo1# chkconfig nfslock off
wo1# chkconfig udev-post off
```

(założenie użytkownika admin)

```
wo1# useradd -u 1000 -c admin -m -n admin
wo1# passwd admin
Changing password for user admin.
New password: <xyz>
Retype new password: <xyz>
passwd: all authentication tokens updated successfully.
```

*(instalację RPMów wymaganych „po cichu” przez **sssd**)*

```
wo1# yum install glibc-devel-2.12-1.107.el6.i686
wo1# yum install libgcc.i686
```

(instalację klienta LDAP)

```
wo1# yum install openldap-clients
```

(instalację oprogramowania Infiniband)

```
wo1# yum groupinstall infiniband
wo1# chkconfig rdma on
```

(zablokowanie odładowywania modułów IB przy zatrzymywaniu WO)

```
wo1# vi /etc/init.d/rdma
...
      echo -n "Unloading OpenIB kernel modules:"
      return 0
...
```

Bez tej modyfikacji WO zawiesi się przy zatrzymywaniu lub restarcie ponieważ rootfs jest na Lustre **fs1** montowanym przez sieć Infiniband – odcięcie filesystemu !!!

(synchronizację HPC tools)

```
adm# hpc_rsync.sh wo1
wo1# vi /etc/profile.d/local.sh
...
export PATH=$PATH:/usr/local/hpc/bin
```

7.2.3 Generacja rootfs

Generacja rootfs odbywa się poprzez skopiowanie poprzez sieć systemu plików z wzorca (*wo1*) na system plików Lustre ***fs1***, który będą montowały bezdyskowe Węzły Obliczeniowe.

Założono, że kolejne wersje rootfs będą zapisywane w katalogu */cfs/fs1/diskless* w odpowiednich podkatalogach:

```
adm# hpc_mount_lustre_fs.sh fs1
adm# ls -F1 /cfs/fs1/diskless
root-sl64-x86_64-class01/
root-sl64-x86_64-class02/
root-sl64-x86_64-class03/
```

Procedura kopiowania rootfs:

*(zamontowanie filesystemu Lustre **fs1**)*

```
adm# hpc_mount_lustre_fs.sh fs1
```

(*utworzenie katalogu na rootfs*)

```
adm# mkdir -p /cfs/fs1/diskless/root-sl64-x86_64-class01
```

(*skopiowanie zdalnego filesystemu poprzez ssh – tylko wypisanie zadań*)

```
adm# copy-rootfs -r wo1 --dryrun \
 /cfs/fs1/diskless/root-sl64-x86_64-class01
Copying wo1:/ to /cfs/fs1/diskless/root-sl64-x86_64-class01/
Copying wo1:/boot to /cfs/fs1/diskless/root-sl64-x86_64-class01/boot
Copying wo1:/dev to /cfs/fs1/diskless/root-sl64-x86_64-class01/dev
```

(*skopiowanie zdalnego filesystemu poprzez ssh – tylko wypisanie poleceń*)

```
adm# copy-rootfs -r wo1 --debug \
 /cfs/fs1/diskless/root-sl64-x86_64-class01
cd /cfs/fs1/diskless/root-sl64-x86_64-class01/; \
 ssh wo1 "cd /; find . -xdev |cpio -o -H crc" |cpio -imud
cd /cfs/fs1/diskless/root-sl64-x86_64-class01/boot; \
 ssh wo1 "cd /boot; find . -xdev |cpio -o -H crc" |cpio -imud
cd /cfs/fs1/diskless/root-sl64-x86_64-class01/dev; \
 ssh wo1 "cd /dev; find . -xdev |cpio -o -H crc" |cpio -imud
```

(*skopiowanie zdalnego filesystemu poprzez ssh*)

```
adm# copy-rootfs -r wo1 \
 /cfs/fs1/diskless/root-sl64-x86_64-class01
```

Procedura modyfikacji rootfs:

Skopiowany rootfs należy dostosować do wymagań montowania w trybie *read-only* na WO.

(utworzenie katalogu /log na log skryptu /init openSIS)

```
adm# cd /cfs/fs1/diskless/root-sl64-x86_64-class01
adm# mkdir log
```

(usunięcie logów wo1)

```
adm# cd /cfs/fs1/diskless/root-sl64-x86_64-class01/var/log
adm# rm -f *log */*log
adm# rm -f spooler* secure* messages* maillog* dmesg* cron* \
 sa/* ConsoleKit/* btmp wtmp
```

(usunięcie zawarości katalogu /var/lock/subsys)

```
adm# cd /cfs/fs1/diskless/root-sl64-x86_64-class01/var/lock/subsys
adm# rm -f *
```

(usunięcie plików z katalogu /var/run)

```
adm# cd /cfs/fs1/diskless/root-sl64-x86_64-class01/var/run
adm# find . -type f -exec rm {} \;
adm# rm -f acpid.socket mcelog-client rpcbind.sock \
 dbus/system_bus_socket portreserve/socket
```

(usunięcie montowań z pliku /etc/fstab poza montowaniem devpts)

```
adm# cd /cfs/fs1/diskless/root-sl64-x86_64-class01/etc
adm# vi fstab
UUID=b85fa7e8-36e8-4b8b-b76e-179279e8d482 /boot ext4 defaults 1 2
tmpfs          /dev/shm         tmpfs   defaults        0 0
sysfs          /sys             sysfs   defaults        0 0
proc           /proc            proc    defaults        0 0
```

(utworzenie katalogów na montowania filesystemów Lustre)

```
adm# cd /cfs/fs1/diskless/root-sl64-x86_64-class01
adm# mkdir -p cfs/fs1 cfs/fs2 cfs/fs3
```

(dodanie do etc/fstab montowania /cfs/fs2)

```
adm# cd /cfs/fs1/diskless/root-sl64-x86_64-class01
adm# vi etc/fstab
...
192.168.20.231@o2ib:192.168.20.232@o2ib:\
192.168.20.233@o2ib:192.168.20.234@o2ib:/fs2 \
 /cfs/fs2 lustre defaults,_netdev 1 1
```

Opcja *_netdev* powoduje że filesystem jest montowany dopiero przez serwis **netfs**

(modyfikacja katalogu /home)

```
adm# cd /cfs/fs1/diskless/root-sl64-x86_64-class01
adm# rmdir home
adm# ln -s /cfs/fs2/home .
```

(utworzenie pliku konfiguracyjnego oneSIS)

```
adm# mv /etc/sysimage.conf /etc/sysimage.conf.ORG
adm# vi /etc/sysimage.conf
DISTRO redhat-el 6.2

LINKDIR /tmp
LINKDIR /var/tmp
LINKDIR /var/run -d
LINKDIR /var/log -d
LINKDIR /var/lock/subsys
LINKDIR /var/empty/sshd -d
LINKDIR /var/lib/sss -d
LINKDIR /var/spool/torque -d
adm# cp /etc/sysimage.conf \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/etc
```

Pliki z modyfikacjami dla wspieranych dystrubucji Linux (parametr DISTRO) znajdują się w katalogu */usr/share/oneSIS/distro-patches* (tu: *redhat-el-6.2.patch*).

(*modyfikacja rootfs – tylko wypisanie zadań*)

```
adm# mk-sysimage --dryrun --debug -c /etc/sysimage.conf \
 /cfs/fs1/diskless/root-sl64-x86_64-class01
```

(*modyfikacja rootfs*)

```
adm# mk-sysimage --debug -c /etc/sysimage.conf \
 /cfs/fs1/diskless/root-sl64-x86_64-class01
DEBUG: patch -p0 -d /cfs/fs1/diskless/root-sl64-x86_64-class01 \
 -i /usr/share/oneSIS/distro-patches/redhat-el-6.2.patch -u -N \
 --no-backup-if-mismatch -r /tmp/res
oneSIS: Applying patch: /usr/share/oneSIS/distro-patches/redhat-el-6.2.patch
oneSIS: Altering file: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/etc/inittab
oneSIS: Altering file: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/etc/fstab
oneSIS: Altering file: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/.autofsck
oneSIS: Altering file: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/etc/sysconfig/\
 network-scripts/ifcfg-eth0
oneSIS: Altering file: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/etc/sysconfig/network
oneSIS: Creating directory: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/ram
oneSIS: Creating directory: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/initrd
oneSIS: Creating directory: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/mnt/oneSIS-disk
oneSIS: Creating directory: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/mnt/oneSIS-root
oneSIS: Sym-linking /etc/mtab to /proc/mounts
oneSIS: Renaming: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/tmp -> \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/tmp.default
oneSIS: Creating symlink: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/tmp --> /ram/tmp
oneSIS: Creating LINKDIR: /cfs/fs1/diskless/root-sl64-x86_64-class01/tmp
```

