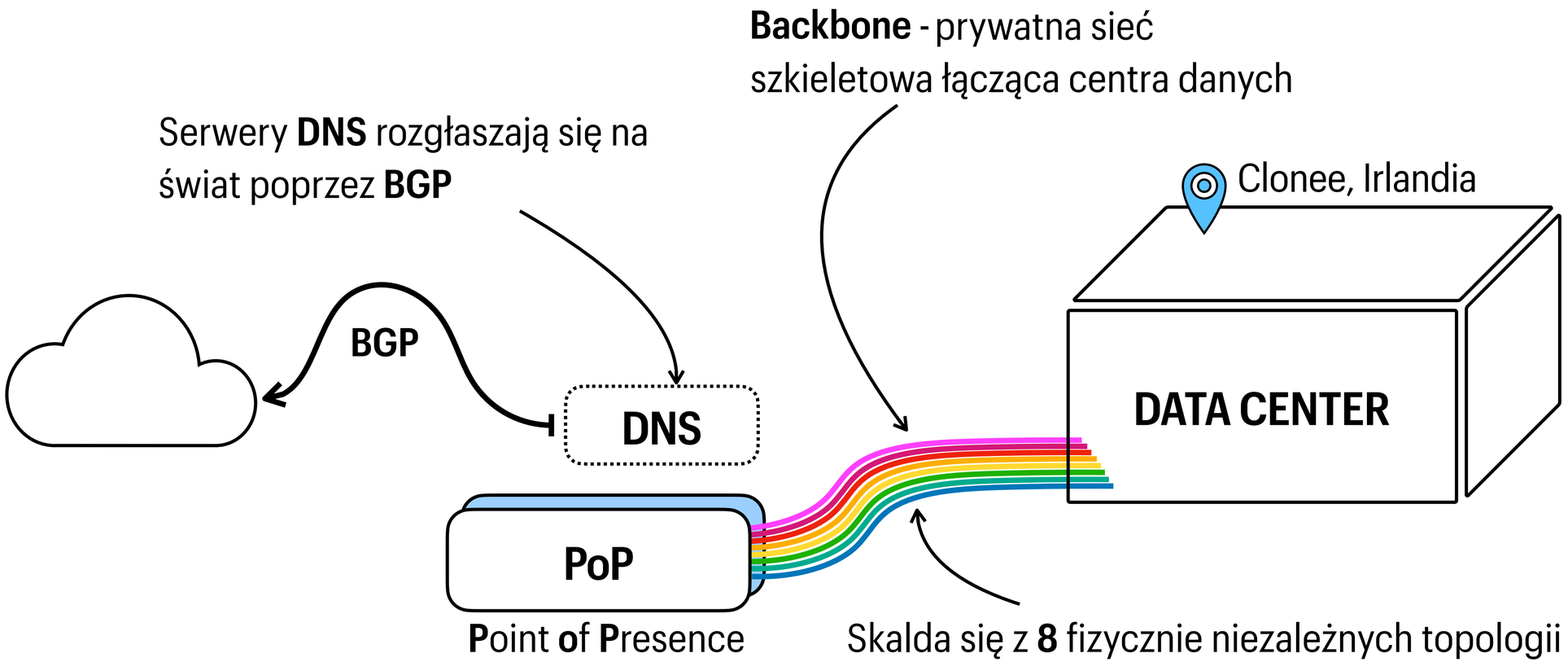
Centra Danych

Za mało i za szybko o bardzo ważnym i coraz ważniejszym

NIKITA MURAVYTSKYI AGH

COOL STORY, BOB

META 2021 OUTAGE

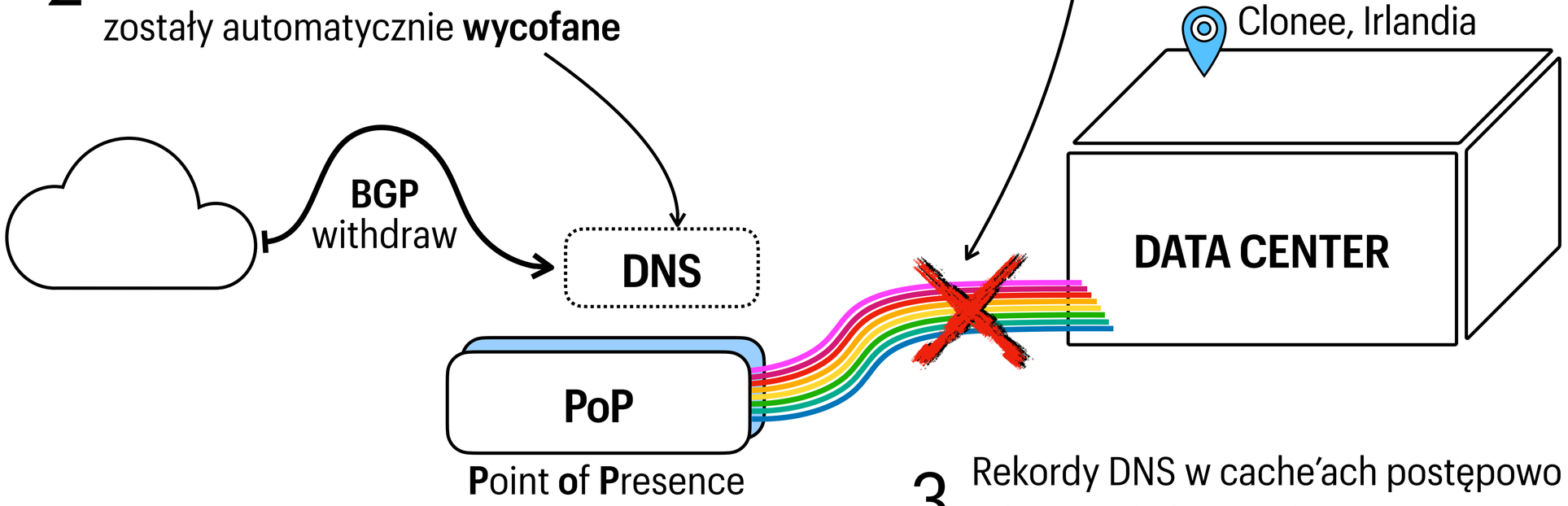


COOL STORY, BOB

META 2021 OUTAGE

1 Pomyłkowe wprowadzenie stanu **maintenance** wszystkich plane'ów na raz sprawiło, że sieć szkieletowa przestała działać

2 Wykrywszy niedostępność DC, wpisy **BGP** rozgłaszające serwery DNS zostały automatycznie **wycofane**

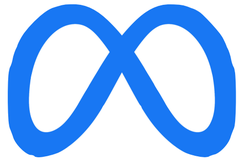


3 Rekordy DNS w cache'ach postępowo znikały, odcinając kolejne usługi - również **wewnętrzne**

KILKA SŁÓW O SOBIE

JESTEM ...

Absolwentem AGH u.



Zajmowałem się automatyzacją DC w Meta,

testowałem sieci razem z NeT RESEARCH

i obecnie nadzoruję produkcję w optiver.

AGENDA

TEMATY KTÓRE NIEKONIECZNIE ZOSTANĄ OMÓWIONE

O samopodobieństwie i niewyczerpalnym pragnieniu skalowalności

O krzemie Broadcom’a, kartach graficznych Nvidii, SmartNIC’ach, Data Processing Unit i 800G

O tym jak topologia Clos’a przekoczowała z centrali telefonicznej do centrum danych

O highwayach współczesnego internetu

O ultima thule współczesnych DC

O robocie śmieciarce WALL-E

O tym że chłodzenie to bardziej o wodę niż o powietrze

O tym jak nie zostawiać kluczyków w aucie

DROGOWSKAZY

JAK SIĘ NAWIGOWAĆ PO TEJ PREZENTACJI?

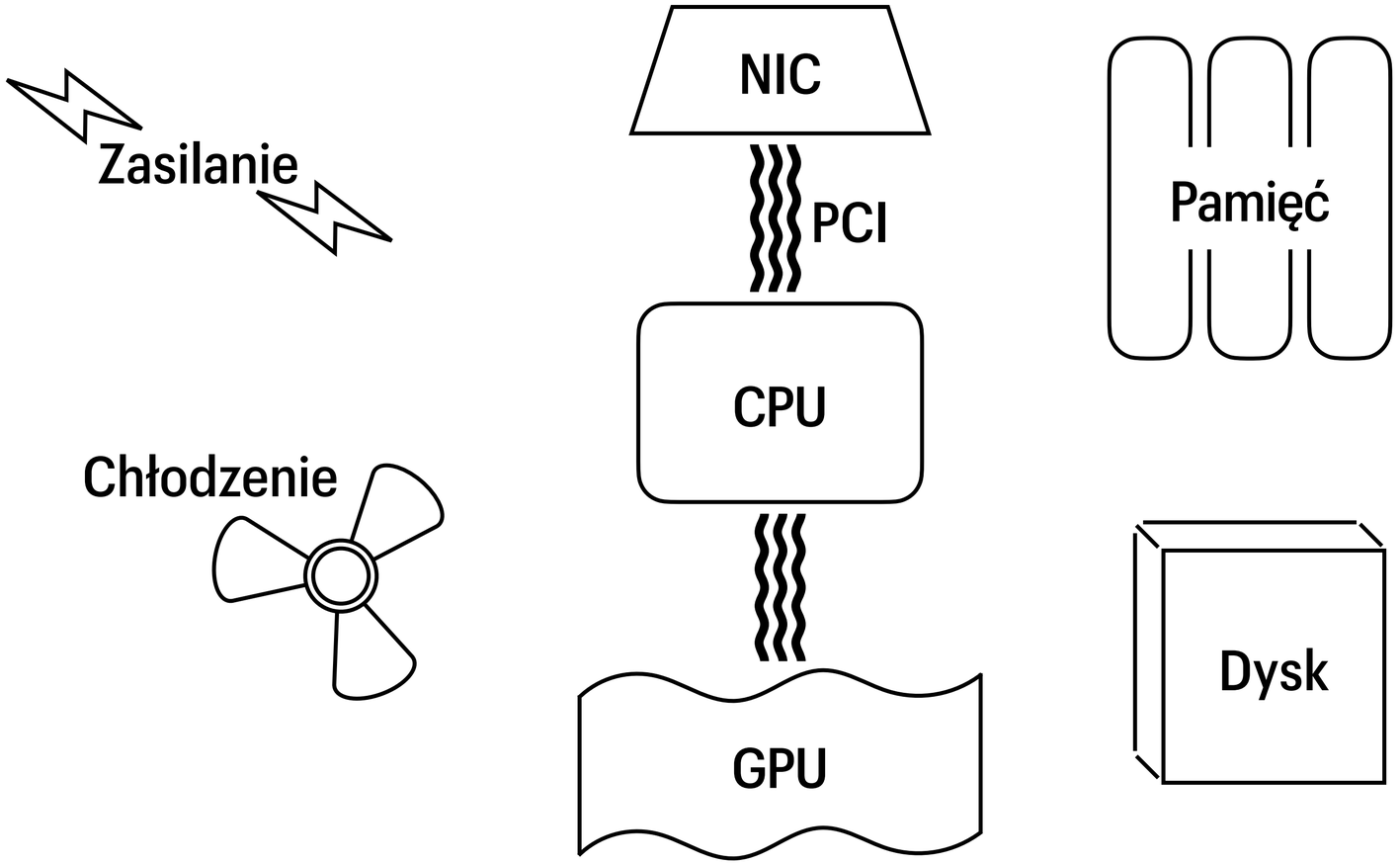
Dana prezentacja jest jedynie przeglądem niektórych zagadnień z tematyki DC.
Na slajdach pozostawione są wskazówki dla samodzielnego studiowania tematu:

 — polecam poczytać, obejrzeć albo posłuchać dla pogłębienia wiedzy;

 — intrygujące pytanie na które nie udzielę odpowiedzi, musicie poszukać sami;

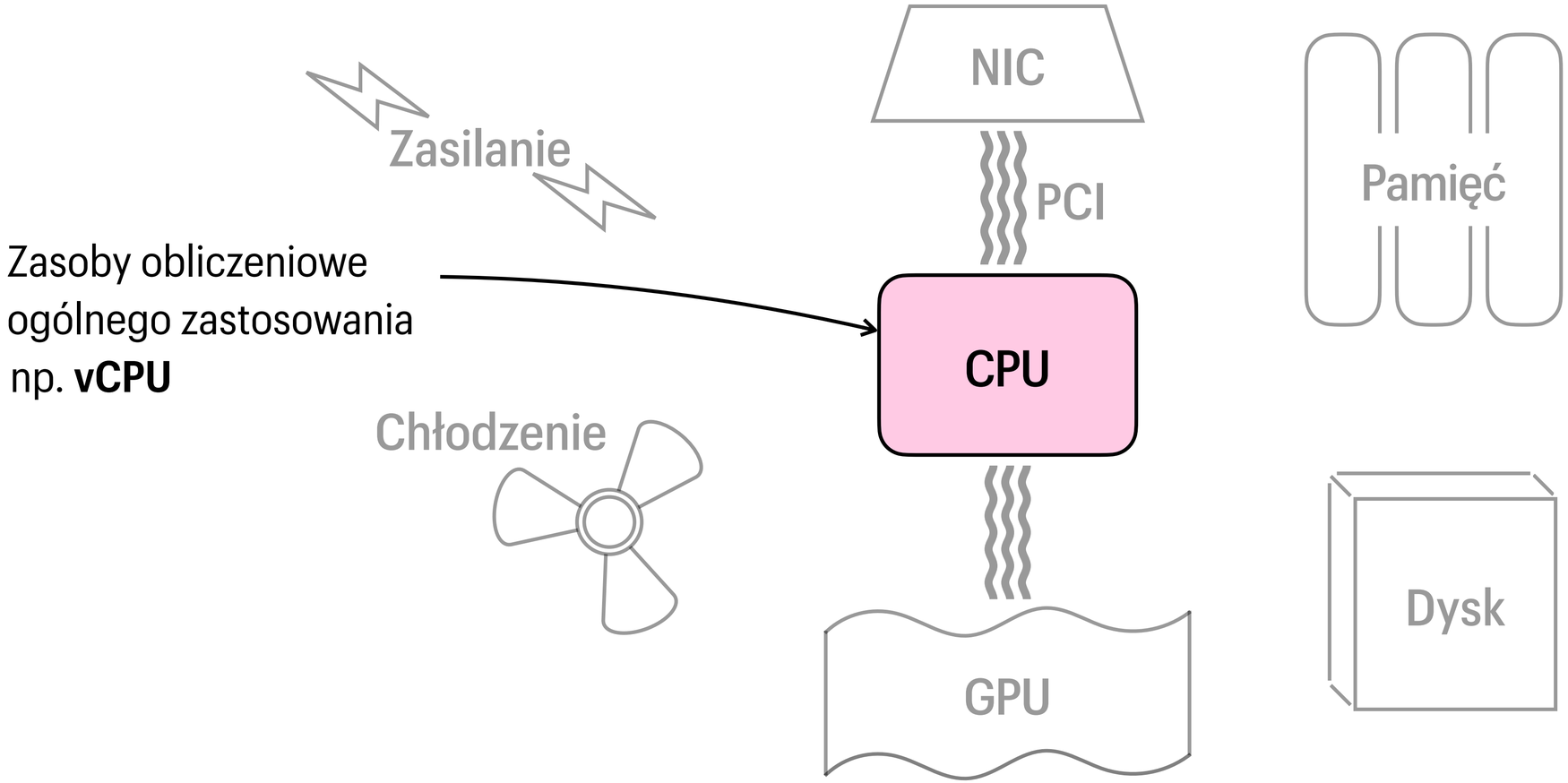
Większość przykładów stanowią rozwiązania Meta'y. Rozwiązania innych firm wykorzystują podobne koncepcje albo nie mają informacji dostępnej publicznie.

