Specyfikacja wymagań technicznych dla rozproszonego systemu składowania plików dla klastra obliczeniowego Galera

Definicje

Klaster obliczeniowy Galera - klaster obliczeniowy złożony z:

- 672 węzły w konfiguracji 2 x Xeon E5345, 8GB RAM, płyta główna Supermicro X7DBT-INF
- sieci InfiniBand: przełączniki Flextronics F-X430044 (24-portowe) oraz Flextronics F-X430081 (120-portowe),

System składowania danych (SSD) - sprzęt składowania danych (macierze dyskowe, kontrolery, serwery) oraz oprogramowanie udostępniające system plików klientom - węzłom klastra obliczeniowego.

Wymagania ogólne

System Składowania danych musi spełniać następujące wymagania:

- obsługa całego klastra obliczeniowego
- obsługa systemu quota
- wsparcie dla programów równoległych i standardu MPI-IO
- skalowalność z zachowaniem architektury systemu i możliwość rozbudowy do co najmniej 2 PB użytkowej przestrzeni dyskowej poprzez dołączenie dodatkowych serwerów danych, zasobów dyskowych i ewentualnie dodatkowych przełączników sieciowych.

System składowania danych musi dostarczać do klastra dwa zasoby danych:

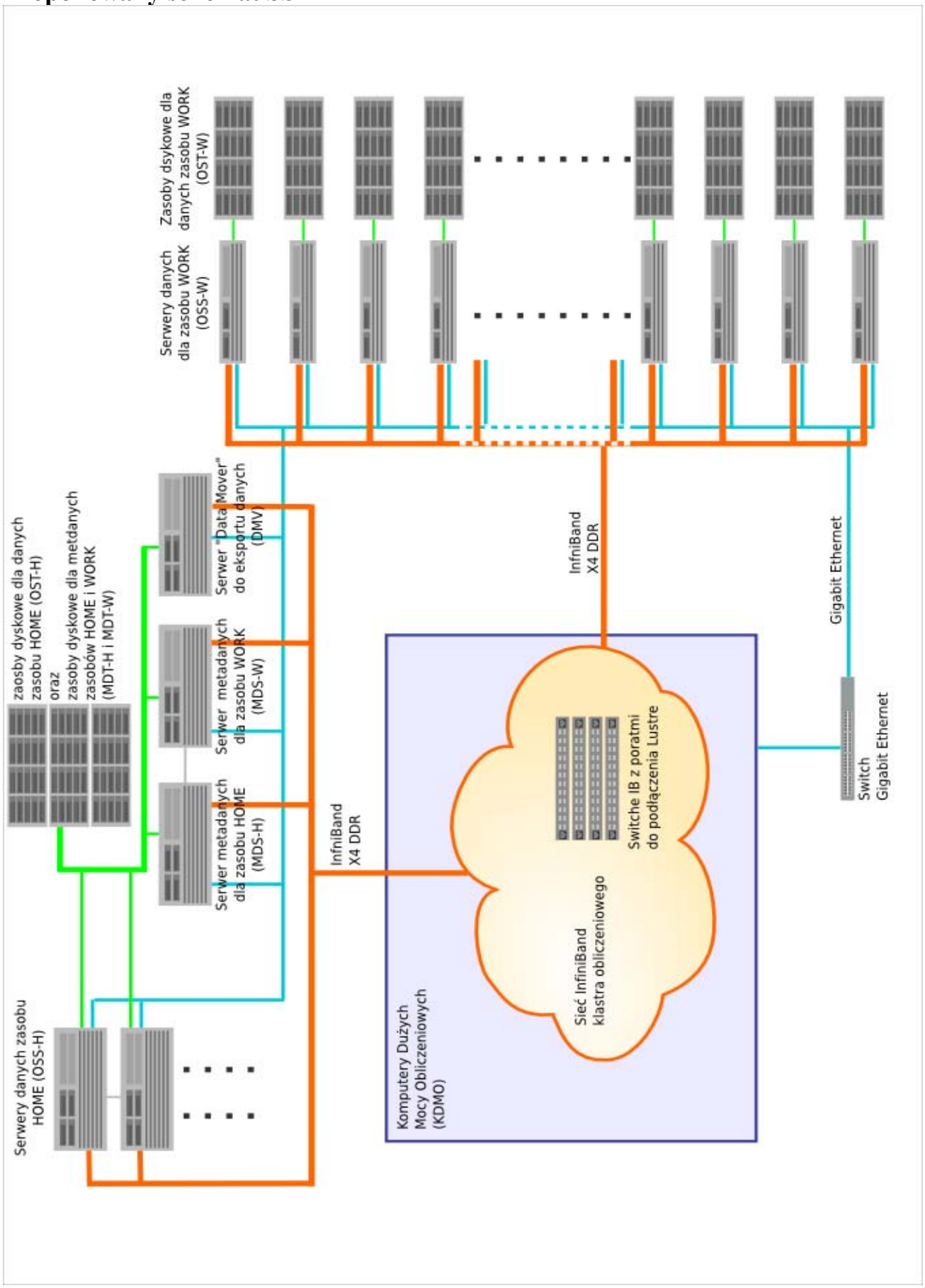
zasób "WORK" - o pojemności co najmniej 0.5 PB surowych dysków,