```
oneSIS: Renaming: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/tmp -> \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/tmp.default
oneSIS: Creating symlink: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/tmp --> /ram/var/tmp
oneSIS: Creating LINKDIR: /cfs/fs1/diskless/root-sl64-x86_64-class01/var/tmp
oneSIS: Renaming: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/run -> \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/run.default
oneSIS: Creating symlink: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/run --> /ram/var/run
oneSIS: Creating LINKDIR: /cfs/fs1/diskless/root-sl64-x86_64-class01/var/run
oneSIS: Renaming: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/log -> \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/log.default
oneSIS: Creating symlink: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/log --> /ram/var/log
oneSIS: Creating LINKDIR: /cfs/fs1/diskless/root-sl64-x86_64-class01/var/log
oneSIS: Renaming: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/lock/subsys  -> \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/lock/subsys.default
oneSIS: Creating symlink: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/lock/subsys --> \
 /ram/var/lock/subsys
oneSIS: Creating LINKDIR: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/lock/subsys
oneSIS: Renaming: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/empty/sshd -> \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/empty/sshd.default
oneSIS: Creating symlink: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/empty/sshd --> \
 /ram/var/empty/sshd
oneSIS: Creating LINKDIR: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/empty/sshd
oneSIS: Renaming: \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/lib/sss -> \
 /cfs/fs1/diskless/root-sl64-x86_64-class01/var/lib/sss.default
...
...
```

7.2.4 Generacja initramfs

Parametry jądra systemu Linux związane z bootowaniem WO wykorzystywanie w initramfs:

Parametr	*Opis*
`cluster=`	Wskazanie nazwy klastra (tu: *cosmo*) – konfiguracja interfejsu Infiniband.
`ping=`	Wskazuje numer IP wykorzystywany przy weryfikacji połączenia sieciowego IB
`lustreroot=`	Wskazanie filesystemu Lustre montowanego jako rootfs.
`shell1`	(Malkom) Wywołanie shella po załadowaniu podstawowych modułów jadra.
`shell2`	(Malkom) Wywołanie shella po konfiguracji sieci.
`shell3`	(Malkom) Wywołanie shella po zamontowaniu systemu plików Lustre.
`shell9`	(Malkom) Wywołanie shella przed wywołaniem procedury *switch_root* .

Parametr *lustreroot* wskazuje lokalizacje filesystemu Lustre który zostanie użyty jako rootfs. Parametr posiada następującą składnie:

```
lustreroot=lustreserver:/lustrefs/lustrepath
```

gdzie:

- *lustreserver* – wskazuje lokalizację postawowego oraz zapasowych serwerów MGS.
- *lustrefs* – wskazuje filesystem Lustre do zamontowania.
- *lustrepath* – wskazuje scieżkę w ramach filesystemu *lustrefs* z obrazem filesystemu root.

Przykład:

```
lustreroot= \
  192.168.20.231@o2ib:192.168.20.232@o2ib: \
  192.168.20.233@o2ib:192.168.20.234@o2ib: \
  /fs1/diskless/root-sl64-x86_64-class01
```

Parametry *shell1* - *shell9* są aktywne podczas inicjalizacji systemu z initramfs oneSIS. Umożliwiają na danym etapie inicjalizacji uruchomienie na konsoli powłoki systemowej i przeprowadzenie dalszej diagnostyki. Funkcjonalność została uzyskana przez modyfikację skryptu *init* wzorca oneSIS *x86_64-glibc-rjh.tar.gz*.(oryginalny skrypt zapisano jako *init.ORG*).

Ustawienie parametru *cluster=cosmo* powoduje że podczas inicjalizacji sieci w initramfs interfejs Infiniband zostanie skonfigurowany z numerem IP w sieci **cmp** (*192.168.20.0/22*) z koncówką z sieci **int** (*192.168.28.0/22*).

Na przykład jeżeli WO dostał od serwera DHCP numer IP w sieci **int** *192.168.28.101/22* (interfejs eth0) to adres WO w sieci **cmp** zostanie skonfigurowany na *192.168.20.101/22* (interfejs ib0).

W celu uzyskania wymienionej funkcjonalności zmodyfikowano skrypt *init* z wzorca oneSIS *x86_64-glibc-rjh.tar.gz* dodając w odpowiednim miejscu dodatkowy kod:

```
...
   elif [ "$cluster" = "cosmo" ]; then
       if [ $INTERFACE != "ib0" ]; then
           # configure ib0
           ib0ip=`ifconfig eth0 | grep 'inet addr:' | \
                  sed -e "s/.*inet addr://" \
                      -e "s/ .*//" \
                      -e "s/^192.168.28./192.168.20./"`
           log echo "oneSIS: mm: cluster $cluster : ib0 ip is $ib0ip"
           ifconfig ib0 $ib0ip netmask 255.255.252.0
       fi
   else
...
```

Nowy wzorzec został zapisany jako *x86_64-glibc-rjh-mm.tar.gz*.

Wzorce initramfs oneSIS znajdują się w katalogu: */usr/share/oneSIS/initramfs-templates*:

```
adm# ls -F1 /usr/share/oneSIS/initramfs-templates
x86_64-glibc-rjh-mm.tar.gz
x86_64-glibc-rjh.tar.gz
x86_64-glibc.tar.gz
x86-uclibc.tar.gz
```

Procedura generacji wzorca x86_64-glibc-rjh-mm.tar.gz:

Nowy wzorzec bazuje na dodatkowym wzorcu *x86_64-glibc-rjh.tar.gz* oneSIS który można skopiować z adresu https://github.com/plaguedbypenguins/oneSIS.

(instalacja wzorca bazowego)

```
adm# mv /var/tmp/sw/x86_64-glibc-rjh.tar.gz \
 /usr/share/oneSIS/initramfs-templates
```

(rozpakowanie wzorca x86_64-glibc-rjh.tar.gz)

```
adm# mkdir /var/tmp/sw/t
adm# cd /var/tmp/sw/t
adm# gzip -dc \
 /usr/share/oneSIS/initramfs-templates/x86_64-glibc-rjh.tar.gz | \
 tar xvf -
```

(modyfikacja pliku init – dopasowanie initramfs do wymogów IMGW)

```
adm# cp init init.ORG
adm# cp /var/tmp/sw/init-2.0.2-rjh-mm-20140211 ./init
```

(modyfikacja pliku etc/modprobe.conf – zdefiniowanie sieci Lnet o2ib)

```
adm# vi etc/modprobe.conf
options lnet networks=o2ib
...
```

(modyfikacja pliku etc/initramfs.conf – DHCP z initramfs po sieci Ethernet)

```
adm# vi etc/initramfs.conf
...
DHCP_INTERFACE: eth0
#DHCP_INTERFACE: ib0
...
```

(utworzenie pliku etc/modules.autoload.cosmo – ładowanie sterowników Ethernet i Infiniband)

```
adm# vi etc/modules.autoload.cosmo
mdio
libcrc32c
bnx2x
ib_addr
ib_core
ib_mad
ib_sa
ib_cm
ib_uverbs
ib_ucm
ib_umad
iw_cm
rdma_cm
rdma_ucm
mlx4_core
mlx4_ib
ib_mthca
ib_ipoib
```

(rozpakowanie klienckich modułów jądra Lustre)

```
adm# rpm2cpio \
 /var/tmp/sw/lustre/lustre-2.4.2/el6/client/RPMS/x86_64/\
 lustre-client-modules-2.4.2-2.6.32_358.23.2.el6.x86_64.x86_64.rpm |\
 cpio -itvd

adm# rpm2cpio \
 /var/tmp/sw/lustre/lustre-2.4.2/el6/client/RPMS/x86_64/\
 lustre-client-modules-2.4.2-2.6.32_358.23.2.el6.x86_64.x86_64.rpm |\
 cpio -ivd
```

(dodanie dodatkowych modułów Infiniband – skopiowanie z wo1)

```
adm# ssh wo1 'cd /; tar cf - \
 lib/modules/2.6.32-358.23.2.el6.x86_64/kernel/drivers/infiniband' |\
 tar tvf -

adm# ssh wo1 'cd /; tar cf - \
 lib/modules/2.6.32-358.23.2.el6.x86_64/kernel/drivers/infiniband' |\
 tar xvf -
```

(dodanie dodatkowych modułów IPv6 – skopiowanie z wo1)

```
adm# ssh wo1 'cd /; tar cf - \
 lib/modules/2.6.32-358.23.2.el6.x86_64/kernel/net/ipv6/ipv6.ko' |\
 tar tvf -

adm# ssh wo1 'cd /; tar cf - \
 lib/modules/2.6.32-358.23.2.el6.x86_64/kernel/net/ipv6/ipv6.ko' |\
 tar xvf -
```

(wygenerowanie zależności pomiędzy modułami)

```
adm# depmod -nv -b /var/tmp/sw/t 2.6.32-358.23.2.el6.x86_64
adm# depmod -v -b /var/tmp/sw/t 2.6.32-358.23.2.el6.x86_64
```

(utworzenie nowego wzorca)

```
adm# tar cvf - * | gzip > \
 /usr/share/oneSIS/initramfs-templates/x86_64-glibc-rjh-mm.tar.gz

adm# ls -F1 /usr/share/oneSIS/initramfs-templates
x86_64-glibc-rjh-mm.tar.gz
x86_64-glibc-rjh.tar.gz
x86_64-glibc.tar.gz
x86-uclibc.tar.gz
```

Procedura generacji initramfs (initramfs-2.6.32-358.23.2.el6.x86_64):

(wylistowanie dostępnych wersji jądra w obrazie rootfs)