IDEALNE CENTRUM DANYCH JEST JAK KOMPUTER



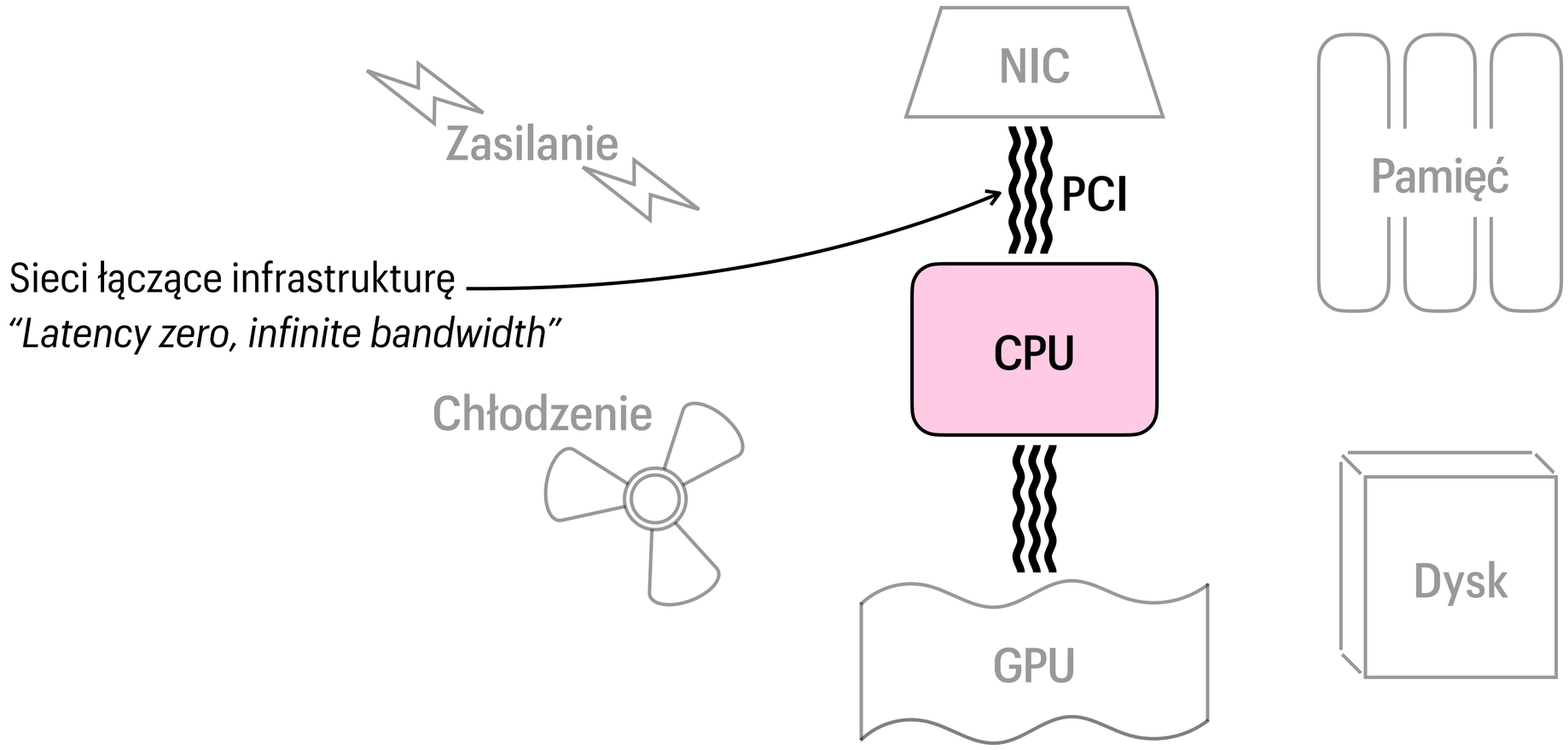
O SAMOPODOBIEŃSTWIE

IDEALNE CENTRUM DANYCH JEST JAK KOMPUTER



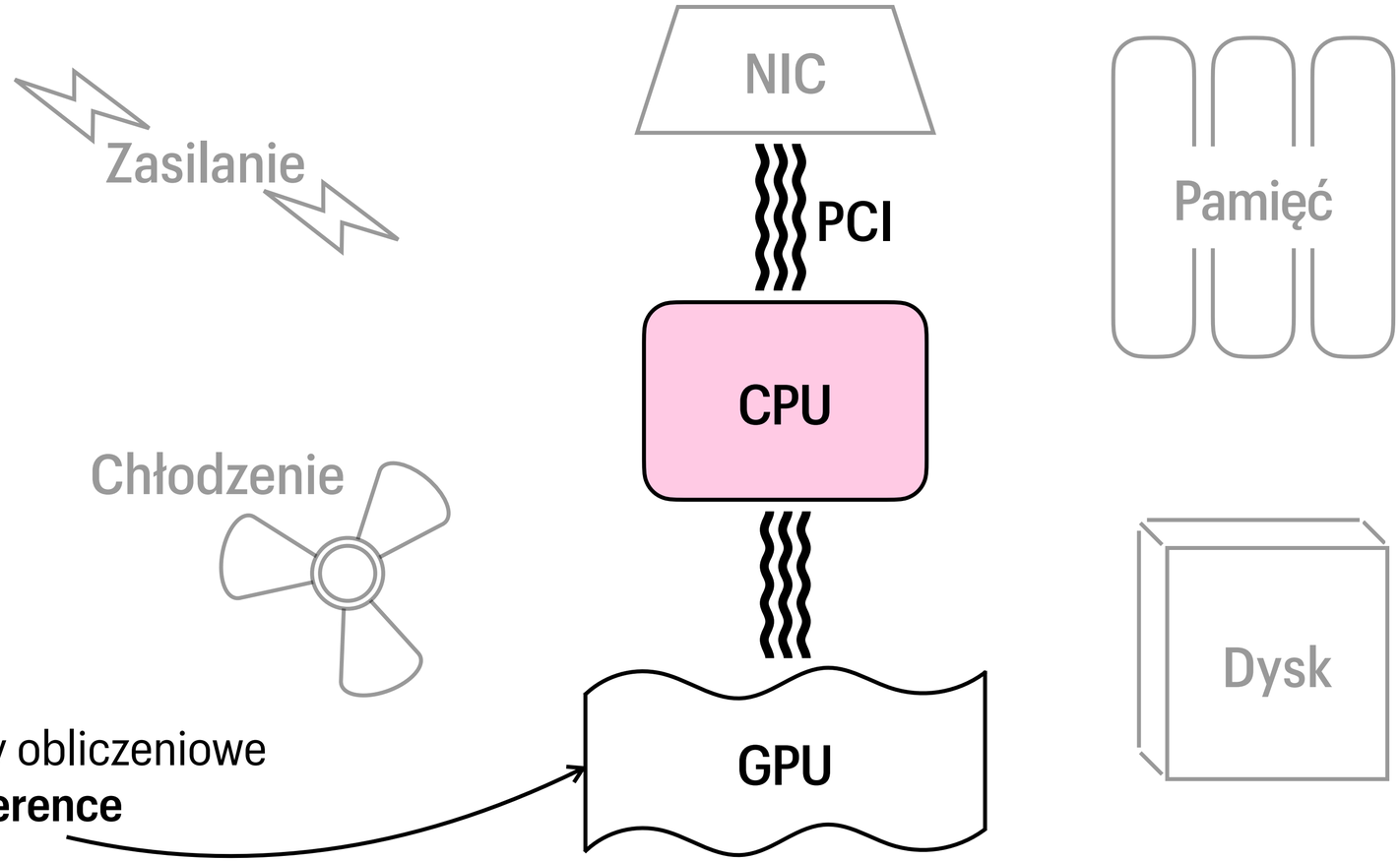
O SAMOPODOBIEŃSTWIE

IDEALNE CENTRUM DANYCH JEST JAK KOMPUTER



O SAMOPODOBIEŃSTWIE

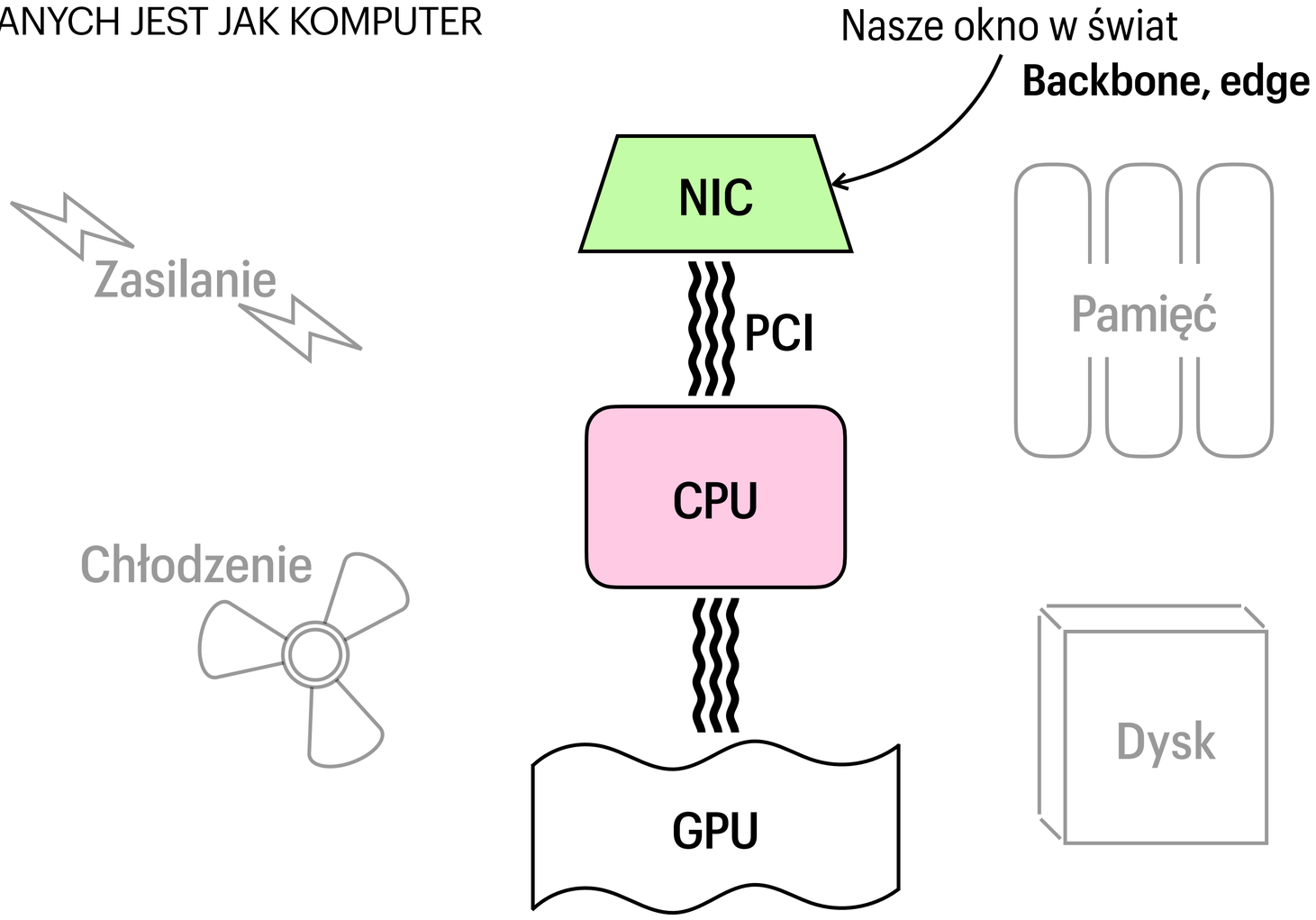
IDEALNE CENTRUM DANYCH JEST JAK KOMPUTER



