Sieć dla superkomputera

Aleksanda Hein, Wiktor Kawka, Błażej Smorawski

11.05.2023

Założenia

Dla tego komputera powstanie nowy budynek o powierzchni około 800m2. Założono, że super komputer będzie miał 2048 węzłów obliczeniowych, które są pogrupowane w 16 grup. Oprócz tego, założono iż będą 4 podwieszane kanały. Istotne jest, aby było jak najmniejsze opóźnienie w przesyłaniu danych między węzłami oraz jak największa przepustowość.

Dodatkowym założeniem jest brak rozbudowy klastra - klaster jest budowany pod kątem maksymalnej wydajności bez planów jakichkolwiek zmian architektury. Lokalność węzłów obliczeniowych jest krytyczna, dlatego nie przewidziano możliwości podłączenia kolejnych szaf, kolejnych grup czy przełączników. Zaplanowany układ oferuje najwydajniejszą komunikację i nie ma możliwości dodania jakiegokolwiek węzła bez zaburzenia symetrii komunikacji między resztą.

Plan budynku

Pomieszczenia

1. Pomieszczenie dla węzłów obliczeniowych
2. Pomieszczenie dla ochrony
3. Magazyn
4. Schowek dla osób dbających o porządek

Sieci

Sieć Obliczeniowa

Celem sieci obliczeniowej jest dostarczenie szybkiego połączenia pomiędzy wszystkimi węzami obliczeniowymi klastra. Podstawowym medium komunikacyjnym jest **InfiniBand** wykorzystując złącze **QSFP56** z prędkością **HDR**. Każda grupa będzie wyposażona w pięć przełączników *NVIDIA MQM8790-HS2R Quantum HDR InfiniBand Switch* w rozmiarze 1U wyposażone w 40 portów wyjściowych.

Urządzenia zostały wybrane ze względu na wykorzystanie w węzłach obliczeniowych akceleratrów firmy NVidia, co umożliwia ścisłą współpracę pomiędzy nimi i wykorzystanie protokołu **SHARP** służącego do wykonywania obliczeń **‘in network‘**. Przykładem jest operacja redukcji buforów między urządeniami, która może zostać wykonana na pierwszym napotkanym przełączniku, co znacznie zmniejsza obciążenie sieci oraz zwalnia zasoby węzłów obliczeniowych.

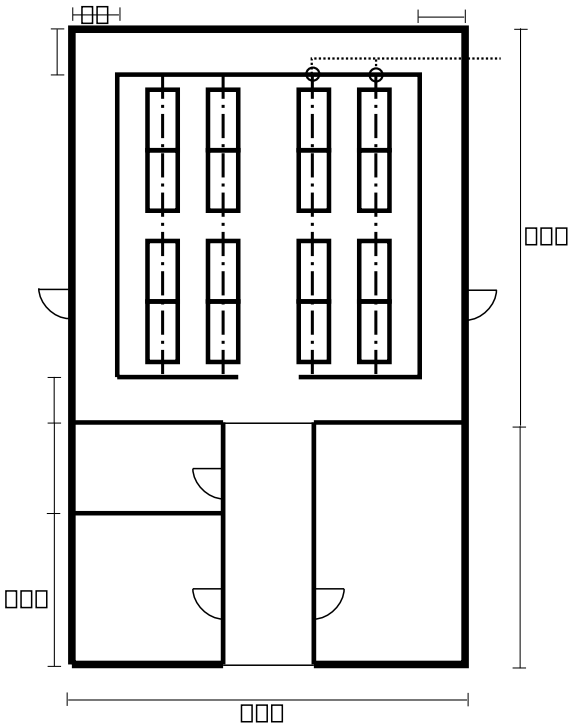

Figure 1: Uproszczony plan budynku

Przełączniki dodatkowo gwarantują nam nieprzerwaną dostępność i możliwość realizowania napraw w trakcie pracy urządzenia:

- 1+1 zasilaczez z możliwością **‘hot plug‘**
- N+1 redundantne wentylatory z możliwością **‘hot plug‘**
- certyfikowane zasilacze **80 gold+**

Cztery przełączniki w grupie 128 węzłów są podłączone do 32 węzłów obliczeniowych każdy, a pozostałe 8 portów jest wykorzystanych do połączenia z pozostałymi przełącznikami w grupie dwoma zaagregowanymi łączami. Piąty przełącznik to przełącznik wyjściowy grupy, którego 8 portów łączy się z przełącznikami grupy, a 30 pozostałych przez łącza optyczne z 15 innymi grupami w gęstym połączeniu każdy z każdym.