```
adm# ls -F1 /cfs/fs1/diskless/root-sl64-x86_64-class01/lib/modules
2.6.32-358.23.2.el6.x86_64/
2.6.32-358.el6.x86_64/
```

(utworzenie initramfs dla danej wersji jądra (template: x86_64-glibc-rjh-mm.tar.gz))

```
adm# cat /usr/local/hpc/bin/hpc_initramfs.sh
#!/bin/sh

rm -f /tmp/initramfs-2.6.32-358.23.2.el6.x86_64

mk-initramfs-oneSIS \
    --template=/usr/share/oneSIS/initramfs-templates/\
         x86_64-glibc-rjh-mm.tar.gz \
    --basedir=/cfs/fs1/diskless/root-sl64-x86_64-class01 \
    --config=/etc/oneSIS/initramfs.conf \
    --verbose \
    --debug \
    -w bnx2x \
    -w mlx4_core \
    -w mlx4_ib \
    --verbose \
    /tmp/initramfs-2.6.32-358.23.2.el6.x86_64 \
    2.6.32-358.23.2.el6.x86_64
adm# hpc_initramfs.sh
```

(skopiowanie initramfs i jądra Linux na serwery bootujące adm[1-2])

```
adm1# mkdir -p /tftpboot/sl/6.4/x86_64/class01
adm2# mkdir -p /tftpboot/sl/6.4/x86_64/class01
adm# scp /tmp/initramfs-2.6.32-358.23.2.el6.x86_64 \
    adm[1-2]:/tftpboot/sl/6.4/x86_64/class01
adm# scp /cfs/fs1/diskless/root-sl64-x86_64-class01/boot/\
    vmlinuz-2.6.32-358.23.2.el6.x86_64 \
    adm[1-2]:/tftpboot/sl/6.4/x86_64/class01
```

7.3 Konfiguracja TFTP/PXE

Serwis TFTP udostępnia Węzłom Obliczeniowym pliki PXE (*pxelinux.0*,*default*) oraz jądro i initramfs systemu Linux.

W celu instalacji serwisu TFTP i PXE na obu komputerach należy wykonać polecenia:

```
# yum install xinetd
# yum install tftp tftp-server
# yum install syslinux
```

W celu konfiguracji serwisu TFTP należy wykonać:

```
# mkdir /tftpboot
# vi /etc/xinetd.d/tftp
...
server_args = -s /tftpboot -v
...
```

W celu aktywacji i uruchomienia serwisu TFTP na obu komputerach należy wykonać polecenia:

```
# chkconfig xinetd on
# service xinetd start
# chkconfig tftp on
```

W celu uruchomienia serwisu TFTP w trybie debug należy wykonać polecenie:

```
# /usr/sbin/in.tftpd -l -s /tftpboot -L -vvvvvv
```

Po modyfikacji zawartości kalalogu */tftpboot* na jednym z serwerów, należy zreplikować konfigurację na drugi serwer (modyfikacje można wykonywać na dowolnym serwerze), np:

```
adm1# /tftpboot/build_tftpboot.sh
Synchronizing /tftpboot to adm2
sending incremental file list
...
```

W celu konfiguracji PXE należy wykonać polecenia:

```
# mkdir -p /tftpboot/sl/6.4/x86_64/class01
# cp /usr/share/syslinux/pxelinux.0 /tftpboot/sl/6.4/x86_64/class01
# mkdir /tftpboot/sl/6.4/x86_64/class01/pxelinix.cfg
```

Plik */tftpboot/sl/6.4/x86_64/class01/pxelinux.cfg/default* jest plikiem konfiguracyjnym PXE:

```
adm1# cat /tftpboot/sl/6.4/x86_64/class01/pxelinux.cfg/default
menu title Welcome to Scientific Linux 6.4 diskless node!
prompt 1
timeout 10
default sl64-class01

label sl64-class01
  menu Diskless class01 node
  kernel vmlinuz-2.6.32-358.23.2.el6.x86_64
  append initrd=initramfs-2.6.32-358.23.2.el6.x86_64 \
     splash=silent showopts cluster=cosmo ping=192.168.20.231 \
     lustreroot=192.168.20.231@o2ib:192.168.20.232@o2ib:\
                192.168.20.233@o2ib:192.168.20.234@o2ib:\
                /fs1/diskless/root-sl64-x86_64-class01 \
     nosoftlockup intel_idle.max_cstate=0 \
     mce=ignore_mce idle=poll

label sl64-class01-serial
  menu Diskless class01 node (serial console)
  kernel vmlinuz-2.6.32-358.23.2.el6.x86_64
  append initrd=initramfs-2.6.32-358.23.2.el6.x86_64 \
     splash=silent showopts cluster=cosmo ping=192.168.20.231 \
     lustreroot=192.168.20.231@o2ib:192.168.20.232@o2ib:\
                192.168.20.233@o2ib:192.168.20.234@o2ib:\
                /fs1/diskless/root-sl64-x86_64-class01 \
     nosoftlockup intel_idle.max_cstate=0 \
     mce=ignore_mce idle=poll \
     vga=normal console=ttyS0,115200n8 shell9
```

W pliku zdefiniowano klauzule *sl64-class01* (standardowy boot) i *sl64-class01-serial* w której komunikaty jądra wyświetlane są na porcie HP iLO4 Virtual Serial Port (COM1 w Linux).

Przekopiowanie do katalogu */tftpboot/sl/6.4/x86_64/class01* jądra systemu Linux (z VM *wo1*) *vmlinuz-2.6.32-358.23.2.el6.x86_64* oraz RAM filesystemu (utworzonego za pomocą oprogramowania oneSIS) *initramfs-2.6.32-358.23.2.el6.x86_64* kończy konfigurację PXE.

7.4 Konfiguracja DHCP

Na serwerach *adm1* i a*dm2* uruchomiona jest usługa DHCP wykorzystywana przez uruchamiające się WO lub dla instalacji po sieci pozostałych serwerów.

W celu instalacji serwisu DHCP na obu komputerach należy wykonać polecenie:

```
# yum install dhcp
```

W celu aktywacji i uruchomienia serwisu DHCP na obu komputerach należy wykonać polecenia:

```
# chkconfig dhcpd on
# service dhcpd start
```

Po modyfikacji konfiguracji na jednym z serwerów, należy zreplikować konfigurację na drugi serwer (modyfikacje można wykonywać na dowolnym serwerze), np:

```
adm1# /etc/dhcp/build_dhcp.sh
Synchronizing /etc/dhcp to adm2
...
```

a następnie zrestartowac serwis DHCP na obu serwerach:

```
adm1# service dhcpd restart
adm2# service dhcpd restart
```

Na obu komputerach konfiguracja znajduje się w katalogu */etc/dhcp*:

```
adm1# ls -F1 /etc/dhcp
build_dhcp.sh*          skrypt replikujący konfigurację pomiędzy serwerami
cosmo/                  katalog z plikami konfiguracyjnymi dla WO/HP iLO4 WO
dhcpd-adm1.conf         główny plik konfiguracyjny dla serwera adm1
dhcpd-adm2.conf         główny plik konfiugracyjny dla serwera adm2
```

```
adm1# ls -F1 /etc/dhcp/cosmo
dhcpd-kvm.conf          konfiguracja dla instalacji VM KVM (adm,mon,hn,wo1)
dhcpd-wo-adm.conf       konfiguracja dla HP iLO4 WO w sieci adm
dhcpd-wo-int.conf       konfiguracja dla uruchamiających się WO w sieci int
dhcpd-ws.conf           konfiguracja dla instalacji cfs1-cfs4,dp i mpdp
```

Każdy serwis DHCP uruchamia się ze swojego pliku konfiguracyjnego:

```
adm1# cat /etc/sysconfig/dhcpd
DHCPDARGS="-cf /etc/dhcp/dhcpd-adm1.conf"

adm1# ps -ef | grep dhcpd
dhcpd  53882  1  0  Feb15  ?  00:00:03 /usr/sbin/dhcpd \
  -user dhcpd -group dhcpd -cf /etc/dhcp/dhcpd-adm1.conf

adm2# cat /etc/sysconfig/dhcpd
DHCPDARGS="-cf /etc/dhcp/dhcpd-adm2.conf"

adm2# ps -ef | grep dhcpd
dhcpd  58504  1  0  Feb15 ?   00:00:02 /usr/sbin/dhcpd \
  -user dhcpd -group dhcpd -cf /etc/dhcp/dhcpd-adm2.conf
```

Plik *dhcpd-adm1.conf* jest głównym plikiem konfiguracyjnym DHCP dla serwera *adm1*:

```
adm1# cat /etc/dhcp/dhcpd-adm1.conf
authoritative;
ignore client-updates;

subnet 192.168.20.0 netmask 255.255.252.0 {
  option subnet-mask 255.255.252.0;
  option ip-forwarding off;
}
subnet 192.168.24.0 netmask 255.255.252.0 {
  option subnet-mask 255.255.252.0;
  option ip-forwarding off;
}
subnet 192.168.28.0 netmask 255.255.252.0 {
  option subnet-mask 255.255.252.0;
  option ip-forwarding off;
}
group {
  next-server 192.168.28.241;
  filename "/sl/6.4/x86_64/pxelinux.0";
  include "/etc/dhcp/cosmo/dhcpd-ws.conf";
  include "/etc/dhcp/cosmo/dhcpd-kvm.conf";
}
group {
  include "/etc/dhcp/cosmo/dhcpd-wo-adm.conf";
}
group {
  use-host-decl-names on;
  next-server 192.168.28.241;
  filename "/sl/6.4/x86_64/class01/pxelinux.0";
  include "/etc/dhcp/cosmo/dhcpd-wo-int.conf";
}
```

Plik *dhcpd-adm2.conf* rożni się wpisem ***next-server*** ustawionym na ***192.168.28.242***.

Plik *dhcpd-wo-adm.conf* ma nstępującą strukturę:

```
# c7000-01 (06)
host mpnode001 {hardware ethernet 9C:B6:54:94:55:8E; fixed-address 192.168.24.1;  }
...
host mpnode006 {hardware ethernet 9C:B6:54:94:65:33; fixed-address 192.168.24.6;  }
# c7000-02 (12)
host mpnode007 {hardware ethernet 9C:B6:54:94:65:B0; fixed-address 192.168.24.7;  }
...
host mpnode018 {hardware ethernet 9C:B6:54:94:39:99; fixed-address 192.168.24.18; }
# c7000-03 (16)
host mpnode019 {hardware ethernet 9C:B6:54:94:A9:12; fixed-address 192.168.24.19; }
...
host mpnode034 {hardware ethernet 9C:B6:54:94:05:5D; fixed-address 192.168.24.34; }
# c7000-04 (16)
host mpnode035 {hardware ethernet 9C:B6:54:94:54:1A; fixed-address 192.168.24.35; }
...
host mpnode050 {hardware ethernet 9C:B6:54:94:29:D0; fixed-address 192.168.24.50; }
# c7000-05 (16)
host mpnode051 {hardware ethernet 9C:B6:54:94:29:C9; fixed-address 192.168.24.51; }
...
host mpnode066 {hardware ethernet F0:92:1C:09:68:EC; fixed-address 192.168.24.66; }
# c7000-06 (16)
host mpnode067 {hardware ethernet F0:92:1C:09:D8:06; fixed-address 192.168.24.67; }
...
host mpnode082 {hardware ethernet 9C:B6:54:94:29:3C; fixed-address 192.168.24.82; }
# c7000-07 (16)
host mpnode083 {hardware ethernet F0:92:1C:09:78:41; fixed-address 192.168.24.83; }
...
host mpnode098 {hardware ethernet F0:92:1C:09:D8:97; fixed-address 192.168.24.98; }
# c7000-08 (16)
host mpnode099 {hardware ethernet 9C:B6:54:94:A4:EE; fixed-address 192.168.24.99; }
...
host mpnode114 {hardware ethernet 9C:B6:54:94:65:A9; fixed-address 192.168.24.114;}
# c7000-09 (12)
host mpnode115 {hardware ethernet 9C:B6:54:94:29:F2; fixed-address 192.168.24.115;}
...
host mpnode126 {hardware ethernet 9C:B6:54:94:29:5A; fixed-address 192.168.24.126;}
# c7000-10 (13)
host mpnode127 {hardware ethernet F0:92:1C:09:98:5F; fixed-address 192.168.24.127;}
...
host mpnode139 {hardware ethernet 9C:B6:54:94:99:B1; fixed-address 192.168.24.139;}
```

Plik *dhcpd-wo-int.conf* ma nstępującą strukturę:

```
# c7000-01 (06)
host node001 {hardware ethernet 9c:b6:54:9b:d3:60; fixed-address 192.168.28.1;  }
...
host node006 {hardware ethernet 9c:b6:54:9a:04:80; fixed-address 192.168.28.6;  }
# c7000-02 (12)
host node007 {hardware ethernet 9c:b6:54:99:6f:98; fixed-address 192.168.28.7;  }
...
host node018 {hardware ethernet 9c:b6:54:9a:df:98; fixed-address 192.168.28.18; }
# c7000-03 (16)
host node019 {hardware ethernet 9c:b6:54:9b:81:68; fixed-address 192.168.28.19; }
...
host node034 {hardware ethernet 9c:b6:54:9b:c2:e8; fixed-address 192.168.28.34; }
# c7000-04 (16)
host node035 {hardware ethernet 9c:b6:54:9a:9f:58; fixed-address 192.168.28.35; }
...
host node050 {hardware ethernet 9c:b6:54:99:3e:d0; fixed-address 192.168.28.50; }
# c7000-05 (16)
host node051 {hardware ethernet 9c:b6:54:9b:00:e0; fixed-address 192.168.28.51; }
...
host node066 {hardware ethernet 9c:b6:54:9b:92:60; fixed-address 192.168.28.66; }
# c7000-06 (16)
host node067 {hardware ethernet f0:92:1c:14:45:d8; fixed-address 192.168.28.67; }
...
host node082 {hardware ethernet 9c:b6:54:9b:80:68; fixed-address 192.168.28.82; }
# c7000-07 (16)
host node083 {hardware ethernet 9c:b6:54:9a:1f:78; fixed-address 192.168.28.83; }
...
host node098 {hardware ethernet 9c:b6:54:9b:a0:48; fixed-address 192.168.28.98; }
# c7000-08 (16)
host node099 {hardware ethernet 9c:b6:54:9b:c0:a8; fixed-address 192.168.28.99; }
...
host node114 {hardware ethernet 9c:b6:54:9a:9f:e0; fixed-address 192.168.28.114;}
# c7000-09 (12)
host node115 {hardware ethernet 9c:b6:54:9b:31:38; fixed-address 192.168.28.115;}
...
host node126 {hardware ethernet 9c:b6:54:9a:20:78; fixed-address 192.168.28.126;}
# c7000-10 (13)
host node127 {hardware ethernet 9c:b6:54:99:de:f0; fixed-address 192.168.28.127;}
...
host node139 {hardware ethernet 9c:b6:54:9b:53:48; fixed-address 192.168.28.139;}
```

Plik *dhcpd-kvm.conf* ma następującą zawartość:

```
adm1# cat /etc/dhcp/cosmo/dhcpd-kvm.conf
host adm { hardware ethernet 52:54:00:d7:6b:7e; fixed-address 192.168.28.243; }
host mon { hardware ethernet 52:54:00:ea:8c:c6; fixed-address 192.168.28.244; }
host hn  { hardware ethernet 52:54:00:99:8f:ab; fixed-address 192.168.28.245; }
host wo1 { hardware ethernet 52:54:00:81:93:61; fixed-address 192.168.28.246; }
```

Konfiguracja aktualnie nie używana.

Plik *dhcpd-ws.conf* ma następującą zawartość:

```
adm1# cat /etc/dhcp/cosmo/dhcpd-ws.conf
host cfs1 { hardware ethernet 9c:b6:54:9b:90:b8; fixed-address 192.168.28.231; }
host cfs2 { hardware ethernet 9c:b6:54:9a:d2:b8; fixed-address 192.168.28.232; }
host cfs3 { hardware ethernet 9c:b6:54:99:be:18; fixed-address 192.168.28.233; }
host cfs4 { hardware ethernet 9c:b6:54:9b:60:10; fixed-address 192.168.28.234; }
host   dp { hardware ethernet 24:be:05:9a:18:01; fixed-address 192.168.28.221; }
host mpdp { hardware ethernet 9c:b6:54:08:07:e0; fixed-address 192.168.24.221; }
```

Konfiguracja aktualnie używana tylko do adresowania HP iLO4 *mpdp*.

7.5 Konfiguracja DNS

Na serwerach *adm1* i a*dm2* uruchomiona jest usługa DNS wykorzystywana wewnętrznie przez klaster do rozwiązywania nazw. Na potrzeby wewnętrzne klastra zdefiniowano domenę DNS *cosmo.local*.

W celu instalacji serwisu DNS na obu komputerach należy wykonać polecenie:

```
# yum install bind
```

W celu aktywacji i uruchomienia serwisu DNS na obu komputerach należy wykonać polecenia:

```
# chkconfig named on
# service named start
```

Po modyfikacji konfiguracji na jednym z serwerów, należy zreplikować konfigurację na drugi serwer (modyfikacje można wykonywać na dowolnym serwerze), np:

```
adm1# /var/named/build_named.sh
Synchronizing /var/named to adm2
...
```

a następnie zrestartować serwis DNS na obu serwerach:

```
adm1# service named restart
adm2# service named restart
```

Na obu komputerach konfiguracja znajduje się w katalogu */var/named*:

```
adm1# ls -F1 /var/named
build_named.sh*          skrypt replikujący konfigurację pomiędzy serwerami
P/                       katalog z plikami konfiguracyjnymi domeny cosmo.local
```

```
adm1# ls -F1 /var/named/P
build_wo.sh              sktypt generujacy konfiguracje DNS dla WO
db.192.168               konfiguracja odwrotnych mapowań dla 192.168.0.0
db.192.168-wo-*          konfiguracja odwrotnych mapowań dla WO
db.cosmo                 konfiguracja domeny cosmo.local
db.cosmo-wo-*            konfiguracja WO w domenie cosmo.local
```

Plik */etc/named.conf* jest głównym plikiem konfiguracyjnym DNS:

```
adm1# cat /etc/named.conf
```

```
options {
     directory "/var/named";

     forward only;
     forwarders {
        10.91.91.13;
        10.91.91.14;
     };
};
zone "cosmo.local" {
     type master;
     file "P/db.cosmo";
};
zone "168.192.in-addr.arpa" {
     type master;
     file "P/db.192.168";
};
```

Plik jest taki sam na obu serwerach. Nazwy które nie mogą być rozwiązane lokalnie przez serwis DNS przesyłana są do serwerów 10.91.91.13-14.

W celu konfiguracji klienckiej na każdym serwerze klastra utworzono plik */etc/resolv.conf*:

```
adm1# cat /etc/resolv.conf
domain cosmo.local
nameserver 192.168.28.241
nameserver 192.168.28.242
```

7.6 Konfiguracja NTP

Na serwerach *adm1* i a*dm2* uruchomiona jest usługa NTP wykorzystywana wewnętrznie przez klaster do synchronizacji czasu.

W celu uruchomienia usługi NTP na *adm1/adm2* wykonano:

```
# cd /etc
# mv ntp.conf ntp.conf.ORG
# vi ntp.conf
server 10.91.91.13 prefer
server 10.91.91.14
# service ntpdate start
# service ntpd start
# chkconfig ntpdate on
# chkconfig ntpd on
# ntpq -p
      remote           refid  st t when poll reach  delay   offset  jitter
==============================================================================
*dns1.imgw.pl 212.244.36.227   2 u   28   64   377  1.218    1.954   0.270
+dns2.imgw.pl 212.244.36.228   2 u   25   64   377  1.247    1.999   0.299
```

W celu uruchomienia usługi NTP na pozostałych serwerach klastra wykonano:

```
# cd /etc
# mv ntp.conf ntp.conf.ORG
# vi ntp.conf
server 192.168.28.241 prefer
server 192.168.28.242
# service ntpdate start
# service ntpd start
# chkconfig ntpdate on
# chkconfig ntpd on
# ntpq -p
         remote       refid  st t when poll reach  delay   offset  jitter
==============================================================================
*adm1.cosmo.loca 10.91.91.13   3 u   21   64   377  0.209   -4.115   2.044
+adm2.cosmo.loca 10.91.91.13   3 u   14   64   377  0.181    0.101   2.964
```

Rozdział 8: Wspólny system plików

WSP udostępnia przestrzeń dyskową dla Węzłów Obliczeniowych i serwera backupu.

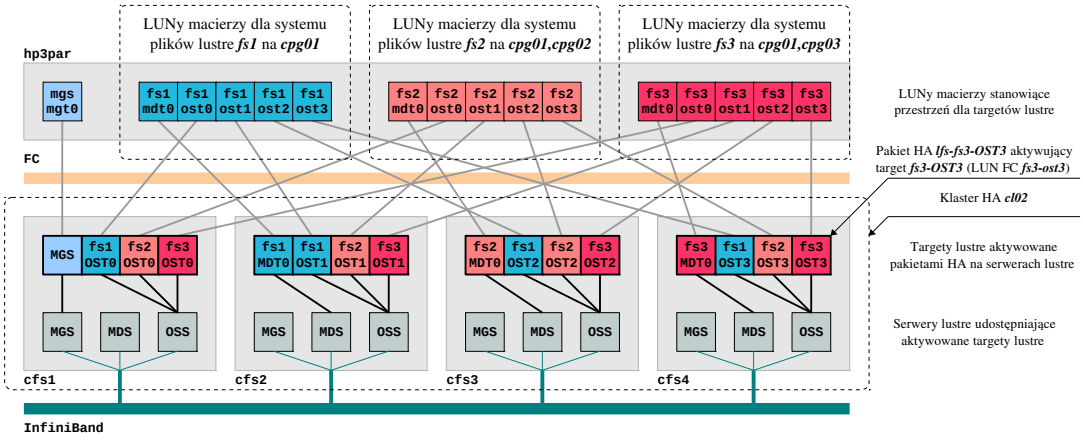
8.1 Architektura

WSP zbudowany jest w oparciu o następujące elementy:

- macierz HP 3PAR 7400
- dwa przełączniki FC 8/24c realizujące sieć FC
- cztery serwery HP BL460c Gen8 (oznaczone jako *cfs1–cfs4*) dołączone do sieci IB z systemem operacyjnym Scientific Linux 6.4 x86_64 i oprogramowaniem klastra HA
- oprogramowanie *Lustre* w wersji 2.4.2

Serwery WSP nie posiadają dysków wewnętrznych i uruchamiają się z LUNów wystawionych z macierzy HP 3PAR 7400. Po uruchomieniu formują czterowęzłowy klastr HA *cl02*.

Targety systemu plików Lustre (MGT,MDT,OST) mogą być aktywowane na dowolnym serwerze WSP (wszystkie LUNy macierzy dostępne są na wszystkich serwerach WSP). Dla każdego targetu Lustre zdefiniowany jest oddzielny pakiet klastra HA który nim zarządza (aktywuje, deaktywuje, przełącza).



Architektura Wspólnego Systemu Plików

Reguły tworzenia targetów Lustre:

- wielkość targetu MGT wynosi 1GB (jeden target MGT na cały WSP, zdefiniowany na dedykowanym LUNie)
- wielkość targetu MDT dla danego systemu plików wynosi 1% jego wielkości (jeden target MDT na system plików)
- wielkość targetu OST dla danego systemu plików wynosi 25% jego wielkości (założono cztery targety OST na system plików)

Parametry zdefiniowanych systemów plików Lustre:

Lustre FS	*Wielkość*	*7400 CPG*	*Opis*	*Uwagi*
fs1	200GB	cpg01	Przestrzeń administracyjna	Systemy główne (*rootfs*) WO.
fs2	34TB	cpg01,cpg02	Przestrzeń produkcyjna	
fs3	34TB	cpg01,cpg03	Przestrzeń backup	

Parametry targetów zdefiniowanych systemów plików Lustre:

Target	*Pakiet HA*	*Punkt Montowania Targetu*	*7400 VV*
MGS	lfs-mgs-MGT0000	/lustre/fs/mgs/MGT0000	mgs-mgt0000
fs1-MDT0000	lfs-fs1-MDT0000	/lustre/fs/fs1/MDT0000	fs1-mdt0000
fs1-OST0000	lfs-fs1-OST0000	/lustre/fs/fs1/OST0000	fs1-ost0000
fs1-OST0001	lfs-fs1-OST0001	/lustre/fs/fs1/OST0001	fs1-ost0001
fs1-OST0002	lfs-fs1-OST0002	/lustre/fs/fs1/OST0002	fs1-ost0002
fs1-OST0003	lfs-fs1-OST0003	/lustre/fs/fs1/OST0003	fs1-ost0003
fs2-MDT0000	lfs-fs2-MDT0000	/lustre/fs/fs2/MDT0000	fs2-mdt0000
fs2-OST0000	lfs-fs2-OST0000	/lustre/fs/fs2/OST0000	fs2-ost0000
fs2-OST0001	lfs-fs2-OST0001	/lustre/fs/fs2/OST0001	fs2-ost0001
fs2-OST0002	lfs-fs2-OST0002	/lustre/fs/fs2/OST0002	fs2-ost0002
fs2-OST0003	lfs-fs2-OST0003	/lustre/fs/fs2/OST0003	fs2-ost0003
fs3-MDT0000	lfs-fs3-MDT0000	/lustre/fs/fs3/MDT0000	fs3-mdt0000
fs3-OST0000	lfs-fs3-OST0000	/lustre/fs/fs3/OST0000	fs3-ost0000
fs3-OST0001	lfs-fs3-OST0001	/lustre/fs/fs3/OST0001	fs3-ost0001
fs3-OST0002	lfs-fs3-OST0002	/lustre/fs/fs3/OST0002	fs3-ost0002
fs3-OST0003	lfs-fs3-OST0003	/lustre/fs/fs3/OST0003	fs3-ost0003

8.2 Konfiguracja serwerów Lustre

8.2.1 Wielościeżkowość

W celu zapewnienia HA (na poziomie redundantnych scieżek do dysków macierzy) na serwerach CFS został skonfigurowany mechanizm wielościeżkowosci *mutipath*. Konfiguracja wielościeżkowosci wykonywana jest poprzez edycje pliku */etc/multipath.conf*.

Macierz HP 3PAR 7400 wymaga dodania do pliku konfiguracyjnego, w odpowiednich sekcjach, następujących wpisów:

```
...
defaults {
     user_friendly_names yes
     polling_interval 10
     max_fds 8192
}
devices {
  device {
     vendor "3PARdata"
     product "VV"
     no_path_retry 18
     features "0"
     hardware_handler "0"
     path_grouping_policy multibus
     getuid_callout "/lib/udev/scsi_id --whitelisted --device=/dev/%n"
     path_selector "round-robin 0"
     rr_weight uniform
     rr_min_io_rq 1
     path_checker tur
     failback immediate
  }
}
...
```

Powyższe modyfikacje są wymagane dla systemu RHEL ≥ 6.2 i HP 3PAR OS ≥ 3.1.1.

W celu listowania dysków wirtualnych macierzy (VV) z poziomu systemu operacyjnego Linux należy na wszystkich serwerach zainstalować narzędzie HP3PARInfo:

```
 # cd /var/tmp/sw
# mkdir HP3PARInfo
# cd HP3PARInfo
# tar xvf ../HP3PARInfo.tar
 # ./unix_local_install.sh
 HP3PARInfo is not installed on this server
 Do you want to install HP3PARInfo v1.2? y or n [y]: y
 Installing HP3PARInfo...
 Copying files...
 Successfully installed HP3PARInfo v1.2
 # ln -s /usr/bin/HP3PARInfo /usr/bin/hp3parinfo
 # hp3parinfo -v
 Linux HP3PARInfo Version 1.2 Jun 10 2013
 Copyright 2012-2013 Hewlett-Packard Development Company, L.P.
```

W celu identyfikacji wszystkich VV wystawionych z macierzy HP 3PAR do danego komputera należy wykonać polcecenie:

```
 # hp3parinfo -v
```

W celu wypisania informacji o dysku na konkretnej ścieżce należy wykonać polecenie:

```
 cfs1# hp3parinfo -f /dev/sda
Device File     : /dev/sda     Serial#              : 1621446
 Host Target     : 00           Code Rev             : 3.1.2 MU3
 Array LUN       : 01           LUN WWN  : 60002ac0000000000000004000053c6
 VV Name         : cfs1_disk01  Size                 : 20480 MB
 Domain Name     : -            Domain Id            : 0
 VV ID           : 4            User CPG Name        : cpg01
 Snap CPG Name   : -            Provisioning Type    : Thinly Provisioned
 TPVV Reclaim    : Supported    Reserved User Size   : 11264 MB
 Used User Size  : 7362 MB      TPVV Allocation Unit : 16 KB
 ATS Support     : Supported    XCOPY Support        : Supported
```

Każdy wirtualny wolumin macierzy (VV) został skonfigurowany z wykorzystaniem mechanizmu multipath z nadaniem aliasu w systemie Linux takiego jak nazwa VV w macierzy.

Na przykład w celu konfiguracji VV *cfs1_disk01* na komputerze *cfs1* należy w pliku konfiguracyjnym *multipath.conf* dodać w odpowiednich sekcjach wpisy:

```
blacklist_exceptions {
     wwid "360002ac0000000000000000004000053c6"
     ...
}
multipaths {
     multipath {
          uid 0
          gid 0
          wwid "360002ac0000000000000000004000053c6"
          mode 0600
          alias cfs1_disk01
     }
     ...
}
```

Po edycji należy aktywować konfigurację poleceniem:

```
cfs1# multipath
```

W celu wylistowania konfiguracji *multipath* należy wykonać polecenie:

```
cfs1# multipath -ll
cfs1_disk01 (360002ac0000000000000000004000053c6) dm-0 3PARdata,VV
size=20G features='1 queue_if_no_path' hwhandler='0' wp=rw
`-+- policy='round-robin 0' prio=1 status=active
  |- 1:0:2:1  sdak 66:64   active ready running
  |- 1:0:3:1  sdbc 67:96   active ready running
  |- 2:0:0:1  sdbu 68:128  active ready running
  |- 1:0:0:1  sda  8:0     active ready running
  |- 1:0:1:1  sds  65:32   active ready running
  |- 2:0:1:1  sdcm 69:160  active ready running
  |- 2:0:2:1  sdde 70:192  active ready running
  `- 2:0:3:1  sddw 71:224  active ready running
...
```

8.2.2 Instalacja oprogramowania serwerowego Lustre

Na oprogramowanie serwerowe Lustre składa się odpowiednio zmodyfikowane jądro systemu, moduły jądra systemu i programy narzędziowe.

Instalacja oprogramowania:

(instalacja wymaganych systemowych pakietów RPM)

```
 # yum install net-snmp-libs openmpi sg3_utils
```

(uaktualnienie e2fsprogs)

```
 # rpm -qa | grep e2fsprogs
 e2fsprogs-libs-1.41.12-14.el6.x86_64
 e2fsprogs-1.41.12-14.el6.x86_64
# cd /var/tmp/sw/lustre/e2fsprogs-1.42.7.wc2/el6/RPMS/x86_64
 # rpm -Uvh --test *.rpm
 # rpm -Uvh *.rpm
# rpm -qa | grep e2fsprogs
 e2fsprogs-static-1.42.7.wc2-7.el6.x86_64
 e2fsprogs-1.42.7.wc2-7.el6.x86_64
 e2fsprogs-devel-1.42.7.wc2-7.el6.x86_64
 e2fsprogs-debuginfo-1.42.7.wc2-7.el6.x86_64
 e2fsprogs-libs-1.42.7.wc2-7.el6.x86_64
```

(instalacja jądra Linux zmodyfikowanego dla Lustre)

```
 # cd /var/tmp/sw/lustre/lustre-2.4.2/el6/server/RPMS/x86_64
# rpm -ivh --test kernel-2.6.32-358.23.2.el6_lustre.x86_64.rpm \
      kernel-firmware-2.6.32-358.23.2.el6_lustre.x86_64.rpm
 # rpm -ivh kernel-2.6.32-358.23.2.el6_lustre.x86_64.rpm \
      kernel-firmware-2.6.32-358.23.2.el6_lustre.x86_64.rpm
```

(instalacja jądra Linux: zainstalować pakiety jeżeli będzie doinstalowywane gcc)

```
 # rpm -ivh --test \
      kernel-headers-2.6.32-358.23.2.el6_lustre.x86_64.rpm
 # rpm -ivh kernel-headers-2.6.32-358.23.2.el6_lustre.x86_64.rpm
```

(instalacja jądra Linux: zainstalować pakiety jeżeli jest taka potrzeba)

```
 # rpm -ivh --test kernel-devel-2.6.32-358.23.2.el6_lustre.x86_64.rpm
 # rpm -ivh kernel-devel-2.6.32-358.23.2.el6_lustre.x86_64.rpm
```

(instalacja serwerowych modułów jądra Lustre)

```
# rpm -ivh --test \
lustre-modules-2.4.2-2.6.32_358.23.2.el6_lustre.x86_64.x86_64.rpm \
lustre-ldiskfs-4.1.0-2.6.32_358.23.2 .el6_lustre.x86_64.x86_64.rpm
# rpm -ivh \
lustre-modules-2.4.2-2.6.32_358.23.2.el6_lustre.x86_64.x86_64.rpm \
lustre-ldiskfs-4.1.0-2.6.32_358.23.2.el6_lu stre.x86_64.x86_64.rpm
```

(instalacja serwerowych narzędzi Lustre)

```
# rpm -ivh --test \
lustre-2.4.2-2.6.32_358.23.2.el6_lustre.x86_64.x86_64.rpm \
lustre-osd-ldiskfs-2.4.2-2.6.32_358.23.2.el6_lustre.x86_64.x86_64.rpm
# rpm -ivh \
lustre-2.4.2-2.6.32_358.23.2.el6_lustre.x86_64.x86_64.rpm \
lustre-osd-ldiskfs-2.4.2-2.6.32_358.23.2.el6_lustre.x86_64.x86_64.rpm
```

(instalacja dodatkowych narzędzi Lustre)

```
# rpm -ivh --test \
lustre-tests-2.4.2-2.6.32_358.23.2.el6_lustre.x86_64.x86_64.rpm \
lustre-iokit-1.4.0-1.noarch.rpm
# rpm -ivh \
lustre-tests-2.4.2-2.6.32_358.23.2.el6_lustre.x86_64.x86_64.rpm \
lustre-iokit-1.4.0-1.noarch.rpm
# rpm -e perf-2.6.32-358.el6.x86_64
# rpm -ivh --test \
perf-2.6.32-358.23.2.el6_lustre.x86_64.rpm \
python-perf-2.6.32-358.23.2.el6_lustre.x86_64.rpm
# rpm -ivh \
perf-2.6.32-358.23.2.el6_lustre.x86_64.rpm \
python-perf-2.6.32-358.23.2.el6_lustre.x86_64.rpm
```

(wylistowanie zainstalowanych pakietów)

```
# rpm -qa | grep lustre
```

(konfiguracja sieci LNET po Inifiniband)

```
# cat /etc/modprobe.d/lustre.conf
options lnet networks=o2ib(ib0)
```

8.3 Konfiguracja klientów Lustre

8.3.1 Instalacja oprogramowania klienckiego Lustre

Na oprogramowanie klienckie Lustre składają się moduły jądra systemu i programy narzędziowe.

Instalacja oprogramowania:

(instalacja wymaganych systemowych pakietów RPM)

```
# yum install net-snmp-libs openmpi sg3_utils
```

(instalacja zwykłego jądra systemu ale W TEJ SAMEJ WERSJI co dla serwera)

```
# cd /var/tmp/sw/kernels
# rpm -ivh --test \
    kernel-2.6.32-358.23.2.el6.x86_64.rpm \
    kernel-firmware-2.6.32-358.23.2.el6.noarch.rpm
# rpm -ivh \
    kernel-2.6.32-358.23.2.el6.x86_64.rpm \
    kernel-firmware-2.6.32-358.23.2.el6.noarch.rpm
```

(instalacja klienckich modułów jądra Lustre)