Dedykowane zasoby obliczeniowe
np. **Big Data, AI, inference**

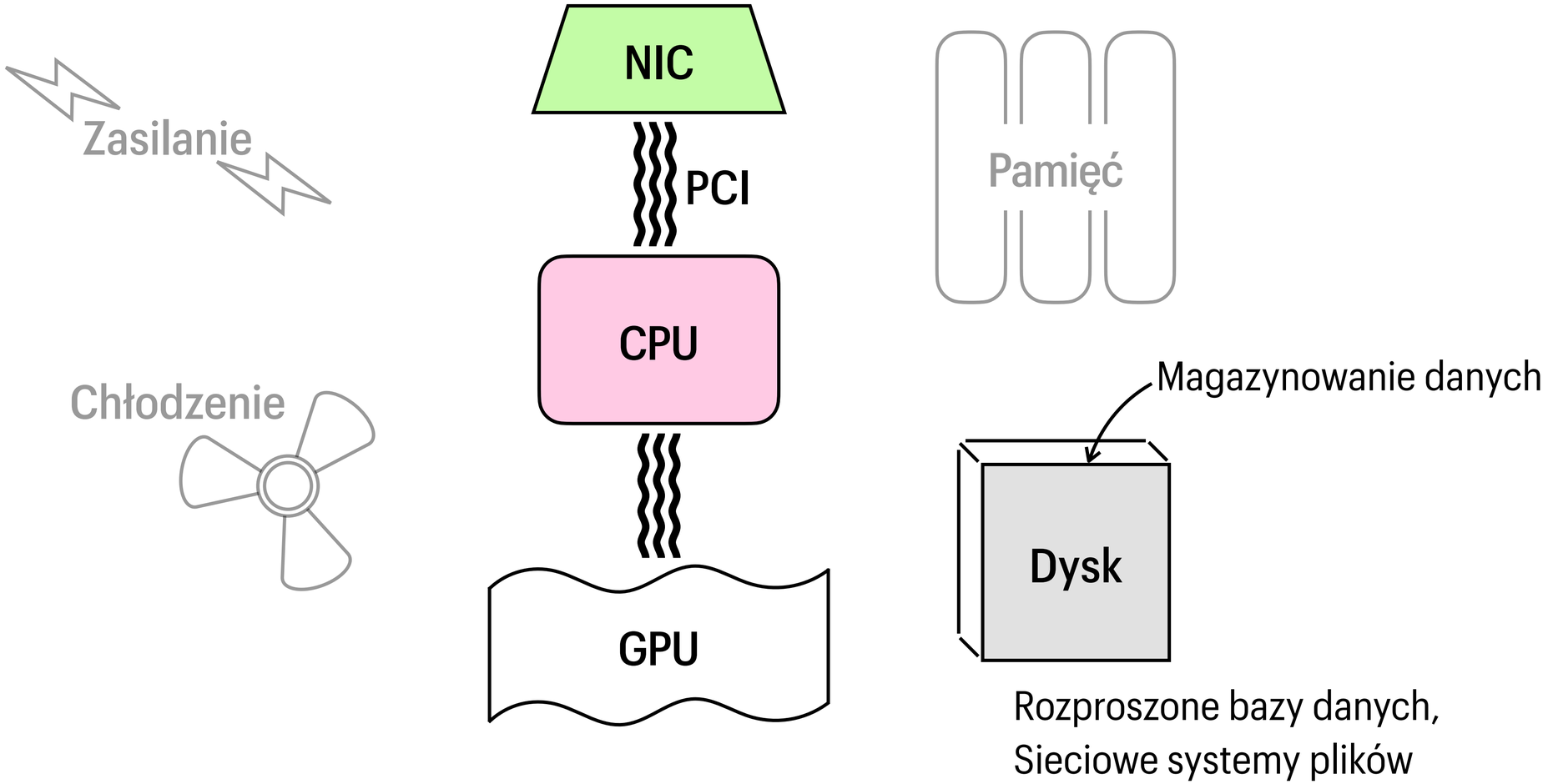
O SAMOPODOBIEŃSTWIE

IDEALNE CENTRUM DANYCH JEST JAK KOMPUTER



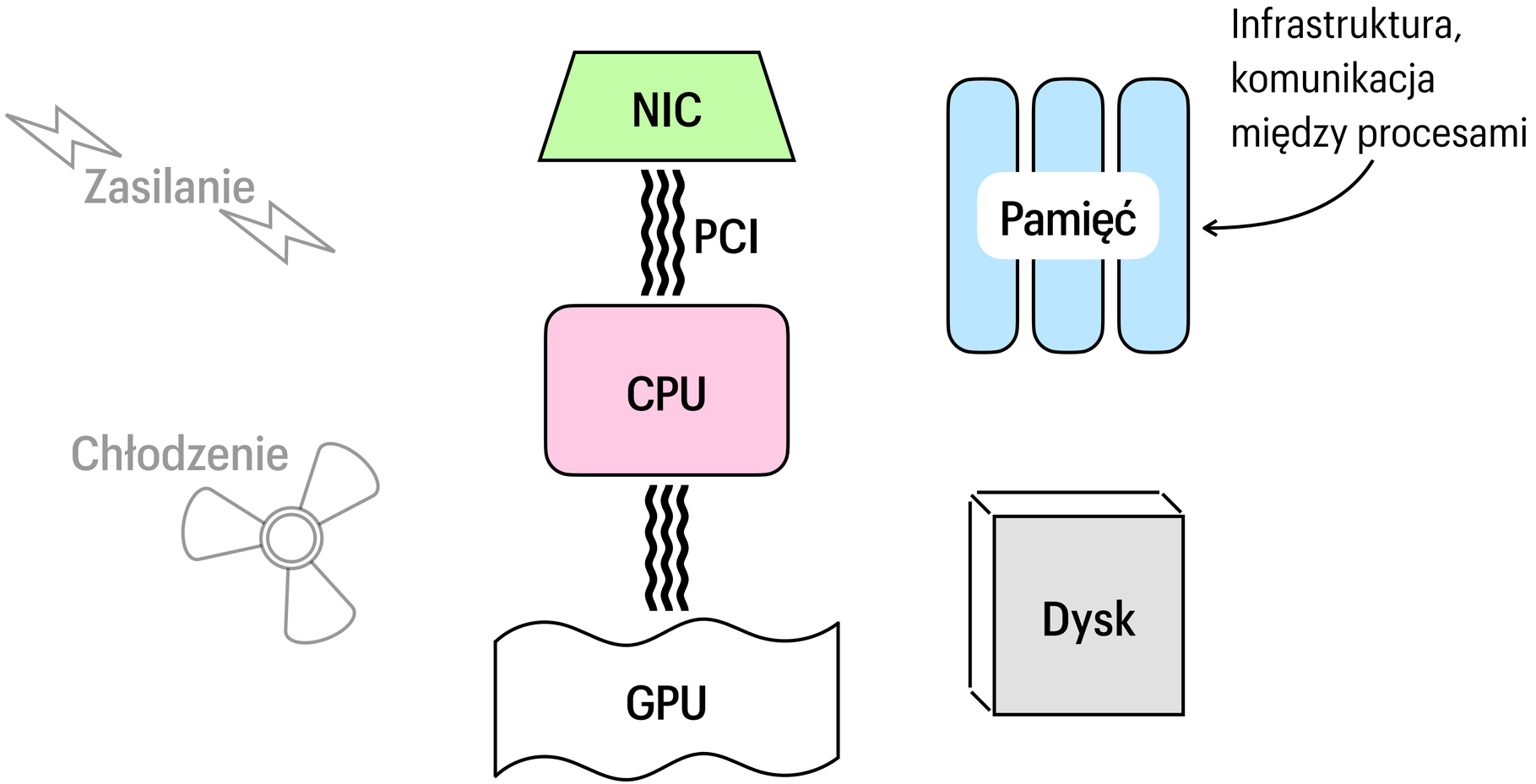
O SAMOPODOBIEŃSTWIE

IDEALNE CENTRUM DANYCH JEST JAK KOMPUTER



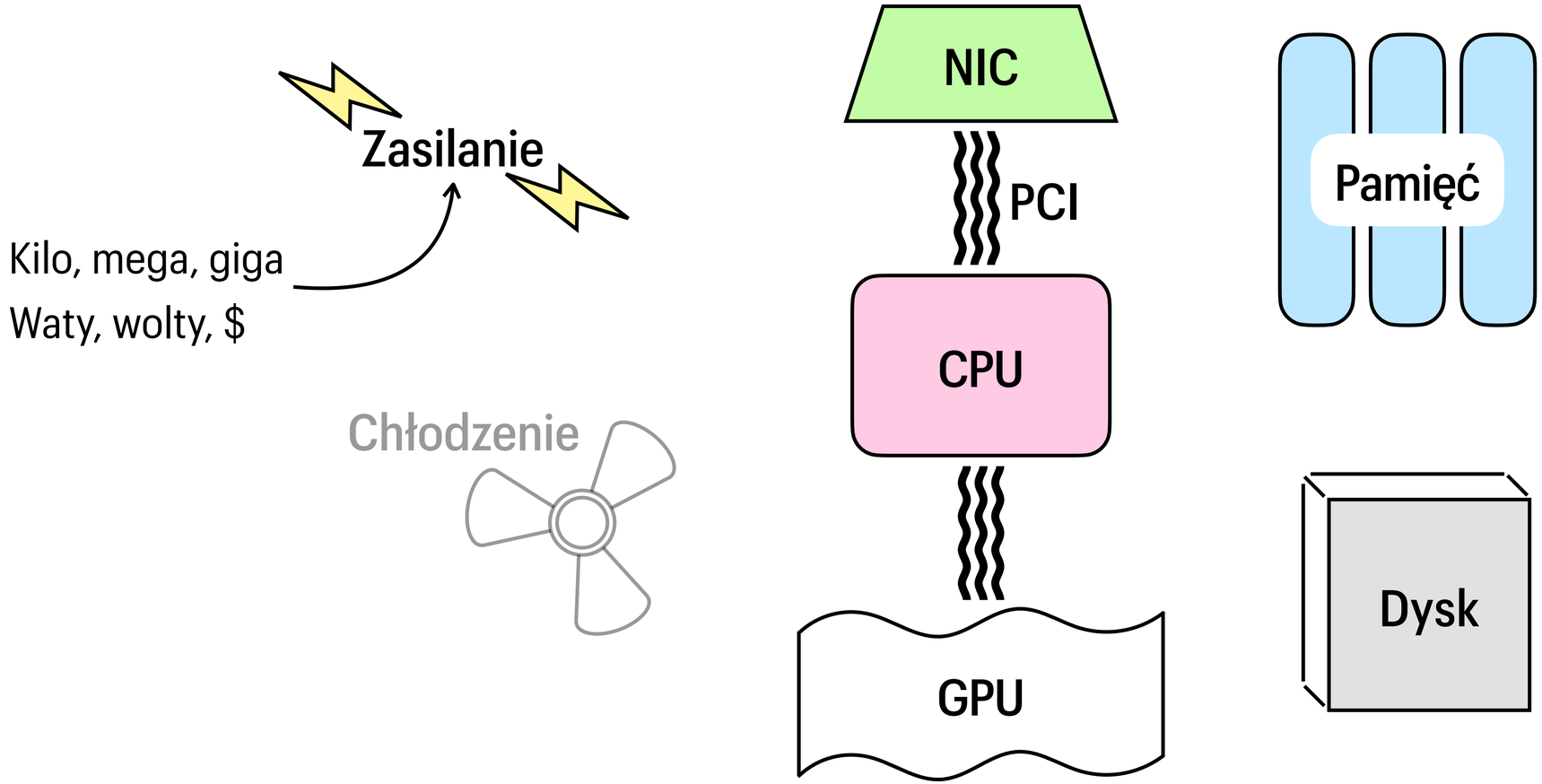
O SAMOPODOBIEŃSTWIE

IDEALNE CENTRUM DANYCH JEST JAK KOMPUTER



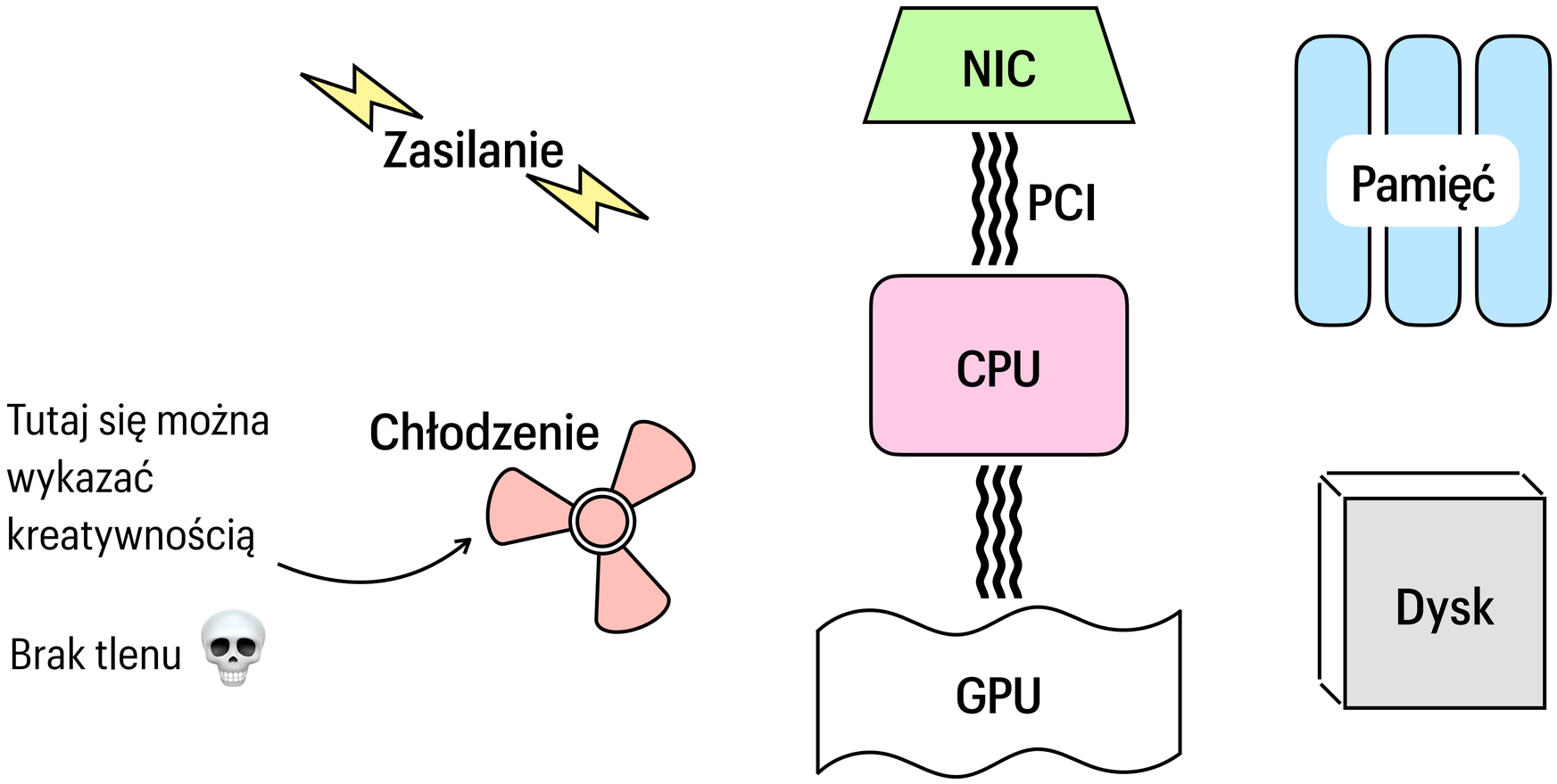
O SAMOPODOBIEŃSTWIE

IDEALNE CENTRUM DANYCH JEST JAK KOMPUTER



O SAMOPODOBIEŃSTWIE

IDEALNE CENTRUM DANYCH JEST JAK KOMPUTER

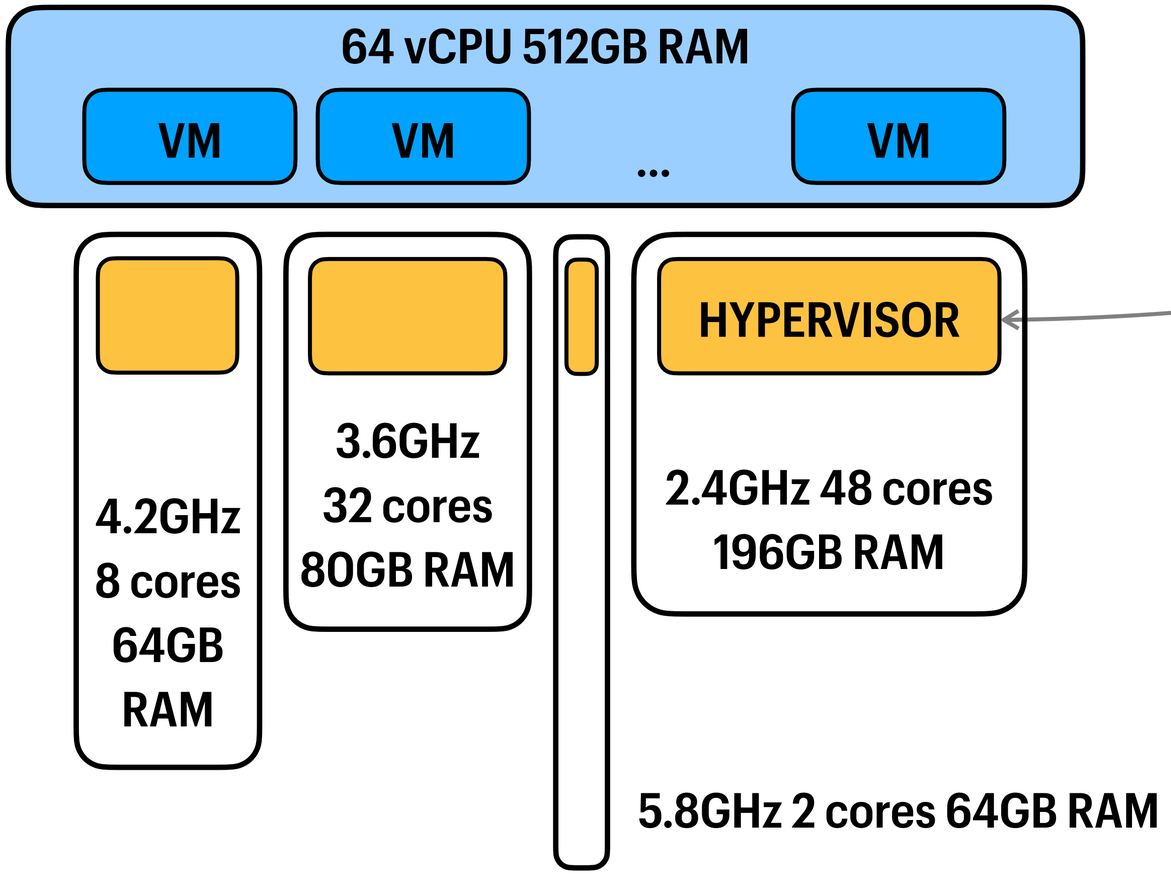


Zasoby obliczeniowe

NIKITA MURAVYTSKYI AGH

O WIRTUALIZACJI

KOMPUTERY JUŻ OD DAWNA ŻYJĄ W METAVERSUM



🤔 Dlaczego dostępna pamięć fizyczna > sumaryczna pamięć VM?

hypervisor zużywa zasoby przeznaczone dla procesów biznesowych

vCPU jest uniwersalną jednostką zasobów obliczeniowych

Pozwala to działać na równomiernej i homogenicznej infrastrukturze bez przywiązania do sprzętu

O BASEBOARD MANAGEMENT CONTROLLER

STEROWANIE I MONITORING NIEZALEŻNY OD CPU, OS I FIRMWARE

Jakie problemy rozwiązuje **BMC**:

- oddzielenie sterowania do osobnej niezależnej warstwy;

- sterowanie systemem niezależnie od systemu operacyjnego;

- wykonywanie operacji, które normalnie wymagają fizycznej interakcji z urządzeniem.