Wybrane zostały przewody firmy NVidia ze względu na:

- Zgodność z **InfiniBand HDR**
- Przepływność **200 Gb/s**
- BER lepsze niż **1e-15**

Potrzebny sprzęt:

- Przełącznik *NVIDIA MQM8790-HS2R Quantum HDR InfiniBand Switch* x 16 x 5 = **80**
- Przewody miedziane *NVIDIA MCP1650-H00AE30 DAC 1m* x 2048 + 16 * 4 * 4 * 2 = **2560**
- Przewody optyczne *NVIDIA MFS1S00-H015V AOC 15m* x 16 x 15 = **240**

Sieć kontrolna

Sieć kontrolna jest siecią odseparowaną fizycznie od sieci, która pozwala nam na kontrolowanie pracy klastra bez wpływu na jego parametry obliczeniowe. Ze względu na mniejsze wymagania wydajnościowe będzie to sieć **Ethernetowa**.

Wybraliśmy przełączniki firmy *NVidia* ze względu na bardzo dobre zdolności telemetryczne, niskie opóźnienia i złożone mechanizmy zarządania ruchem.

Każda grupa będzie wyposażona w pięć 64. portowych ethernetowych przełączników *NVIDIA MSN4600-CS2F Spectrum-3 100GbE 2U Open Ethernet Switch*. Tak jak wcześniejsze przełączniki są wyposażone w 2 zasilacze i N+1 wymienialnych wentylatorów.

Wybrane przewody to *NVIDIA MCP2M00-A005E26L DAC 5m* oraz *NVIDIA MCP1650-V001E30 DAC 1M*.

Potrzebny sprzęt:

- Przełącznik *NVIDIA MSN4600-CS2F Spectrum-3 100GbE 2U Open Ethernet Switch* x 16 x 4 = **64**
- Przewody miedziane *NVIDIA MCP1650-V001E30 DAC 1m* x **2560**

- Przewody miedziane *NVIDIA MCP2M00-A005E26L DAC 5m* x 16 x 15 = **480**

Sieć administracji sprzętu sieciowego

Sieć łącząca wszsytkie przełączniki i routery z powyższych sieci w celach konfiguracyjnych. Wykorzysuje przewody miedziane i złącza RJ-45. Ze względu na niskie wymagania będzie oparta o przełączniki *Cisco CBS350-48T-4X-EU Managed 48-port GE* x 16.

Bezpieczeństwa

Do zadań realizowanych przez sieć bezpieczeństwa należą: - monitorowanie budynku za pomocą kamer - kontrola dostępu do budynku, jak i jego poszczególnych pomieszczeń za pomocą kart - monitorowanie i regulacja temperatury w pomieszczeniu, w którym znajduje się superkomputer - monitorowanie i powiadamianie o włamaniach do budynku - wykrywanie i powiadamianie o pożarze w budynku - wyświetlanie obrazu z kamer i danych z czujników na komputerach pracowników

Potrzebne urządzenia: - ok. 25 kamer do monitoringu, - ok. 7 czytników NFC oraz kart dla każdego pracownika budynku, - ok. 20 000 czujników temperatury, - ok. 8 urządzeń chłodzących i 8 urządzeń pobierających ciepło, - ok. 40 czujników antywłamaniowych, - ok. 20 czujników dymu.

Wybranym medium komunikacyjnym jest kabel Ethernet kategorii 5E. Dokonano takiego wyboru, gdyż medium to zapewnia wystarczającą przepustowość oraz zasięg.

Potrzebny sprzęt: - przewody miedziane Ethernet kat.5E -> 3800m - 6 przełączników *Cisco CBS350-48T-4X-EU Managed 48-port GE* - 6 podwójnych gniazdek Ethernet *Alantec 2xRJ45 IP54*

Dostępowa

Celem sieci dostępowej jest zapewnienie dostępu z sieci internet do dwóch węzłów obliczeniowych pełniących rolę węzłów dostępowych. Firma telekomunikacyjna dostarcza dostarcza nam 8 łącz ethernetowych **200Gb**. Routery są bezpośrednio połączone dwoma optycznymi przewodami *NVIDIA MFS1S00-H015V AOC 15m* do każdego węzła dostępowego. Taka instalacja pozwala nam na osiągnięcie łącza rzędu **400Gb/s** do jednego węzła dostępowego, co może być konieczne ze względu na wielu użytkowników oraz dużą zajętość danych przetwarzanych w takim klastrze.

Router został wybrany ze względu na:

- Dużą moc obliczeniową - dwa procesory *Intel Xeon scalable processor Platinum*
- 1+1 redundantne zasilacze klasy **80 Plus platinum**

Potrzebny sprzęt:

- Router *NVIDIA MGA100-HS2 Skyway* x 2
- Przewody optyczne *NVIDIA MFS1S00-H015V AOC 15m* x 4

Plan adresacji

Sieć obliczeniowa

Adresy w postaci 10.10.grupa.węzeł/16, gdzie węzły są pogrupowane według przynależności do szaf kolejno od 1 do 128.

Sieć Kontrolna

Adresy w postaci 40.40.grupa.węzeł/16, gdzie węzły są pogrupowane według przynależności do szaf kolejno od 1 do 128.