```
# cd /var/tmp/sw/lustre/lustre-2.4.2/el6/client/RPMS/x86_64
# rpm -ivh --test \
    lustre-client-modules-2.4.2-2.6.32_358.23.2.el6.x86_64.x86_64.rpm
# rpm -ivh \
    lustre-client-modules-2.4.2-2.6.32_358.23.2.el6.x86_64.x86_64.rpm
```

(instalacja klienckich narzędzi Lustre)

```
# rpm -ivh --test \
    lustre-client-2.4.2-2.6.32_358.23.2.el6.x86_64.x86_64.rpm
# rpm -ivh \
    lustre-client-2.4.2-2.6.32_358.23.2.el6.x86_64.x86_64.rpm
```

(instalacja klienckich narzędzi Lustre)

```
# rpm -ivh --test \
    lustre-client-2.4.2-2.6.32_358.23.2.el6.x86_64.x86_64.rpm
# rpm -ivh \
    lustre-client-2.4.2-2.6.32_358.23.2.el6.x86_64.x86_64.rpm
```

(instalacja dodatkowych narzędzi Lustre)

```
# rpm -ivh --test \
    lustre-tests-2.4.2-2.6.32_358.23.2.el6_lustre.x86_64.x86_64.rpm \
    lustre-iokit-1.4.0-1.noarch.rpm
# rpm -ivh \
    lustre-tests-2.4.2-2.6.32_358.23.2.el6_lustre.x86_64.x86_64.rpm \
    lustre-iokit-1.4.0-1.noarch.rpm
```

(wylistowanie zainstalowanych pakietów)

```
# rpm -qa | grep lustre
lustre-client-modules-2.4.2-2.6.32_358.23.2.el6.x86_64.x86_64
lustre-client-2.4.2-2.6.32_358.23.2.el6.x86_64.x86_64
lustre-client-tests-2.4.2-2.6.32_358.23.2.el6.x86_64.x86_64
lustre-iokit-1.4.0-1.noarch
```

(konfiguracja sieci LNET po Inifiniband)

```
# cat /etc/modprobe.d/lustre.conf
options lnet networks=o2ib(ib0)
```

8.4 Procedury administracyjne Lustre

W celu ręcznego zamontowania systemu plików Lustre należy wykonać polecenia:

```
# mkdir -p /cfs/fs1 /cfs/fs2 /cfs/fs3
# mount -t lustre 192.168.20.231@o2ib:192.168.20.232@o2ib:\
    192.168.20.233@o2ib:192.168.20.234@o2ib:/fs1 /cfs/fs1
# mount -t lustre 192.168.20.231@o2ib:192.168.20.232@o2ib:\
    192.168.20.233@o2ib:192.168.20.234@o2ib:/fs2 /cfs/fs2
# mount -t lustre 192.168.20.231@o2ib:192.168.20.232@o2ib:\
    192.168.20.233@o2ib:192.168.20.234@o2ib:/fs3 /cfs/fs3
```

W celu ręcznego zamontowania systemu plików Lustre z ***/etc/fstab*** należy wykonać polecenia:

```
# mkdir -p /cfs/fs1 /cfs/fs2 /cfs/fs3
# vi /etc/fstab
...
192.168.20.231@o2ib:192.168.20.232@o2ib:\
    192.192.168.20.233@o2ib:192.168.20.234@o2ib:/fs1 /cfs/fs1 \
    lustre defaults,noauto,localflock 1 1
192.168.20.231@o2ib:192.168.20.232@o2ib:\
    192.168.20.233@o2ib:192.168.20.234@o2ib:/fs2 /cfs/fs2 \
    lustre defaults,noauto 1 1
192.168.20.231@o2ib:192.168.20.232@o2ib:\
    192.168.20.233@o2ib:192.168.20.234@o2ib:/fs3 /cfs/fs3 \
    lustre defaults,noauto 1 1
# mount /cfs/fs1
# mount /cfs/fs2
# mount /cfs/fs3
```

W celu odmontowania systemu plików Lustre należy wykonać polecenia:

```
# umount /cfs/fs1
# umount /cfs/fs2
# umount /cfs/fs3
```

W celu weryfikacji stanu serwerów MDS/OSS zamontowanego systemu plików Lustre należy wykonać polecenia:

```
# df -t lustre -h
Filesystem            Size  Used Avail Use% Mounted on
192.168.20.231@o2ib:192.168.20.232@o2ib:192.168.20.233@o2ib:192.168.20.234@o2ib:/fs1
                      200G  7.9G  182G   5% /cfs/fs1
192.168.20.231@o2ib:192.168.20.232@o2ib:192.168.20.233@o2ib:192.168.20.234@o2ib:/fs2
                       32T  976G   30T   4% /cfs/fs2
192.168.20.231@o2ib:192.168.20.232@o2ib:192.168.20.233@o2ib:192.168.20.234@o2ib:/fs3
                       32T  977G   30T   4% /cfs/fs3
# lfs check servers
fs1-OST0003-osc-ffff88040a626800: active
fs1-OST0000-osc-ffff88040a626800: active
fs1-OST0002-osc-ffff88040a626800: active
fs1-MDT0000-mdc-ffff88040a626800: active
fs1-OST0001-osc-ffff88040a626800: active
fs2-OST0000-osc-ffff880411071400: active
fs2-OST0001-osc-ffff880411071400: active
fs2-OST0002-osc-ffff880411071400: active
fs2-OST0003-osc-ffff880411071400: active
fs2-MDT0000-mdc-ffff880411071400: active
fs3-OST0003-osc-ffff880414c6c400: active
fs3-OST0000-osc-ffff880414c6c400: active
fs3-OST0002-osc-ffff880414c6c400: active
fs3-MDT0000-mdc-ffff880414c6c400: active
fs3-OST0001-osc-ffff880414c6c400: active
```

W celu wylistowania sieci LNET należy wykonać polecenie:

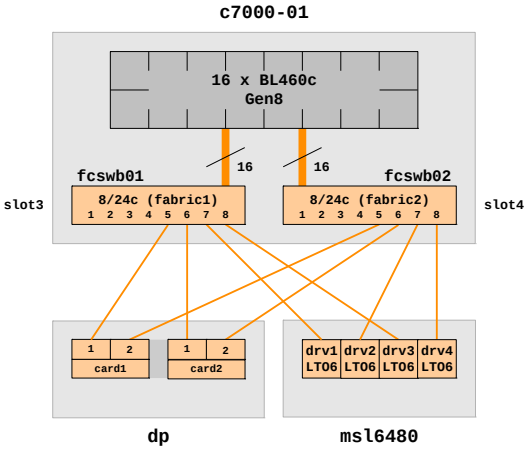
```
# lctl list_nids
192.168.20.243@o2ib
```

Rozdział 9: Backup

9.1 Architektura

System wykonywania kopii zapasowych działa w dwóch krokach:

- Cykliczna synchronizacja systemu plików Lustre *fs2* (przestrzeń produkcyjna) z *fs3* (przestrzeń backupu) za pomocą programu *lustre_rsync* na komputerze *dp*.
- Backup systemu plików *fs3* za pomocą oprogramowania HP Data Protector 8.0



Podłączenie serwera backupu do sieci FC

Serwer backupu *dp* podłączony jest do sieci FC, IB, wewnętrznej (*int*) i zarządzającej (*adm*).

Dodatek A: Wstępna konfiguracja systemu ScientificLinux 6.4

Po instalacji systemu SL 6.4 należy wykonać wstępną konfigurację:

(usunięcie logów instalacyjnych)

```
# rm /root/anaconda-ks.cfg
# rm /root/install.log*
```

(stworzenie dodatkowych katalogów)

```
# mkdir /cdrom
```

(modyfikacja środowiska)

```
# vi /etc/profile.d/local.sh
```

```
PS1='\h\$ '

alias rm='rm -i'
alias cp='cp -i'
alias mv='mv -i'

alias lsip='/sbin/ifconfig -a|egrep "Link|inet|^$"'
alias lspf='/sbin/iptables -L -v -n|grep -v limit'
alias lsis='netstat -an -A inet'
alias lsha='clustat'
```

(zmiana runlevelu)

```
# init 3
# vi /etc/inittab
id:3:initdefault:
```

(modyfikacja grub)

```
# vi /boot/grub/menu.lst
#splashimage=(hd0,0)/grub/splash.xpm.gz
#hiddenmenu
...
module ... rhgb (usunac)
...
```

(zablokowanie SELINUX)

```
# vi /etc/selinux/config
SELINUX=disabled
```

(konfiguracja DNS)

```
# vi /etc/resolv.conf
domain cosmo.local
nameserver 192.168.28.241
nameserver 192.168.28.242
```

(zablokowanie serwisów)

```
# chkconfig cups off
# chkconfig certmonger off
# chkconfig autofs off
# chkconfig postfix off
# chkconfig cpuspeed off
# chkconfig iptables off
# chkconfig ip6tables off
# chkconfig kdump off
```

(konfiguracja yum: modyfikacja konfiguracji ogólnej)

```
# vi /etc/yum.conf
[main]
group_package_types=mandatory default optional
```

(konfiguracja yum: zablokowanie standardowych repozytoriów)