- zapewniający ciągły zapis danych ze wszystkich węzłów klastra w jednym czasie z prędkością sumaryczną co najmniej 10 GB/s,
- zabezpieczony na poziomie serwerów danych i serwerów metadanych mechanizmami typu RAID 5 z jednym dyskiem zapasowym na każdą grupę RAID.
- złożony z co najmniej 10 serwerów danych

zasób "HOME" - o pojemności co najmniej 28 TB surowych dysków

- zapewniający ciągły zapis danych z prędkością co najmniej 1.5 GB/s,
- w pełni zabezpieczony przed awarią każdego z elementów zasobu.
- złożony z co najmniej 2 serwerów danych

Proponowany schemat SSD



przyjęte oznaczenia:

- SSD - System Składowania Danych
- OSS-H - serwer danych dla zasobu HOME
- OSS-W - serwer danych dla zasobu WORK
- MDS-H - serwer metadanych dla zasobu HOME
- MDS-W - serwer metadanych dla zasobu WORK
- OST-H - zasoby dyskowe dla OSS-H
- OST-W - zasoby dyskowe dla OSS-W
- MDT-H - zasoby dyskowe dla MDS-H
- MDT-W - zasoby dyskowe dla MDS-W
- DMV - serwer (Data Mover) do przerzutu i eksportu zasobów HOME i WORK poza SSD

Specyfikacja sprzętu

serwery		
typ	opis/wymagane wyposażenie	wymagana ilość (minimum)
OSS-H	**serwer danych dla zasobu HOME** – minimum 4 procesory czterordzeniowe o architekturze x86-64 lub em64t – minimum 32GB pamięci RAM z kontrolą błędów ECC – sprzętowy kontroler SCSI, RAID 0,1,5 – co najmniej cztery dyski SCSI lub SAS na system operacyjny, o pojemności min. 140 GB każdy, prędkość obrotowa talerzy min. 10000 RPM, dołączone do kontrolera RAID w trybie mirror (2 dyski + spare) – odpowiednie przyłącze do zewnętrznych zasobów dyskowych – urządzenia i oprogramowanie pozwalające na przejęcie funkcji serwera przez serwer zapasowy w razie awarii serwera podstawowego – redundantne zasilacze N+1 gdzie N co najmniej 1 – co najmniej 4 porty Gigabit Ethernet o prędkości 10/100/1000 Mbit/s (medium: miedź, RJ45) – co najmniej 2 porty InfiniBand 4X DDR, 20 Gb/s, umożliwiające podpięcie węzła do przełącznika pracującego w technologii InfiniBand, udostępnione do połączeń z klastrem obliczeniowym, – zainstalowane i skonfigurowane oprogramowanie *LUSTRE*	2
OSS-W	**serwer danych dla zasobu WORK** – minimum 2 procesory czterordzeniowe o architekturze x86-64 lub em64t – minimum 8GB pamięci RAM z kontrolą błędów ECC – co najmniej 4 porty Gigabit Ethernet o prędkości 10/100/1000 Mbit/s (medium: miedź, RJ45) – co najmniej 2 porty InfiniBand 4X DDR, 20 Gb/s, umożliwiające podpięcie węzła do przełącznika pracującego w technologii InfiniBand, udostępnione do połączeń z klastrem obliczeniowym, – zainstalowane i skonfigurowane oprogramowanie *LUSTRE*	12