Sieć administracyjna

Adresy w postaci 80.80.grupa.przełącznik/16, gdzie przełączniki są pogrupowane według przynależności do szaf kolejno od 1 do 8. Adres 80.80.grupa.255/16 to adres przełącznika administracyjnego. Adresy routerów to 80.80.255.[254,255]/16.

Sieć bezpieczeństwa

Sieć	Podsieć	Adres
sieć bezpieczeństwa	monitoring	192.168.0.0/27
192.168.0.0/24	temperatura	192.168.0.32/27
	pożar	192.168.0.64/27
	dostęp i włamania	192.168.0.128/26

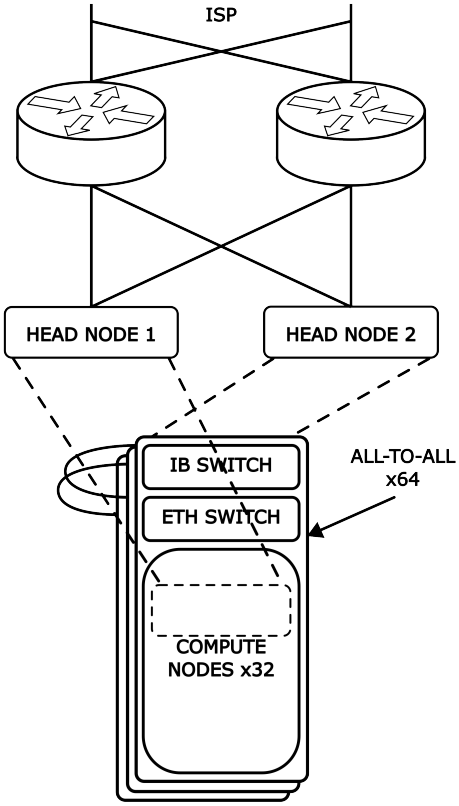

Figure 2: Projekt sieci logicznej

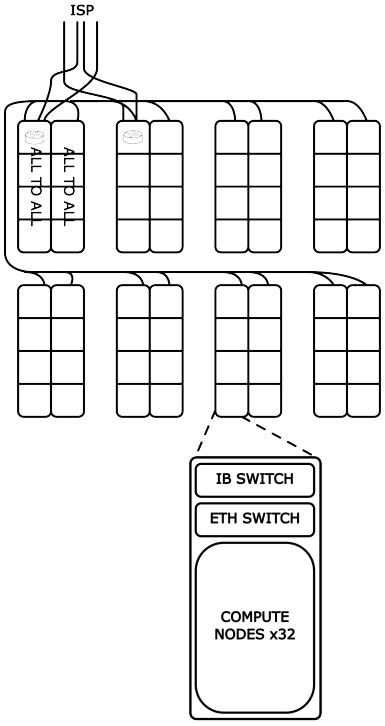

Figure 3: Projekt sieci fizycznej

Niezawodność

Niezawodność została osiągnięta przez:

- Redundancję na poziomie:
  - Łącza dostarczane przez dostawcę usług sieciowych
  - Routerów - 2
  - Przełączników - każdy węzęł podłączony do dwóch
  - Sieci - dwie niezależne sieci podłączone do każdego węzła
  - Interferjsów sieciowych - minimum dwa interfejsy sieciowe w węźle
  - Zasilacze i wentylatory wewnątrz sprzętu sieciowego.
- Zastosowanie różnych standardów połączeń - *Infiniband* + *Ethernet*
- Sprzęt sieciowy pozwalający na naprawy w trakcie pracy
- Przewody sieciowe o bardzo niskiej stopie błędów

Skalowalność

Sieć nie została zaprojektowana pod kątem skalowania w przyszłości.

Kosztorys

	Liczba	Cena	Suma
NVIDIA MQM8790-HS2R Quantum HDR InfiniBand Switch	80	$34,073.00	$2,725,840.00
NVIDIA MSN4600-CS2F Spectrum-3 100GbE 2U Open Ethernet Switch	80	$65,159.00	$5,212,720.00
NVIDIA MCP1650-H00AE30 DAC 1m	2560	$351.00	$898,560.00
NVIDIA MFS1S00-H015V AOC 15m	244	$2,282.00	$556,808.00
NVIDIA MCP2M00-A005E26L DAC 5m	240	$222.00	$53,280.00
NVIDIA MCP1650-V001E30 DAC 1m	2560	$314.00	$803,840.00
*NVIDIA MGA100-HS2 Skyway**	2	$55,795.00*	$111,590.00
Cisco WS-C3850-48XS-E	9	$64,156.00	$577,404.00

	Liczba	Cena	Suma
Licencja do przełącznika	9	$15,877.00	$142,893.00
Molex UTP Cat 5e LSOH	5500m	$0.50	$2750.00
Alantec 2xRJ45 IP54 wall socket	6	$130.00	$780.00
Raspbery Pi 3B+	32	$110.00	$3520.00

* Cena routera przed przejęciem firmy *Mellanox* przez *NVidia*

Suma: $10,362,638.00 + $727,347.00