O BASEBOARD MANAGEMENT CONTROLLER

STEROWANIE I MONITORING NIEZALEŻNY OD CPU, OS I FIRMWARE

Jakie problemy rozwiązuje **BMC**:

- oddzielenie sterowania do osobnej niezależnej warstwy;

- sterowanie systemem niezależnie od systemu operacyjnego;

- wykonywanie operacji, które normalnie wymagają fizycznej interakcji z urządzeniem.

Np. agent, monitorujący system, przestaje wysyłać **metryki** jak system się przeciąży - wtedy kiedy tego potrzebujemy najbardziej

W przypadku błędów albo braku systemu operacyjnego

Np. fizycznie wywołany reboot - naciśnięciem guzika na płycie

🤔 Jak zdalnie zainstalować OS na serwerze?

O SMART NIC'ACH

CZYLI ZWYKŁA KARTA SIECIOWA JEST GŁUPIA?

Po co custom'owa logika na karcie sieciowej?

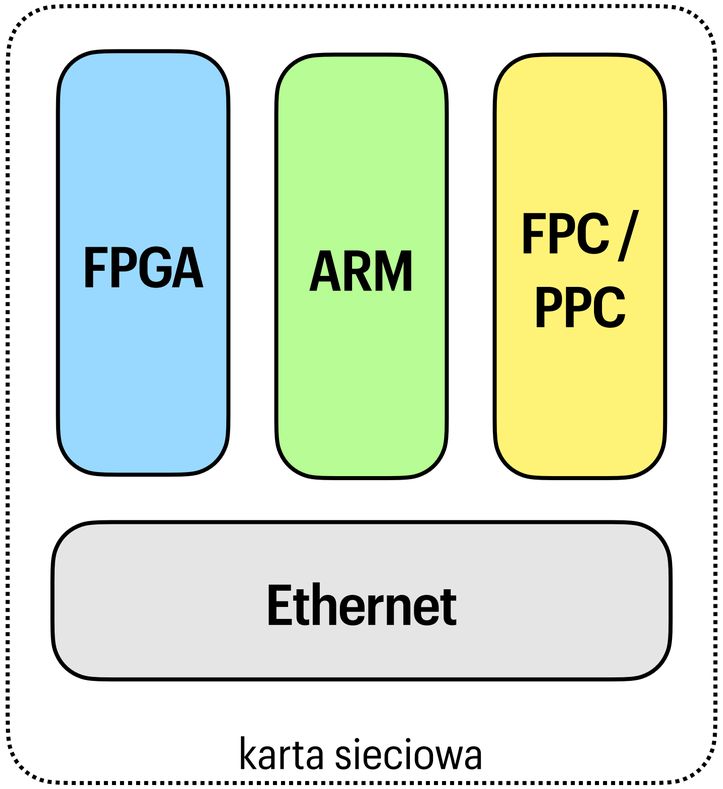
- przeniesienie logiki sieciowej z CPU na dedykowany krzem tzw. **offloading**;
- większa wydajność, mniejsze opoźnienia;
- większa elastyczność, dostosowanie pod konkretne zastosowanie.

Np. Offloading load balancer'a

Przecież możemy po prostu wziąć szybszą kartę sieciową!

O SMART NIC'ACH

PROGRAMOWALNA KARTA SIECIOWA



Każdy wybiera co mu pasuje:

Broadcom - NetXtreme - ARMs oraz ~~TruFlow~~ FPC

Nvidia - ConnectX - ARMs oraz ConnectX

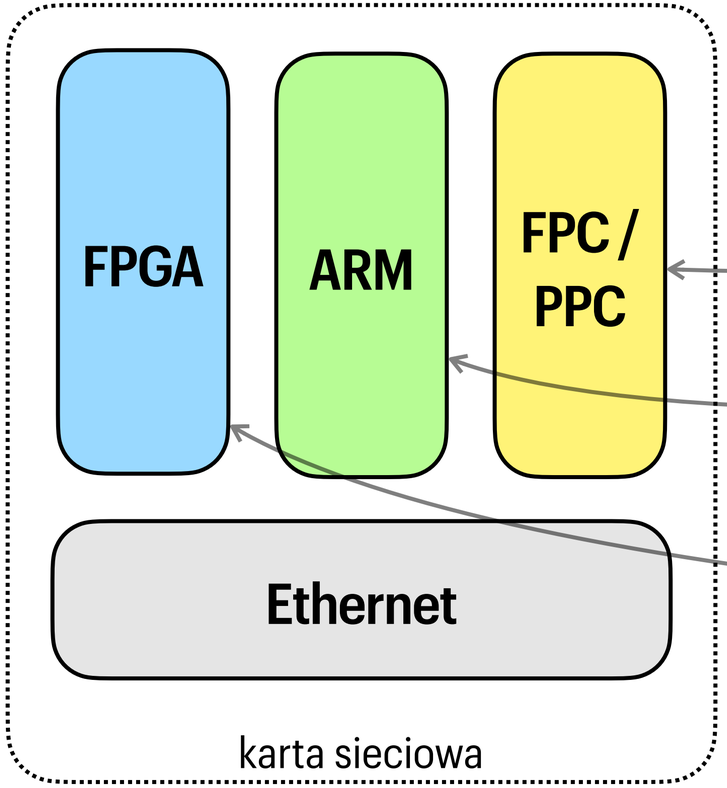
Intel - N3000 - x86 oraz FPGA

Netronome - NFP4000 - ARM oraz FPC/PPC

O SMART NIC'ACH

PROGRAMOWALNA KARTA SIECIOWA

🤔 Dlaczego ARM, nie x86?



Flow / Packet Processing Core - programowalne rdzenie do przetwarzania pakietów i przepływów

ARM do zadań **warstwy zarządzania** - lądowanie kodu na FCP, zbieranie metryk, logów itd.

FPGA do skomplikowanych zadań **warstwy danych** - tunele, szyfrowanie, wektorowe przetwarzanie pakietów itd.

O DATA PROCESSING UNIT

DLACZEGO WSPÓŁCZESNE SERWERY TO NESCAFE 3 W 1

Co jeszcze możemy odpalić na waszej karcie?

Tak.

O DATA PROCESSING UNIT

WSZYSTKO CO NIE GENERUJE ZYSKU OPUSZCZA CPU

Filozofia prosta - **offload'ujemy** na DPU wszystko co nie jest biznes procesem:

- operacje wejścia/wyjścia;
- hypervisor;
- security.

Np. Nitro cards by AWS

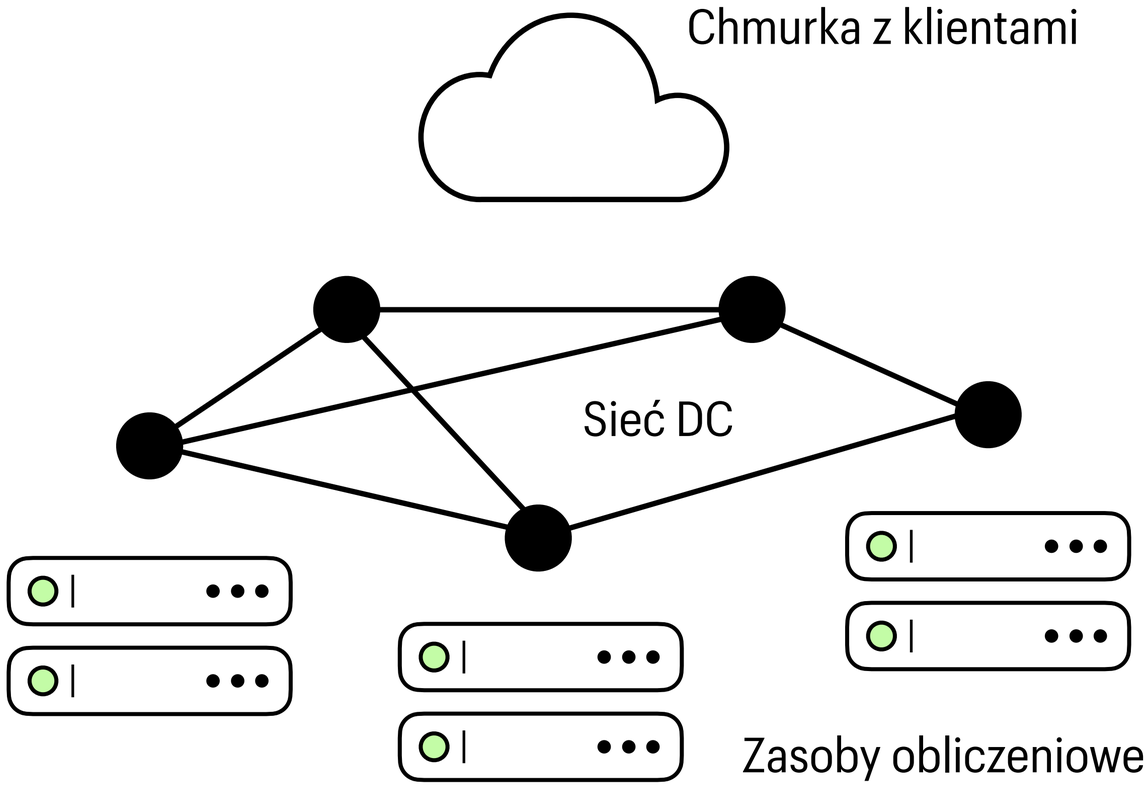
Sieciówka, system plików, silnik bazy danych

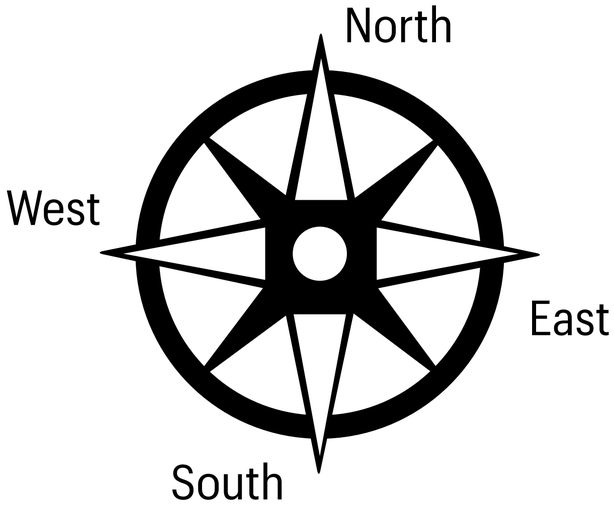
Wykorzystujemy dedykowany krzem gdzie to tylko jest możliwe.

Sieciówka

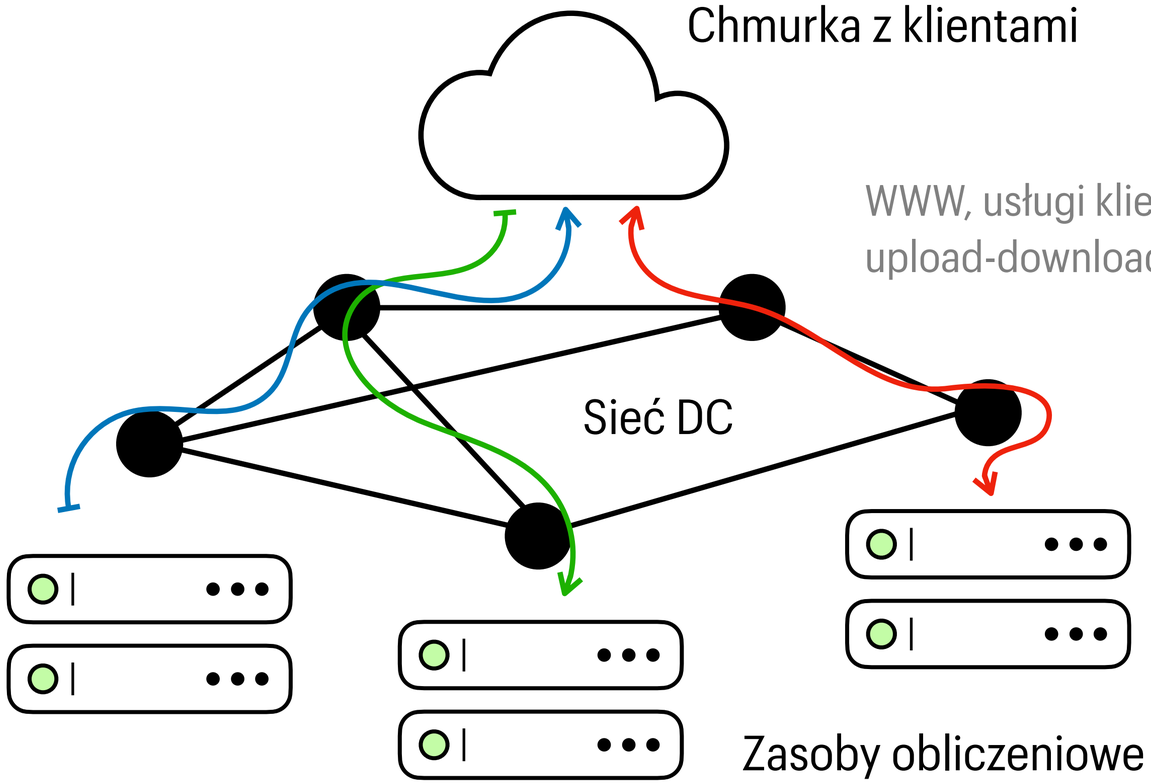
NIKITA MURAVYTSKYI AGH

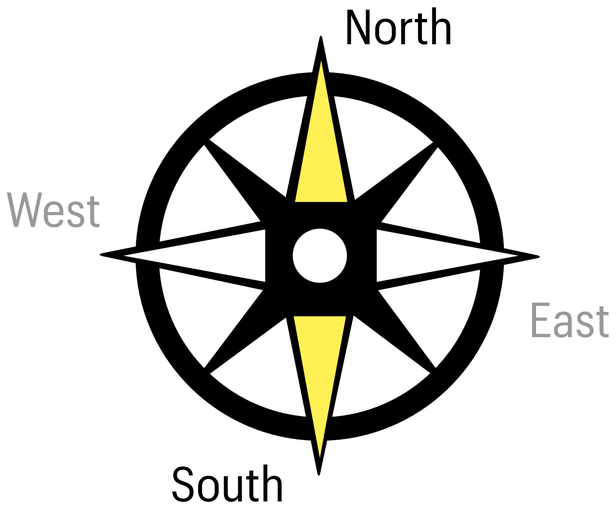
CO POKAZUJE KOMPAS W CENTRUM DANYCH?