MDS-H	**serwer metadanych dla zasobu HOME** – minimum 4 procesory czterordzeniowe o architekturze x86-64 lub em64t – minimum 32GB pamięci RAM z kontrolą błędów ECC – sprzętowy kontroler SCSI, RAID 0,1,5 – co najmniej cztery dyski SCSI lub SAS na system operacyjny, o pojemności min. 140 GB każdy, prędkość obrotowa talerzy min. 10000 RPM, dołączone do kontrolera RAID w trybie mirror (2 dyski + spare) – odpowiednie przyłącze do zewnętrznych zasobów dyskowych – urządzenia i oprogramowanie pozwalające na przejęcie funkcji serwera przez serwer zapasowy w razie awarii serwera podstawowego – redundantne zasilacze N+1 gdzie N co najmniej 1 – co najmniej 4 porty Gigabit Ethernet o prędkości 10/100/1000 Mbit/s (medium: miedź, RJ45) – co najmniej 2 porty InfiniBand 4X DDR, 20 Gb/s, umożliwiające podpięcie węzła do przełącznika pracującego w technologii InfiniBand, udostępnione do połączeń z klastrem obliczeniowym, – zainstalowane i skonfigurowane oprogramowanie *LUSTRE*	1
MDS-W	**serwer metadanych dla zasobu WORK** – minimum 4 procesory czterordzeniowe o architekturze x86-64 lub em64t – minimum 32GB pamięci RAM z kontrolą błędów ECC – sprzętowy kontroler SCSI, RAID 0,1,5 – co najmniej cztery dyski SCSI lub SAS na system operacyjny, o pojemności min. 140 GB każdy, prędkość obrotowa talerzy min. 10000 RPM, dołączone do kontrolera RAID w trybie mirror (2 dyski + spare) – odpowiednie przyłącze do zewnętrznych zasobów dyskowych – urządzenia i oprogramowanie pozwalające na przejęcie funkcji serwera przez serwer zapasowy w razie awarii serwera podstawowego – redundantne zasilacze N+1 gdzie N co najmniej 1 – co najmniej 4 porty Gigabit Ethernet o prędkości 10/100/1000 Mbit/s (medium: miedź, RJ45) – co najmniej 2 porty InfiniBand 4X DDR, 20 Gb/s, umożliwiający podpięcie węzła do przełącznika pracującego w technologii InfiniBand, udostępnione do połączeń z klastrem obliczeniowym, – zainstalowane i skonfigurowane oprogramowanie *LUSTRE*	1
DMV	**serwer metadanych dla zasobu HOME** – minimum 4 procesory czterordzeniowe o architekturze x86-64 lub em64t – minimum 32GB pamięci RAM z kontrolą błędów ECC – sprzętowy kontroler SCSI, RAID 0,1,5 – co najmniej cztery dyski SCSI lub SAS na system operacyjny, o pojemności min. 140 GB każdy, prędkość obrotowa talerzy min. 10000 RPM, dołączone do kontrolera RAID w trybie mirror (2 dyski + spare) – odpowiednie przyłącze do zewnętrznych zasobów dyskowych – urządzenia i oprogramowanie pozwalające na przejęcie funkcji serwera przez serwer zapasowy w razie awarii serwera podstawowego – redundantne zasilacze N+1 gdzie N co najmniej 1 – co najmniej 4 porty Gigabit Ethernet o prędkości 10/100/1000 Mbit/s (medium: miedź, RJ45) – co najmniej 2 porty InfiniBand 4X DDR, 20 Gb/s, umożliwiający podpięcie węzła do przełącznika pracującego w technologii InfiniBand, udostępnione do połączeń z klastrem obliczeniowym, – zainstalowane i skonfigurowane oprogramowanie *LUSTRE*	

macierze dyskowe		
typ	opis/wymagane wyposażenie	wymagana ilość (minimum)
OST-H MDT-H MDT-W	**macierz dyskowa dla metadanych i danych zasobu „HOME”** – łączna surowa pojemność min. 40 TB – co najmniej 12 TB w dyskach 15000 obr./min. dla serwerów metadanych: MDS-W i MDS-H – co najmniej 28 TB w dyskach 10000 obr./min. dla serwerów danych OSS-H – obsługa RAID 0,1,5,6 – redundantny kontroler RAID – redundantne zasilacze N+1 gdzie N co najmniej 1 – przyłącze do serwerów MDS-W, MDS-H, DMV oraz OSS-H – możliwość podłączenia przez nią innych macierzy (tzw. wirtualizacja) – możliwość tworzenia kopii migawkowych (tzw. snapshot) wewnętrznymi mechanizmami macierzy – możliwość partycjonowania (wydzielania) zasobów – minimum 64GB cache – minimum 10 000 operacji IO/s	1
OST-W	**macierz dyskowa dla danych zasobu „WORK”** zespół dysków o parametrach: – łączna surowa pojemność min. 40 TB – dyski twarde o pojemność min. 700 GB, z przyłączem SATA lub SAS, prędkość obrotowa minimum 7200 obr./min. – redundantne zasilanie N+1 gdzie N co najmniej 1 Zamawiający dopuszcza rozwiązanie, w którym serwer OSS-W i macierz OST-W stanowią jedno urządzenie (*wspólna obudowa i zasilanie*)	12