```
# cd /etc/yum.repos.d
modyfikacja definicji repozytoriów na: enabled=0
```

(konfiguracja yum: definicje lokalnych repozytoriów instalacyjnych)

```
# cd /etc/yum.repos.d
# vi sl-local.repo
[base-local]
name=SL $releasever - Base Local
baseurl=http://192.168.28.241/sw/linux/sl/6.4/x86_64/install-dvd
        http://192.168.28.242/sw/linux/sl/6.4/x86_64/install-dvd
gpgcheck=0
enabled=1
```

(instalacja dodatkowych pakietów RPM)

```
# yum install lsscsi
# yum install telnet
# yum install xorg-x11-xauth
# yum install ipmitool
# yum install firefox icedtea-web
```

(instalacja HP STK Tools – HP Scripting Toolkit Tools)

```
# yum install libstdc++-4.4.7-3.el6.i686
# yum install libxml2.i686 libzip.i686
# cd /var/tmp/sw
# rpm -ivh --test hp-scripting-tools-9.50-97.rhel6.i686.rpm
# rpm -ivh hp-scripting-tools-9.50-97.rhel6.i686.rpm
```

Dodatek B: Konfiguracja autoryzacji SSH

W celu autoryzacji użytkowników do logowania się bez hasła pomiędzy węzłami obliczeniowymi i maszyną *hn* należy wykonać na nich następującą konfigurację:

(modyfikacja konfiguracji serwera sshd)

```
# vi /etc/ssh /sshd_config
...
RhostsRSAAuthentication yes
HostbasedAuthentication yes
IgnoreRhosts no                          dla autoryzacji użytkownika root
...
# service sshd restart
```

(modyfikacja konfiguracji klienta ssh)

```
# vi /etc/ssh/ssh_config
...
Host *
   HostbasedAuthentication yes
   EnableSSHKeysign yes
   StrictHostKeyChecking no
...
```

(wygenerowanie na hn plików /etc/ssh/ssh_known_hosts, /etc/hosts.equiv i /root/.shosts)

```
hn# hpc_ssh_keyscan.sh
```

a następnie rozkopiować je na pozostałe maszyny.

Dodatek C: Instalacja Intel MPI Runtime

W celu instalacji Intel MPI Runtime należy wykonać:

```
# cd /var/tmp/sw/intel/mpi
# gzip -dc l_mpi-rt_p_4.1.3.045.tgz | tar xvf -
# cd l_mpi-rt_p_4.1.3.045
# ./install.sh
```

a następnie wybierać domyślne opcje.

Jeżeli podczas uruchamiania skryptu otrzymamy błąd:

```
wo1# ./install.sh
CPU is not supported.
```

oznacza to że nasz procesor nie jest wspierany przez Intel MPI.

W przypadku maszyny wirtualnej KVM *wo1* oznacza to zbyt słabą emulację procesora:

```
wo1# cat /proc/cpuinfo | grep -E 'model|family'
cpu family      : 6
model           : 13
model name      : QEMU Virtual CPU version (cpu64-rhel6)
```

Intel MPI wymaga co najmniej *cpu_family=6* i *model=14*.

W celu rozwiązania problemu w maszynie wirtualnej KVM *wo1* należy zmienić emulację procesora na SandyBridge:

```
wo1# cat /proc/cpuinfo | grep -E 'model|family'
cpu family      : 6
model           : 42
model name      : Intel Xeon E312xx (Sandy Bridge)
```

Dodatek D: Test HPLinpack

Test HPLinpack wymaga dobrania opowiednich parametrów, w przypadku klastra *cosmo* zdecydowano się na następujące ustawienia:

Parametr	*Wartość*	*Uwagi*
N	1231776	Wielkość zadania, zapewnia wykorzystanie 80% RAM klastra.
NB	224	Blocking factor.
PxQ	52x53	Siatka, 52*53=2756 węzłów MPI (procesów, wykorzystanych corów).

Parametr N definiuje wielkość zadania która w bezpośredni sposób zależy od ilości pamięci. Zdecydowano na dobranie parametru N tak by zadanie zajmowało 80% całej pamieci RAM WO biorących udział teście (każdy WO posiada 128GB RAM, do obliczeń wybrano 138 WO ze względu na wielkość siatki).

Całą pamieć WO w bajtach wyliczamy następująco:

```
# echo "128*1024*1024*1024*138" | bc
18966575579136
```

Każdy element macierzy równania liniowego jest zdefiniowany jako typ *double* który w jezyku C zajmuje 8 bajtów co daje nam następującą ilość zmiennych (dla 100%RAM WO):

```
# echo "18966575579136/8" | bc
2370821947392
```

Parametr N jest bokiem macierzy kwadratowej więc możemy go obliczyć jako:

```
# echo "sqrt(2370821947392)" | bc
1539747
```

Założono jednak wykorzystanie 80% pamięci RAM WO biorących udział w teście:

```
# echo "1539747*0.8" | bc
1231797.6
```

Dokonano optymalizacji N zapewniając że jest on wielokrotnością NB:

```
# echo "(1231797.6/224)*224" | bc
1231776
```

Parametry PxQ definiują wykorzystywaną siatkę która powinna być możliwie zbliżona do kwadratu. Z tego względu zdecydowano na wybór siatki 52x53=2756 (24 cory pozostają nie wykorzystane).

Plik konfiguracyjny testu HPLinpack wygląda następująco:

```
node001$ cat HPL-cosmo-138-128-0.8-224.dat
HPLinpack benchmark input file
Innovative Computing Laboratory, University of Tennessee
HPL.out      output file name (if any)
6            device out (6=stdout,7=stderr,file)
1            # of problems sizes (N)
1231776      Ns
1            # of NBs
224          NBs
0            PMAP process mapping (0=Row-,1=Column-major)
1            # of process grids (P x Q)
52           Ps
53           Qs
16.0         threshold
1            # of panel fact
2            PFACTs (0=left, 1=Crout, 2=Right)
1            # of recursive stopping criterium
4            NBMINs (>= 1)
1            # of panels in recursion
2            NDIVs
1            # of recursive panel fact.
1            RFACTs (0=left, 1=Crout, 2=Right)
1            # of broadcast
1            BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
1            # of lookahead depth
0            DEPTHs (>=0)
2            SWAP (0=bin-exch,1=long,2=mix)
64           swapping threshold
0            L1 in (0=transposed,1=no-transposed) form
0            U  in (0=transposed,1=no-transposed) form
1            Equilibration (0=no,1=yes)
8            memory alignment in double (> 0)
```

Testy HPLinpack były uruchamiane z serwera *hn* (head node) jako zlecenia managera kolejek jako użytkownik *admin*:

```
hn# su - admin
hn$ cat qm_hpl.sh
#!/bin/sh

#PBS -N hpl
#PBS -l nodes=138:ppn=20,walltime=240:00:00
#PBS -q batch

#set -x

. /opt/intel/impi/4.1.3/bin64/mpivars.sh

/bin/cp -f HPL-cosmo-138-128-0.8-224.dat HPL.dat

# Intel binary
# mpirun -n 2756 -ppn 20 -f $PBS_NODEFILE \
    /usr/local/bin/xhpl_intel64_dynamic

# Local binary
mpirun -n 2756 -ppn 20 -f $PBS_NODEFILE \
    /usr/local/bin/hpl_intel64_dynamic
hn$ qsub qm_hpl.sh
hn$ qrun 26.hn
hn$ qstat
Job ID              Name        User             Time Use S Queue
------------------ ----------- ---------------- -------- - -----
26.hn               hpl         admin            112:27:4 R batch
```

Podczas testów uzyskano następujące wyniki:

Test 01 – binaria HPL Intela (/usr/local/bin/xhpl_intel64_dynamic)

```
================================================================================
T/V                N    NB     P     Q               Time          Gflops
--------------------------------------------------------------------------------
WR01C2R4   1231776   224    52    53           20565.52     6.05850e+04
HPL_pdgesv() start time Fri Feb 14 23:16:18 2014
HPL_pdgesv() end time   Sat Feb 15 04:59:03 2014
--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=0.0007658 ...... PASSED
================================================================================
```

Test 02 – binaria HPL Intela (/usr/local/bin/xhpl_intel64_dynamic)

```
================================================================================
T/V                N    NB     P     Q               Time          Gflops
--------------------------------------------------------------------------------
WR01C2R4   1231776   224    52    53           20403.25     6.10668e+04
HPL_pdgesv() start time Sun Feb 16 02:11:08 2014
HPL_pdgesv() end time   Sun Feb 16 07:51:11 2014
--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=0.0007658 ...... PASSED
================================================================================
```

Test 03 – binaria HPL skompilowane lokalnie (/usr/local/bin/hpl_intel64_dynamic)

```
================================================================================
T/V                N    NB     P     Q               Time          Gflops
--------------------------------------------------------------------------------
WR01C2R4   1231776   224    52    53           20393.29      6.110e+04
HPL_pdgesv() start time Mon Feb 17 14:01:28 2014
HPL_pdgesv() end time   Mon Feb 17 19:41:21 2014
--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=0.0007658 ...... PASSED
================================================================================
```