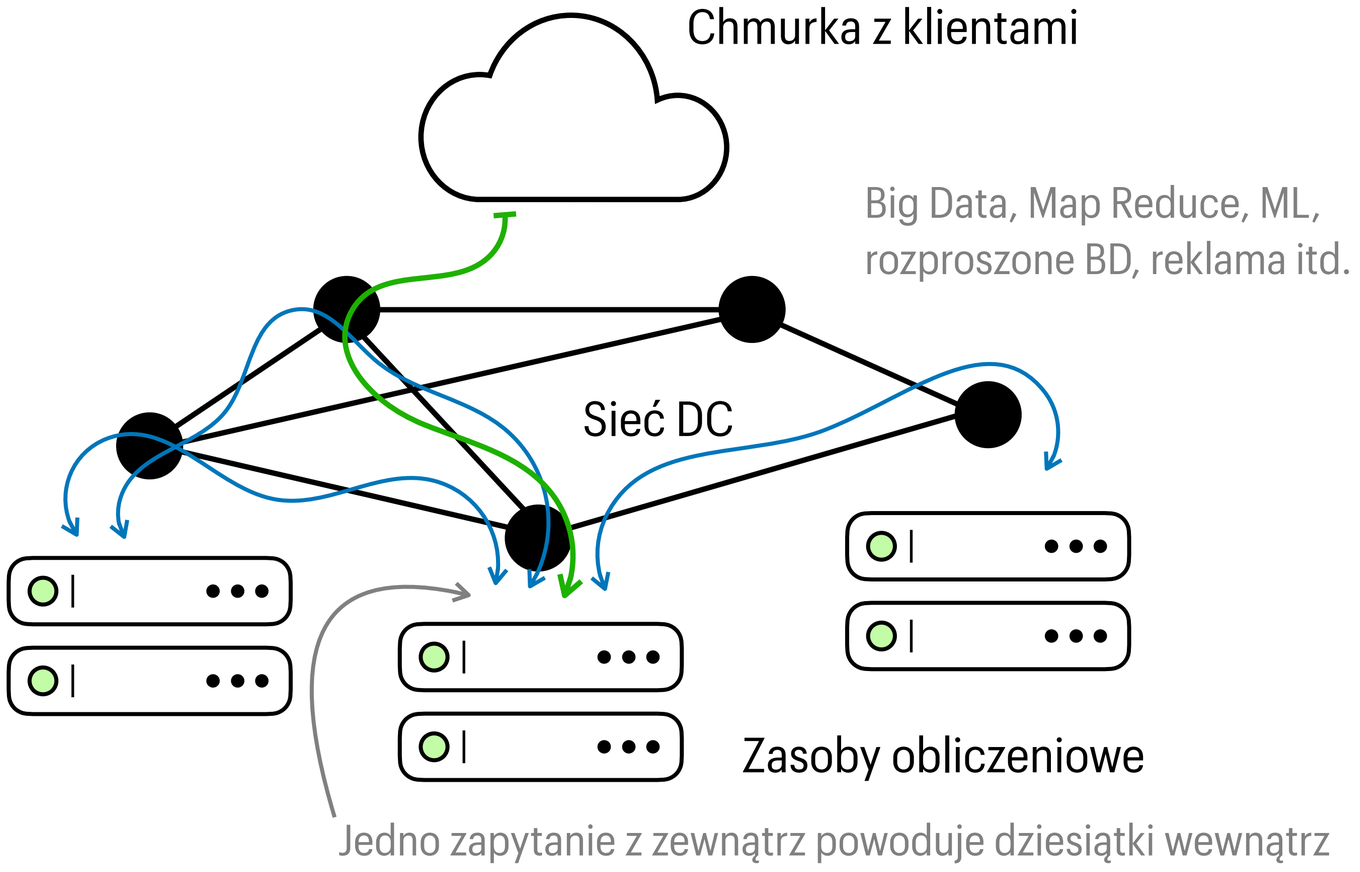
CO POKAZUJE KOMPAS W CENTRUM DANYCH?

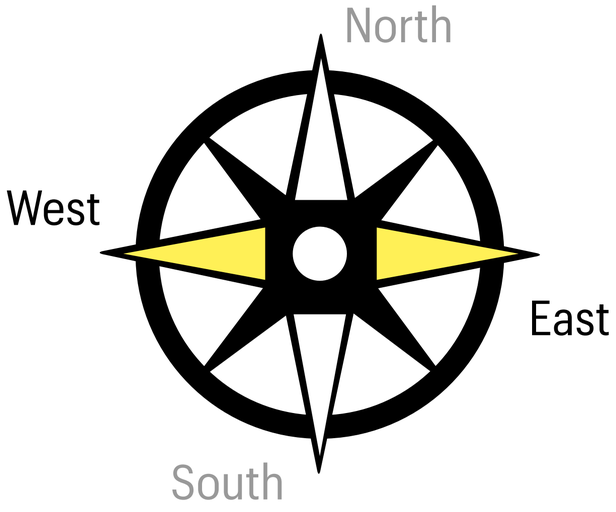




Ruch w płaszczyźnie
serwery - wejście/wyjście DC

CO POKAZUJE KOMPAS W CENTRUM DANYCH?





Ruch pomiędzy serwerami w DC

Keep in mind:

- sieć musi być niezależna od warstwy aplikacji;

- liczba portów w przełącznikach tzw. **radix** jest ograniczona;

- największego ruchu się spodziewamy w płaszczyźnie East-West;

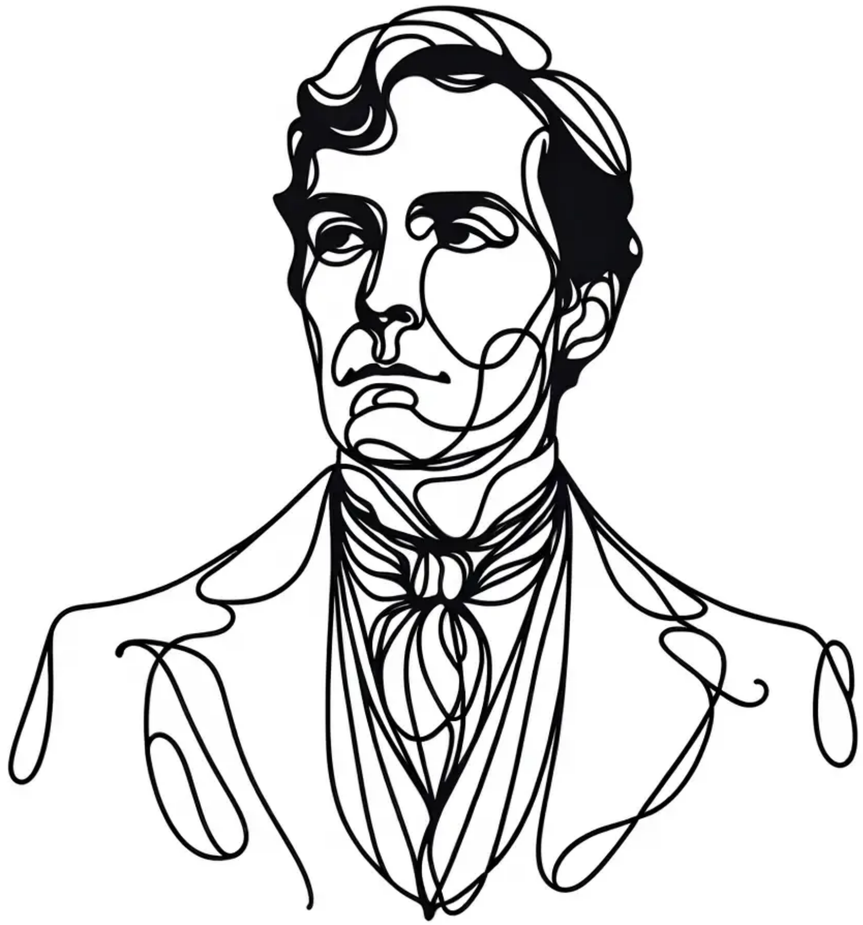
- sieć nie powinna być problemem dla skalowania.

O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

NIEBLOKOWALNE POŁĄCZENIA W SIECI Z KOMUTACJĄ KANAŁÓW

Topologia Clos'a:

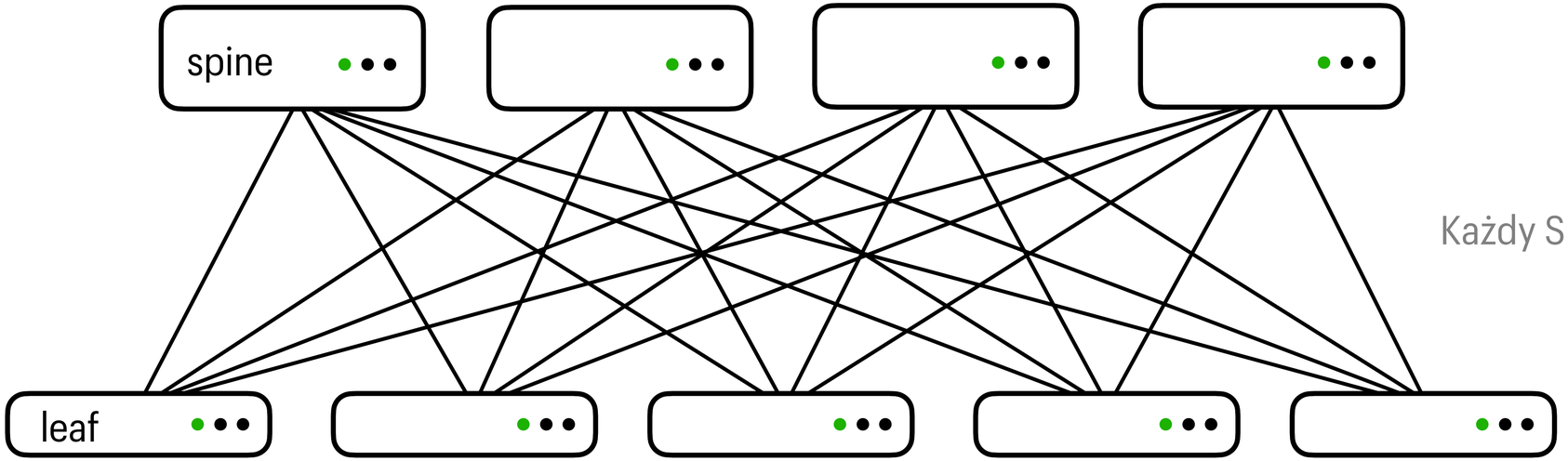
- rozwiązuje problem połączenia węzłów, gdzie każdy może komunikować się z każdym;
- Pierwotnie stworzona do komutacji połączeń w centralach telefonicznych;
- zakłada równomierny dostęp do pasma.



O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

STANDARD TOPOLOGII FIZYCZNYCH W CENTRUM DANYCH

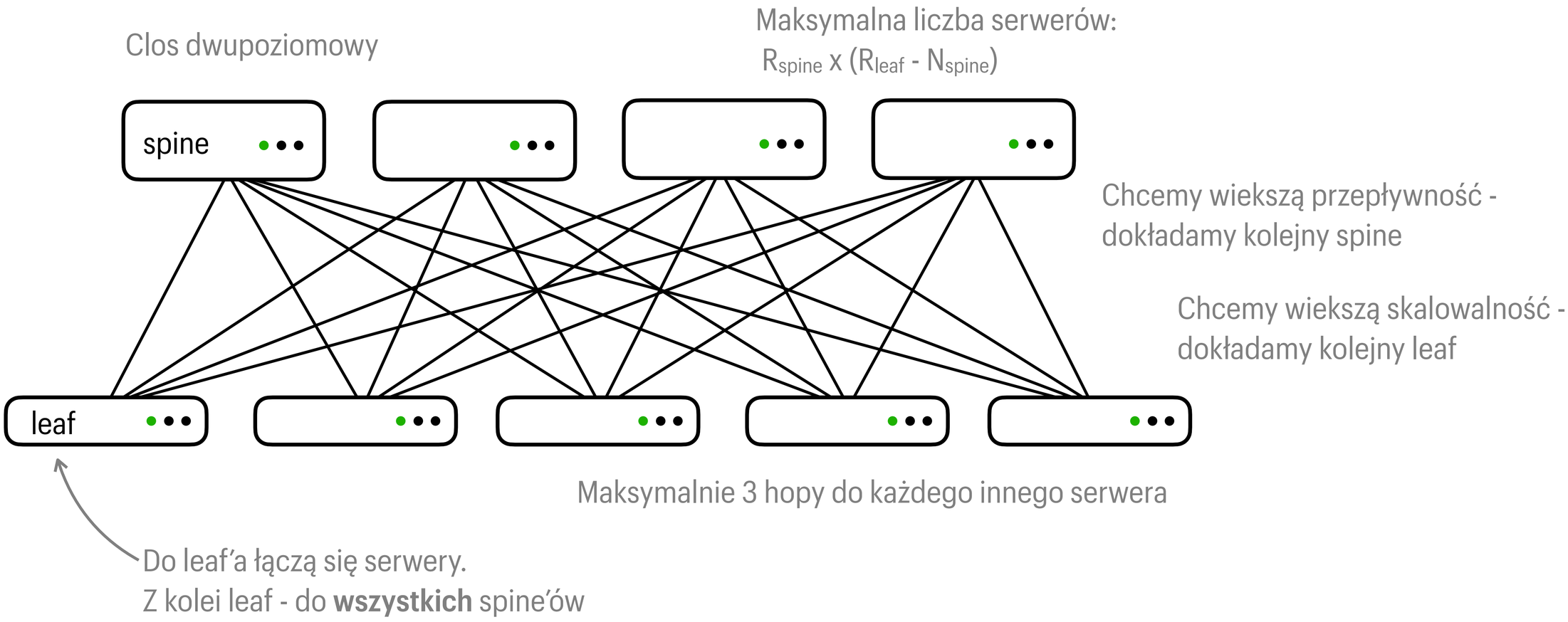
Clos dwupoziomowy



Każdy Spine do każdego Leaf’a

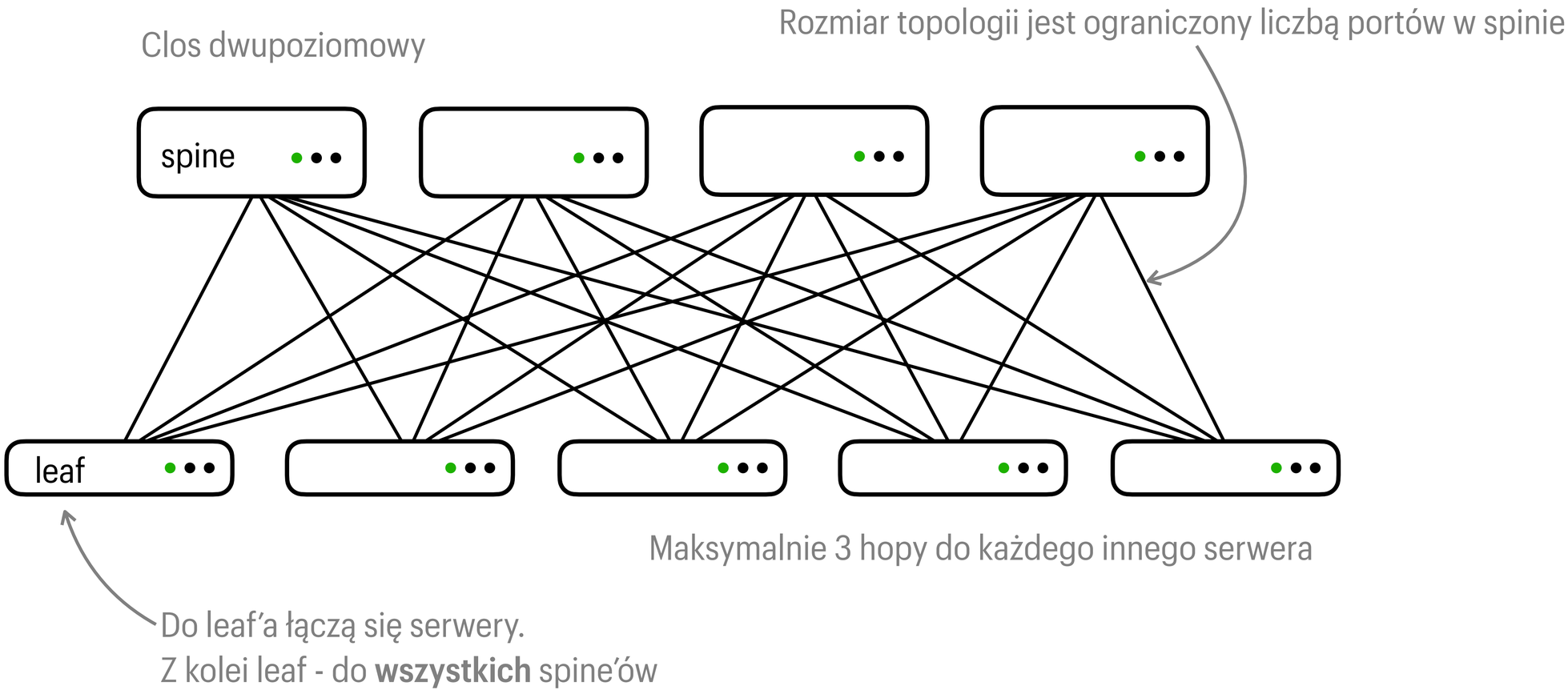
O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