Przełączniki		
typ	opis/wymagane wyposażenie	wymagana ilość (minimum)
Gigabit Ethernet	**Przełącznik sieciowy sieci zarządzającej** – 48-portowy przełącznik Gigabit Ethernet Cisco Catalyst 2960 (*typ przełącznika wymuszony koniecznością zapewnienia zgodności z siecią zarządzającą klastrem obliczeniowego Galera i zasobami CI TASK)*	1

Wymagania dotyczące architektury SSD

- wszystkie serwery OSS-H, MDS-H, MDS-W i DMV muszą być sprzętowo identyczne
- wszystkie serwery OSS-W, OSS-H muszą mieć tą samą architekturę procesora
- wszystkie serwery OSS-W muszą być sprzętowo identyczne
- serwer MDS-H ma być serwerem zapasowym dla MDS-W
- serwer MDS-W ma być serwerem zapasowym dla MDS-H
- serwery OSS-H mają być wzajemnie serwerami zapasowymi

- serwery zapasowe muszą automatycznie przejmować obowiązki serwera podstawowego, np. za pomocą mechanizmu *HEARTBEAT* z wykorzystaniem portu RS-232

Ogólne wymagania dotyczące wszystkich serwerów

- obudowa umożliwiająca montowanie w szafie typu rack 19 z systemem wentylacji zapewniającym pobór powietrza z przodu obudowy (szafy) i wydmuchiwaniem do tyłu
- maksymalna wysokość: 4U
- złącze RS-232
- złącze USB 2.0
- złącze VGA
- 1 wolny slot umożliwiający zainstalowanie karty rozszerzeń w standardzie PCI-Express x8,
- możliwość bootowania systemu operacyjnego przez sieć (DHCP + TFTP),
- możliwość bootowanie systemu operacyjnego z urządzenia USB typu *pendrive* lub karty CF,
- pełne wsparcie systemu Linux 64-bit, jądro minimum 2.6, dla węzła i zainstalowanych w nim kart rozszerzeń,
- możliwość zdalnego załączania i wyłączania serwera oraz monitorowanie parametrów pracy (temperatura, napięcia, prędkości wentylatorów),
- możliwość dostępu do konsoli szeregowej przez sieć Ethernet (Serial-over-LAN)
- zestaw do zamontowania serwera na wysuwanych szynach w szafie teleinformatycznej 19”

Wymaganie dotyczące oprogramowania

Zainstalowane oprogramowanie

- musi być w pełni zgodne z systemem Lustre 1.6 lub Lustre 1.8
- działające na elementach systemu składowania
- posiadać klienta działającego na wszystkich węzłach klastra obliczeniowego (system Debian GNU/Linux etch, jądro Linux 2.6)

Okablowanie, szafy, zasilanie, organizacja okablowania

Szafy

Całość dostarczanego sprzętu musi zostać zainstalowana w nie więcej niż 4 szafach (*jeżeli urządzenie nie stanowi odrębnej szafy wolnostojącej to powinno być zainstalowane w szafie teletechnicznej typu rack o szerokości do 80 cm, głębokości nie większej niż 100cm i wysokości nie większej niż 210cm*), o łącznej pojemności odpowiedniej dla całości sprzętu objętego dostawą (serwery, macierze, akcesoria szaf, itd.), **z zachowaniem wymogu maksymalnego poboru mocy 8 kW na szafę**

Wymagania dotyczące wyposażenia szafy typu rack:

- drzwi przednie i tylne blaszane perforowane
- wysuwany cokół zabezpieczający przed przewróceniem się szafy

- wymagane zdejmowane drzwi przednie i tylne
- dach pełny
- perforacja drzwi musi być wykonana na całości powierzchni (oprócz ramy konstrukcyjnej) przy zachowaniu jak największych otworów dla maksymalnej cyrkulacji powietrza
- wymagana możliwość demontażu szafy na czas transportu
- wymagana możliwość trwałego złączenia szaf bokami w celu zapewnienia stabilności i estetyki zespołu szaf
- osłony boczne pełne zdejmowane