STANDARD TOPOLOGII FIZYCZNYCH W CENTRUM DANYCH



O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

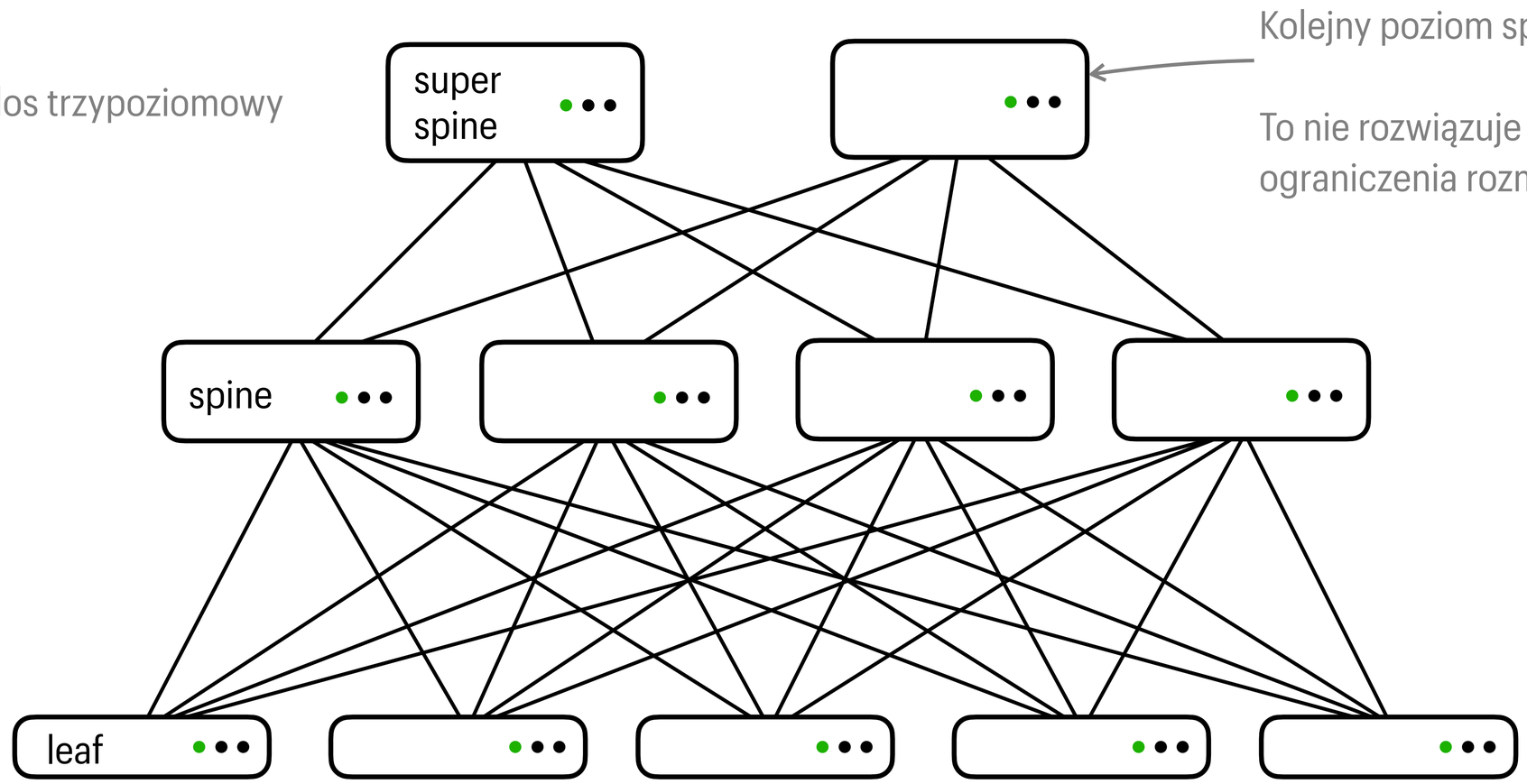
STANDARD TOPOLOGII FIZYCZNYCH W CENTRUM DANYCH



O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

JAKOŚ DUŻO TYCH KRESEK SIĘ ZROBIŁO

Clos trzypoziomowy



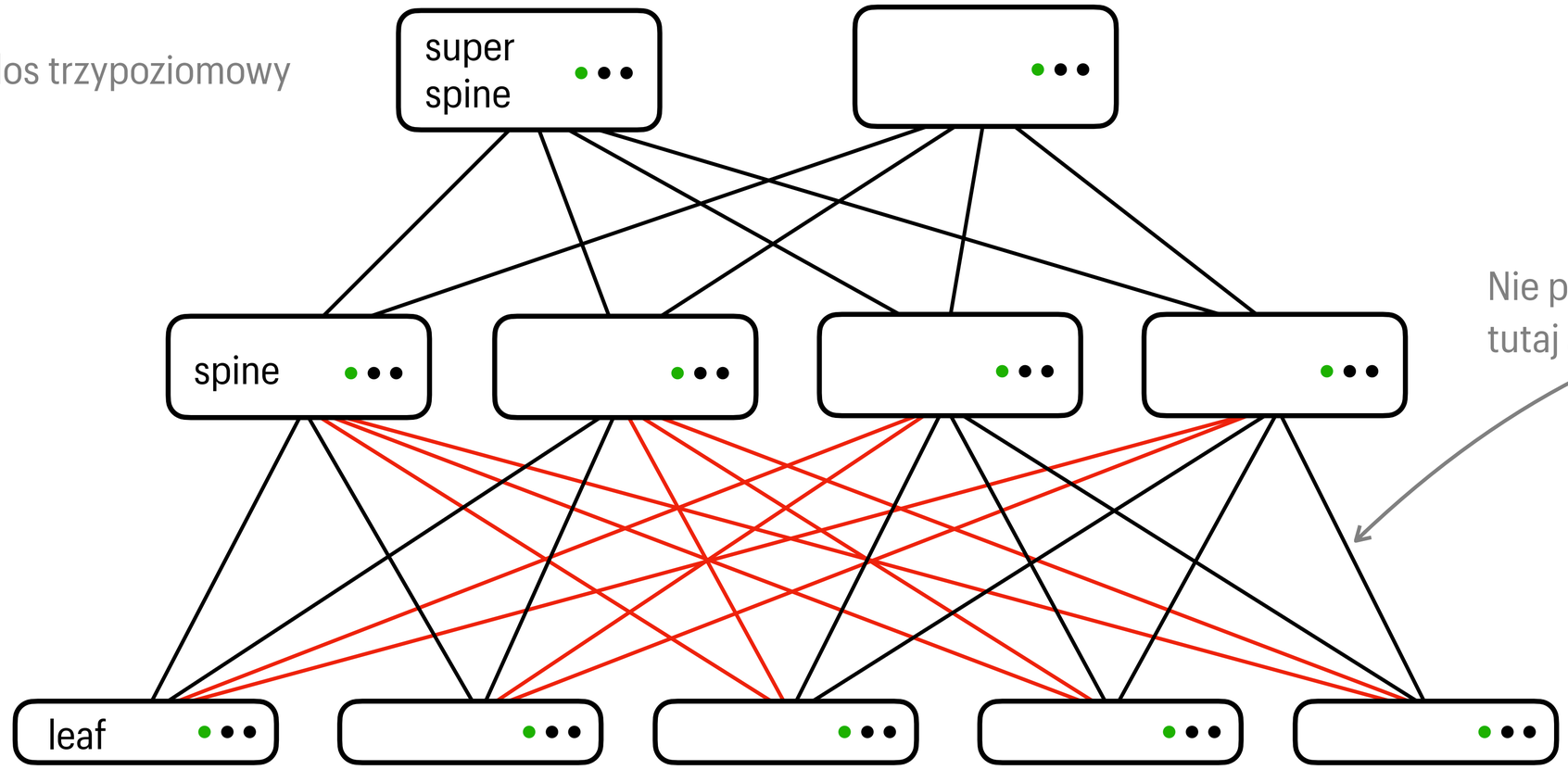
Kolejny poziom spinelów.

To nie rozwiązuje problemu
ograniczenia rozmiaru topologii

O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

JAKOŚ DUŻO TYCH KRESEK SIĘ ZROBIŁO

Clos trzypoziomowy

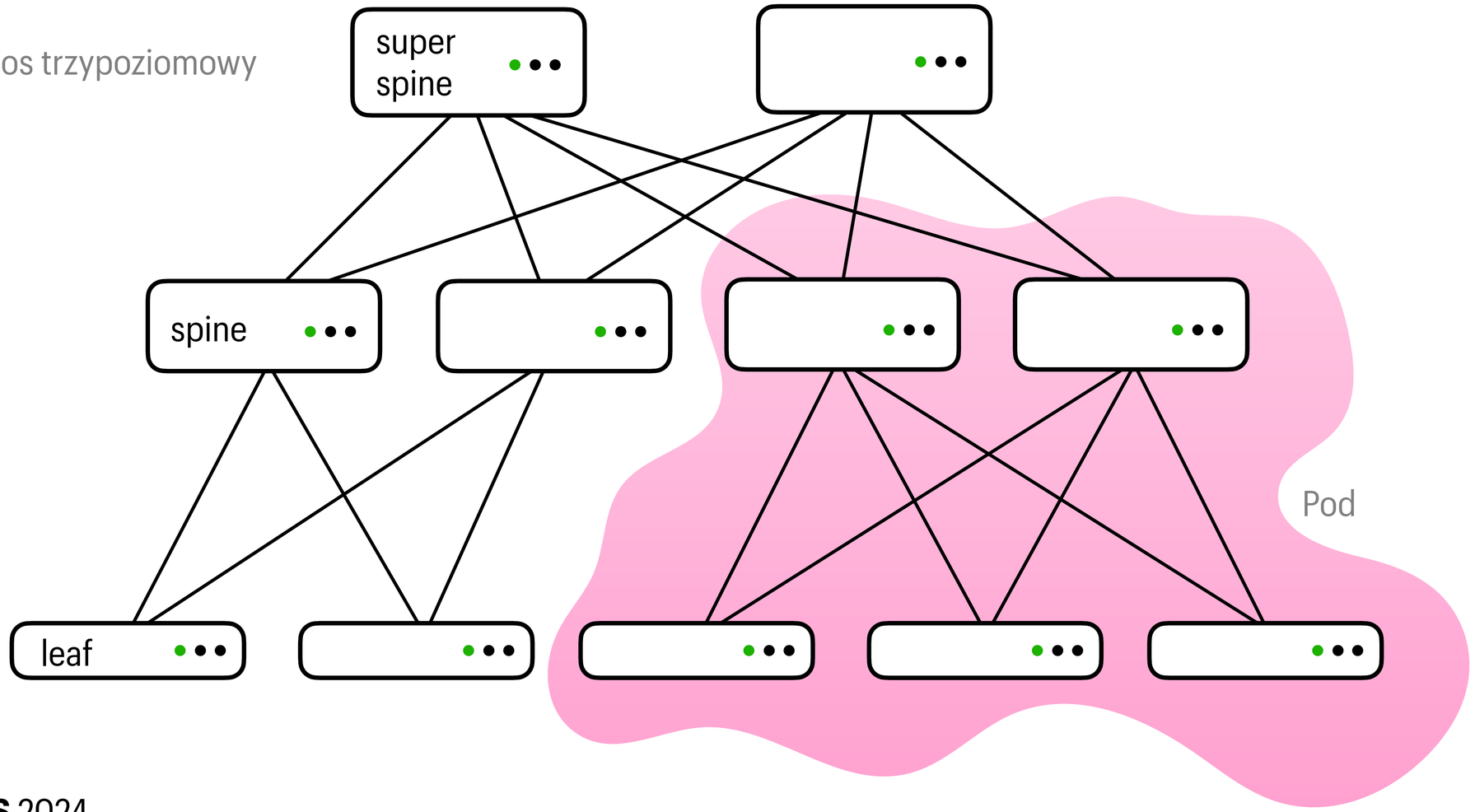


Nie potrzebujemy tylu przepływności tutaj aka. **oversubscription**

O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

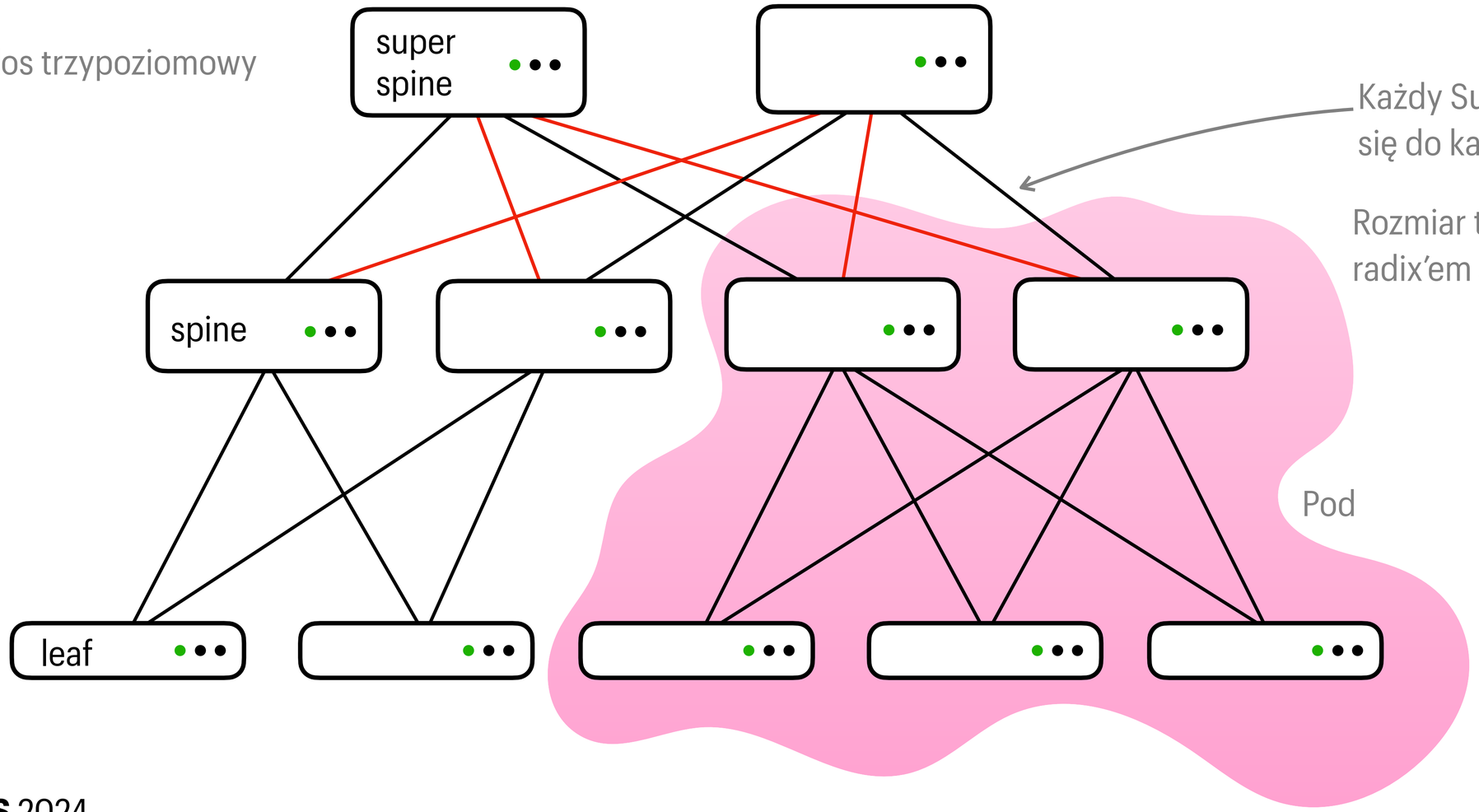
POD - KLOCEK LEGO W TOPOLOGII CENTRUM DANYCH

Clos trzypoziomowy



POD - KLOCEK LEGO W TOPOLOGII CENTRUM DANYCH

Clos trzypoziomowy



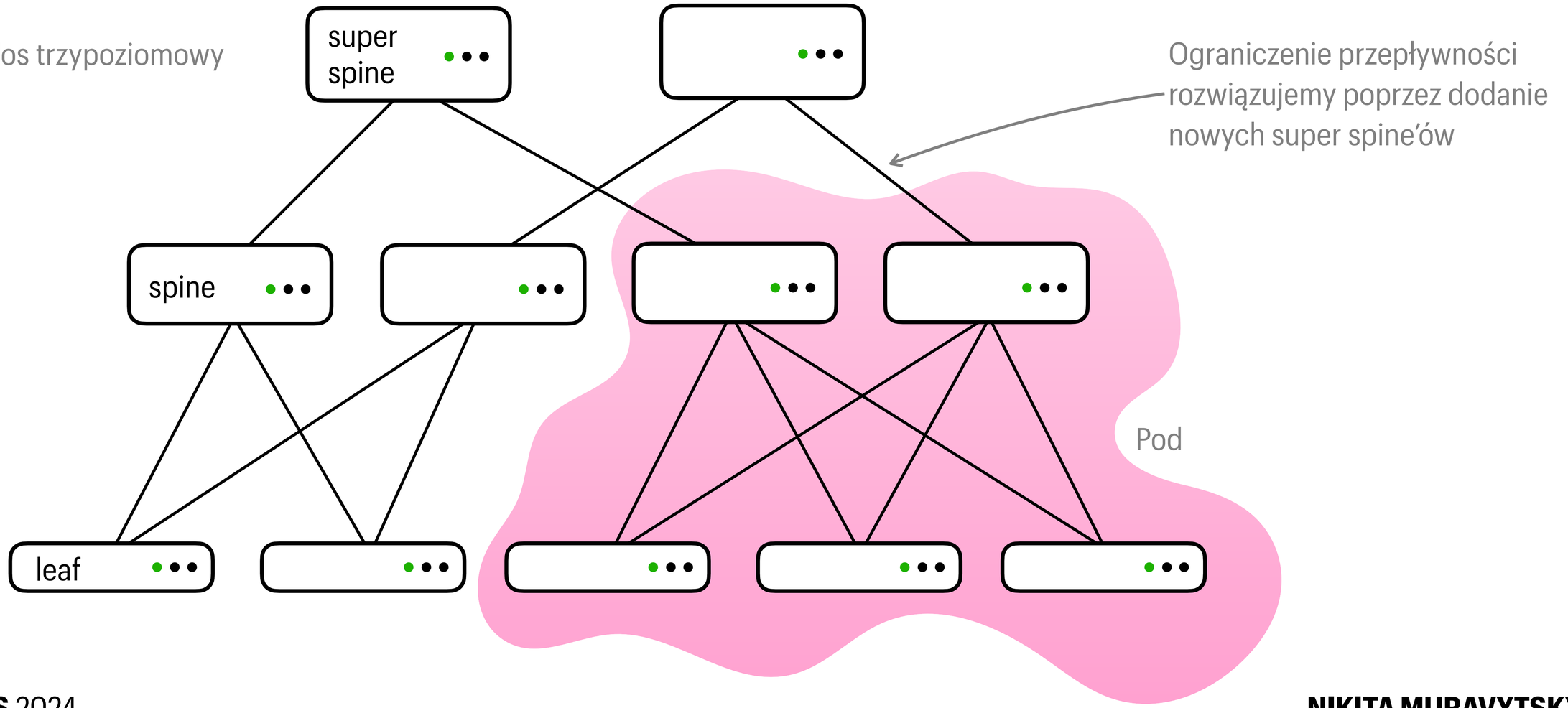
Każdy Super spine łączy się do każdego spine'a.

Rozmiar topologii jest ograniczony radix'em super spine'ów

O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

TOPOLOGIA CLOSA NA REDUKCJI

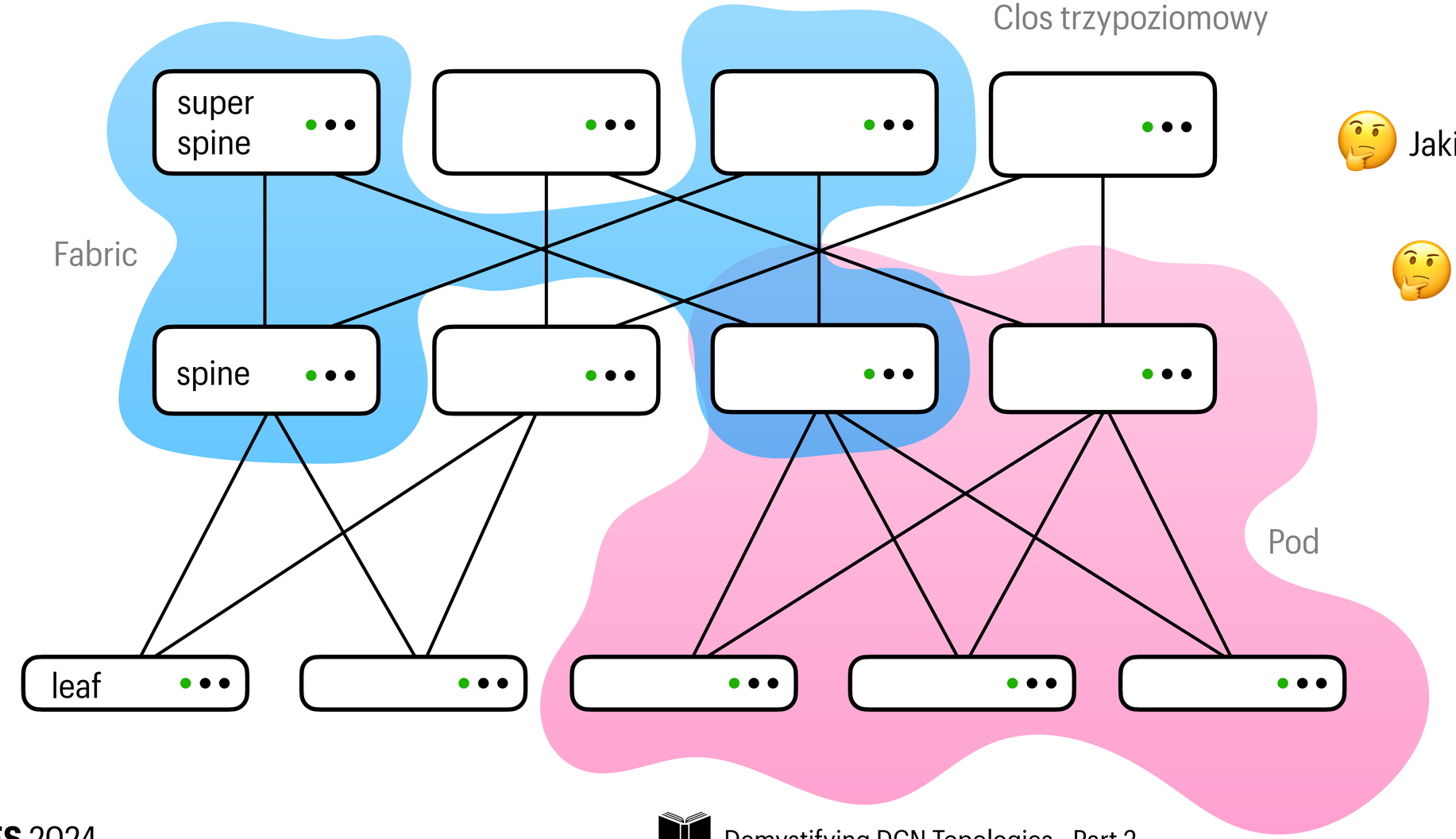
Clos trzypoziomowy



Ograniczenie przepływności rozwiązujemy poprzez dodanie nowych super spine'ów

O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

FABRIC - ŁĄCZNOŚĆ POMIĘDZY POD'AMI



🤔 Jakie są wady Clos'a?

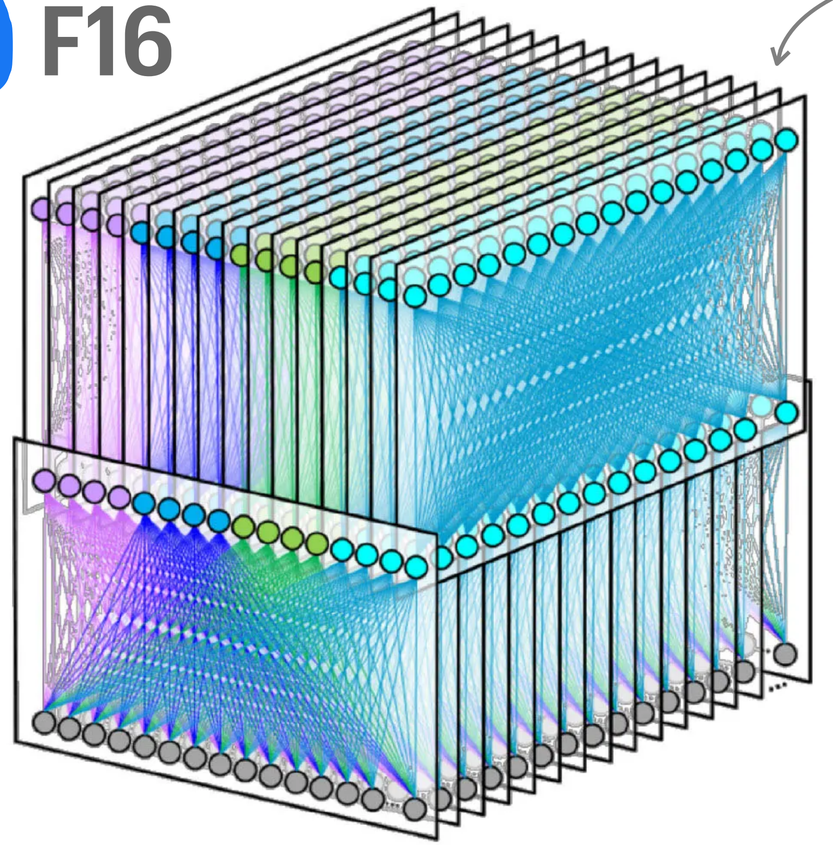
🤔 Czy są alternatywy?

tip: google Dragonfly

O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

CASE STUDY - INFRASTRUKTURA METAVERSUM

F16



Każdy kolor to osobny control plane

Super spine
Spine SW

Spine
Fabric SW

Leaf
Top-of-rack
Rack SW

- 16 x 100Gbps z ToR do Spine

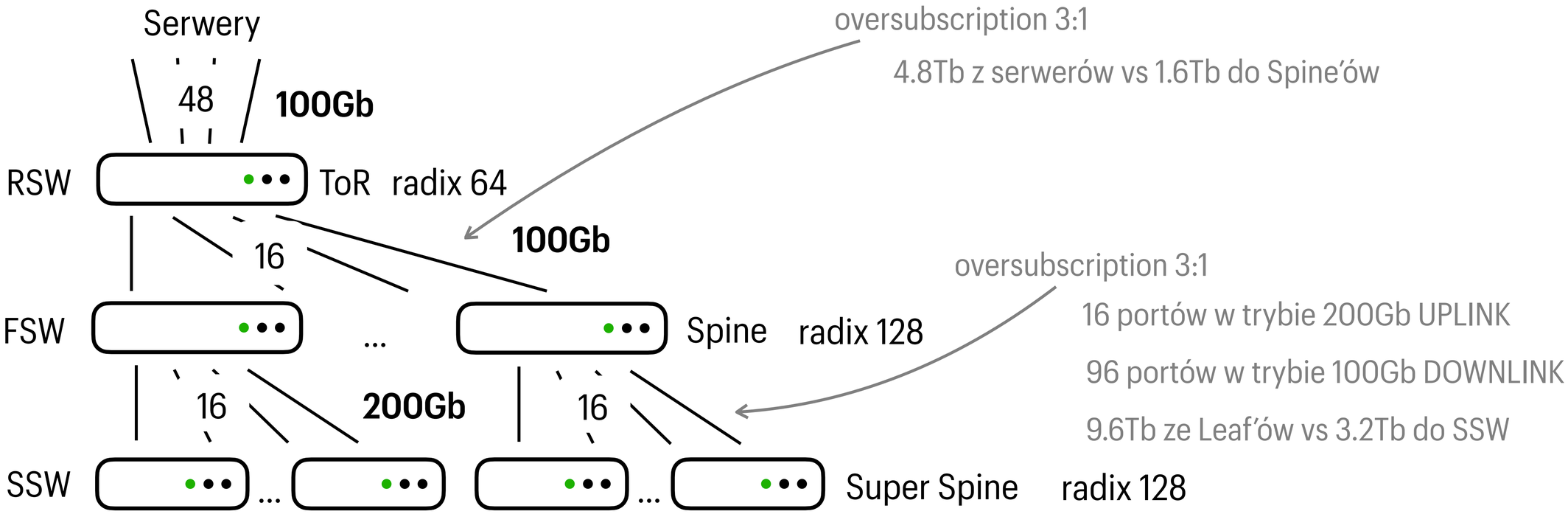
- 16 x 100Gbps ze Spine do Fabric

- mieści 500k - 1000k serwerów

- zbudowany na Broadcom Tomahawk 3
  - 12.8 Tb/s
  - 25.6 Tb/s Tomahawk 4
  - 51.2 Tb/s Tomahawk 5

O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

CASE STUDY - INFRASTRUKTURA METAVERSUM



Maksymalna pojemność: 48 serwerów x 96 RSW per Pod x 64 Fabric ~ 300k per F16

O TYM JAK PAN CLOS PRZEKOCZOWAŁ Z CENTRALI TELEFONICZNYCH DO DC

CASE STUDY - INFRASTRUKTURA METAVERSUM

Clonee, Irlandia

Największe Centrum Danych Mety w Europie

5 budynków

1.2 miliona serwerów

600MW

O TYM JAK BGP PODBIŁ ŚWIAT DC

JAK TO WSZYSTKO TERAZ SKONFIGUROWAĆ?