Zasilanie

- Cała instalacja może pobierać co najwyżej 30 KW mocy, przy czym w pojedynczej szafie pobór nie może przekroczyć 8 KW
- w każdej szafie musi być zapewniona odpowiednia liczba paneli zasilających, przyłączanych linią trójfazową, umożliwiających:
  - zdalne monitorowanie parametrów zasilania i zdalne sterowanie załączaniem i wyłączaniem urządzeń,
  - zarządzanie przez Ethernet
  - obsługę protokółu SNMP, HTTP
- jeżeli szafa nie posiada odpowiedniego wbudowanego panelu zasilającego, należy użyć panelu pełni zgodnego z używanymi z Klastrze Obliczeniowym panelami APC AP7957
- Wykonawca ustali z Zamawiający co najmniej na dwa tygodnie przed planowaną dostawą sposób przyłączenia dostarczanych urządzeń do sieci energetycznej Zamawiającego

Okablowanie

- odpowiednia ilość patchcordów cat5e lub lepszych z wtykami RJ45 (montowane fabrycznie, długość i kolory do ustalenia przed dostawą) umożliwiająca połączenie serwerów z przełącznikiem Gigabit Ethernet,
- odpowiednia ilość opasek na rzepy do spinania kabli,
- odpowiednia ilość trwałych etykiet z tworzywa sztucznego do oznaczenia kabli sieciowych ehternetowych na obu końcach, zawierające: numer szafy, pozycję węzła w szafie,
- odpowiednia ilość kabli w standardzie InifiBand 4X, optycznych, o długości co najmniej 30 m umożliwiających połączenie serwerów z przełącznikami InfiniBand
- korytka podtrzymujące kable prowadzone pomiędzy szafami i przełącznikami w obrębie szaf oraz do szaf klastra obliczeniowego według potrzeb,
- odpowiednia ilość zaślepień do szaf 19", w razie nie wykorzystania całej przestrzeni w szafach

Zamawiający zapewnia

- 42 porty InfiniBand 4xDDR umożliwiające połączenie z klastrem obliczeniowym
- jeden port Gigabit Ethernet umożliwiający połączenie do sieci zarządzającej klastra
- zasilanie 3-fazowe z dwóch niezależnych źródeł

Szczegółowy zakres prac będących przedmiotem zamówienia

- rozładunek urządzeń,
- fizyczny montaż urządzeń w miejscu wskazanym przez Zamawiającego,
- wykonanie niezbędnych połączeń elektrycznych i logicznych w uzgodnieniu z Zamawiającym. Zamawiający wymaga określenia, z co najmniej dwutygodniowym wyprzedzeniem przed planowanym dniem dostawy, typu przyłącza elektrycznego szafy rack, oraz wymagań co do prowadzenia okablowania pomiędzy szafami,
- dokładne i jednoznaczne oznakowanie wszystkich połączeń
- instalacja i konfiguracja na każdym z dostarczonych serwerów systemu operacyjnego Linux. Konfiguracja systemu obejmuje w swym zakresie wszelkie czynności niezbędne do uruchomienia oprogramowania wymienionego dalej,
- instalacja konfiguracja i strojenie systemu składowania danych,
- instalacja i konfiguracja oprogramowania do zarządzania infrastrukturą serwerów,
- instalacja, uruchomienie i przeprowadzenie testu wydajności systemu plików
- przeszkolenie pracowników Zamawiającego,
- opracowanie i dostarczenie dokumentacji powykonawczej,

Szkolenie

Zamawiający wymaga przeszkolenia minimum 3 pracowników w zakresie administracji dostarczonym systemem na poziomie podstawowym, w lokalizacji ustalonej z Wykonawcą, Miejsce, termin oraz szczegółowa tematyka szkolenia zostaną ustalone przez Strony w terminie do 4 tygodni od daty podpisania protokołu odbioru sprzętu,

Dokumentacja powykonawcza

Zamawiający wymaga dostarczenia po zakończeniu całości projektu dokumentacji powykonawczej systemu zawierającej co najmniej:

- dokumentację techniczną dostarczonych urządzeń,
- szczegółowy opis konfiguracji dostarczonego oprogramowania oraz urządzeń,
- procedury eksploatacyjne,
- procedury awaryjne.