L2 wlecze nierównomierne obciążenie linków

- tylko **IP** czyli routing na wszystkich linkach, oprócz może serwer - Top-of-the-Rack;

ECMP - tlen dla topologii Clos'a

- dla równomiernego obciążania linków w górę - **ECMP**;

- ze względu na elastyczność, agregację prefiksów i polityki rutingu - **BGP**;

De facto standard dla DC

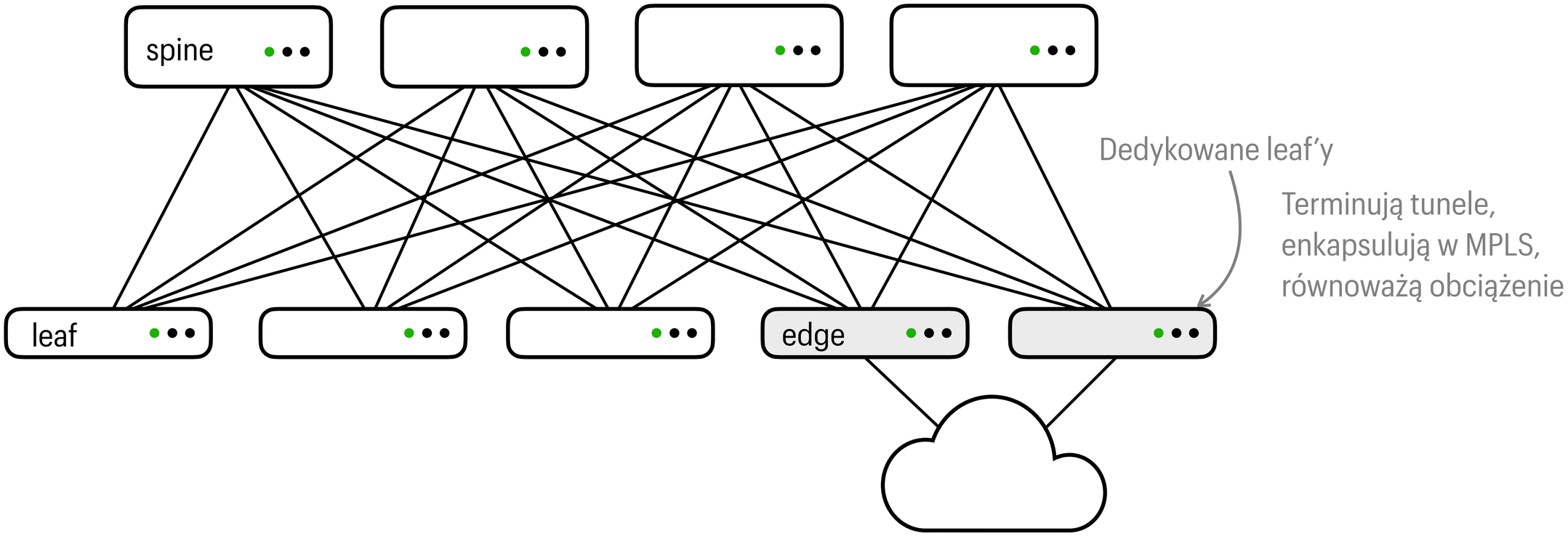
- sieci nakładkowe jeśli potrzebujemy izolacji.

Edge, Backbone

NIKITA MURAVYTSKYI AGH

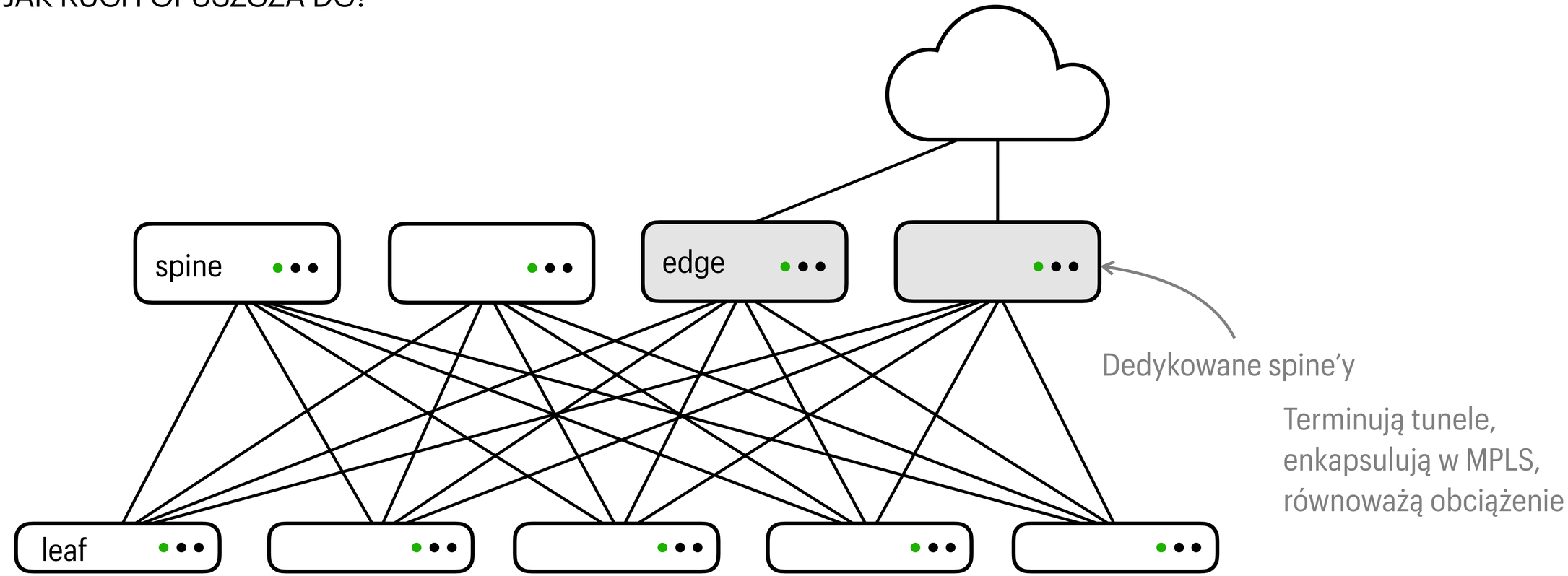
O HIGHWAY'ACH WSPÓŁCZESNEGO INTERNETU

JAK RUCH OPUSZCZA DC?



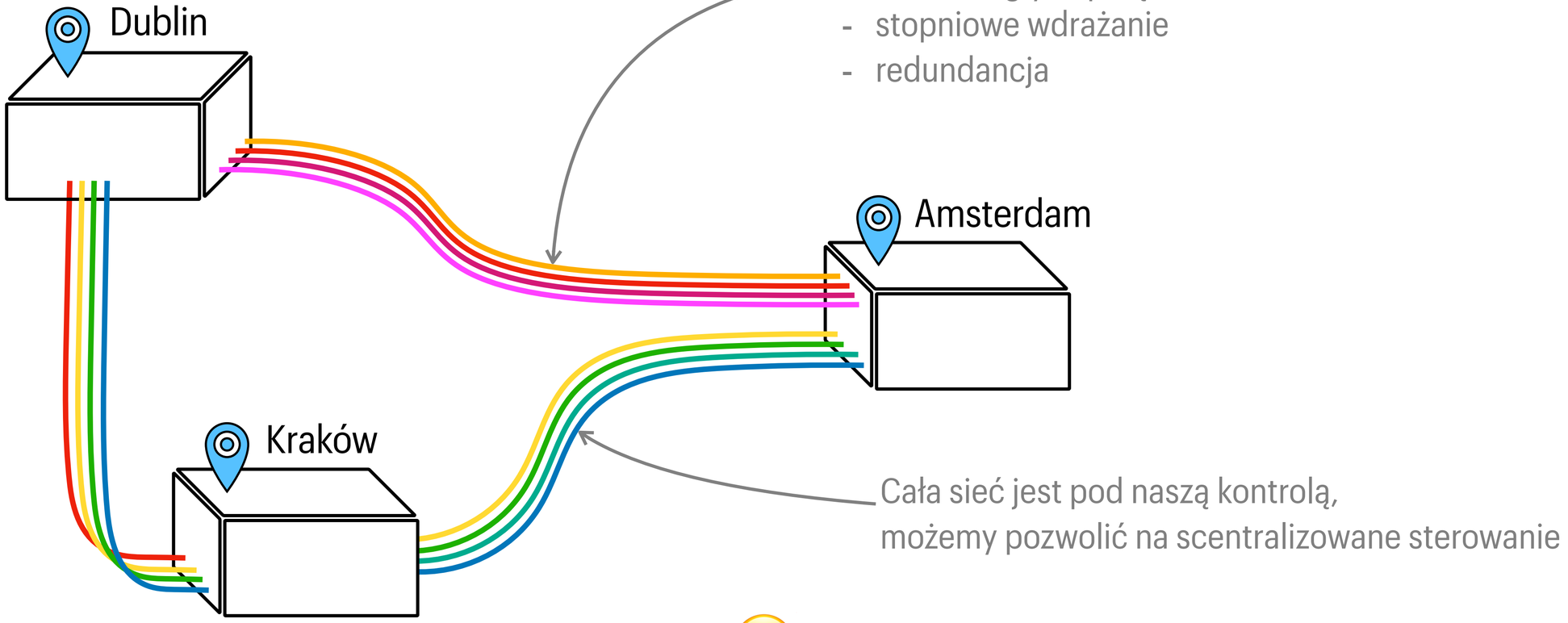
O HIGHWAY'ACH WSPÓŁCZESNEGO INTERNETU

JAK RUCH OPUSZCZA DC?



O HIGHWAY'ACH WSPÓŁCZESNEGO INTERNETU

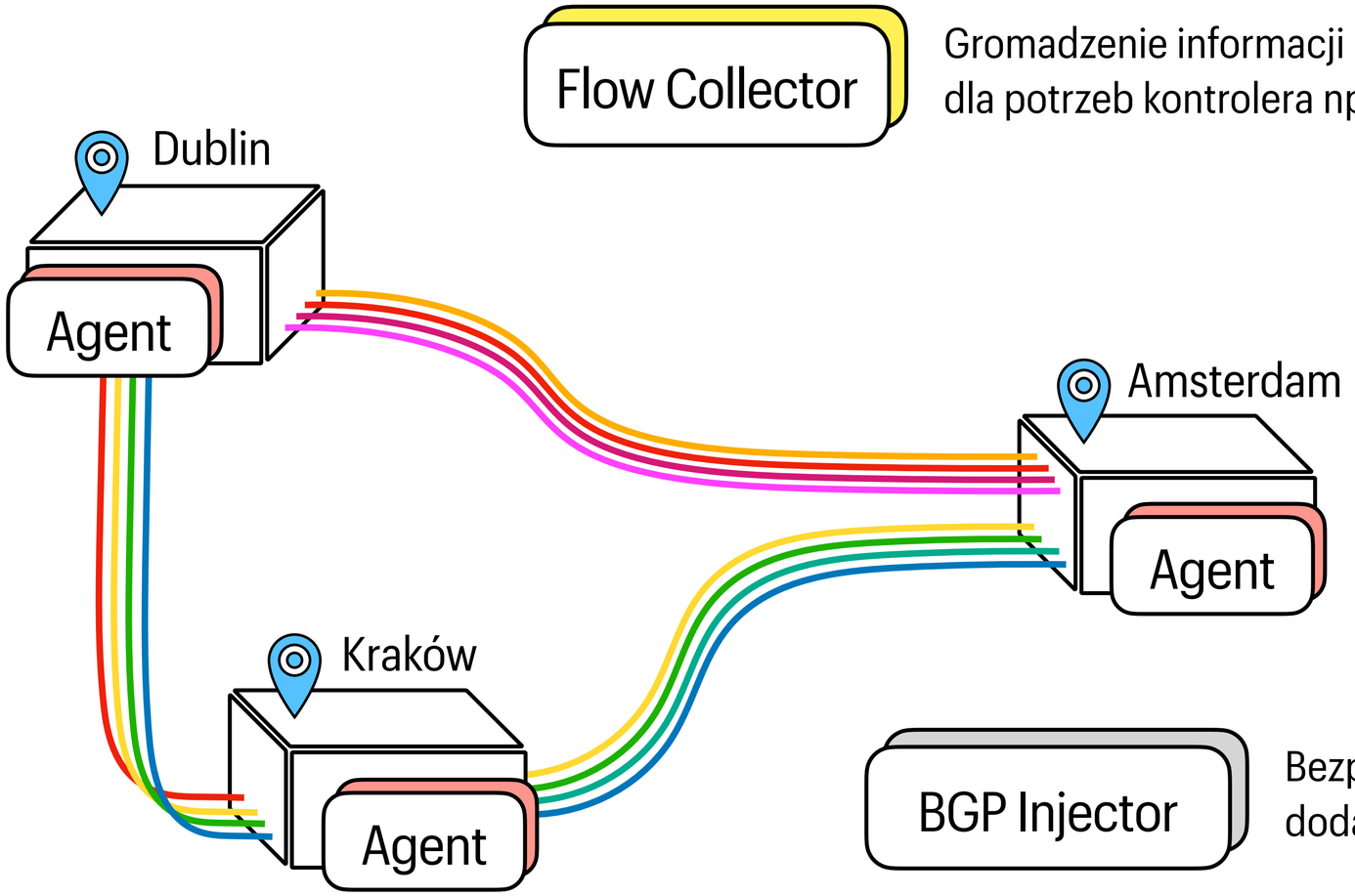
A CO JAK TRZEBA RUCH PRZESŁAĆ DO INNEGO DC?



🤔 Dlaczego rozproszony traffic engineering tutaj nie zadziała?

O HIGHWAY'ACH WSPÓŁCZESNEGO INTERNETU

CZY TO JUŻ SD-WAN?



Flow Collector

Gromadzenie informacji o przepływach w sieci
dla potrzeb kontrolera np. oblicza macierz ruchu

Controller

Obliczanie ścieżek na podstawie
stanu sieci, komunikacja z agentami

BGP Injector

Bezpośrednie sterowanie ruchem poprzez
dodawanie i wycofywanie prefiksów

AI, Big Data

NIKITA MURAVYTSKYI AGH

O ULTIMA THULE WSPÓŁCZESNYCH DC

WYZWANIA STAWIANE PRZEZ BIG DATA I AI

Chcemy wprowadzić rozróżnienie jakimi zasobami dysponujemy

cluster - jednostka logicznego podziału w DC

Niektóre zadania jest lepiej liczyć na dedykowanym sprzęcie

- to psuje nam abstrakcję stworzoną przez wirtualizację;

Dużo ciężkich przepływów, przewidywalne zachowanie

- stawiane są zupełnie inne wymagania wobec sieci;

- stawiane są zupełnie inne wymagania wobec fizycznej budowy DC.

Nie chcemy żeby przepływy z np. AI obciążały szkielet sieci

🤔 Dlaczego topologia Clos'a nie wystarcza na clustry AI?

Trendy rozwoju:

- alternatywny stack sieciowy - Infiniband, Nvidia NVLink;

- coraz większa gęstość GPU per serwer;

- GPU i storage w jednym rozwiązaniu;

- w przypadku AI podział na dwa układy - uczenie i inferencja.

Automatyzacja

NIKITA MURAVYTSKYI AGH

O ROBOCIE ŚMIECIARCE WALL-E

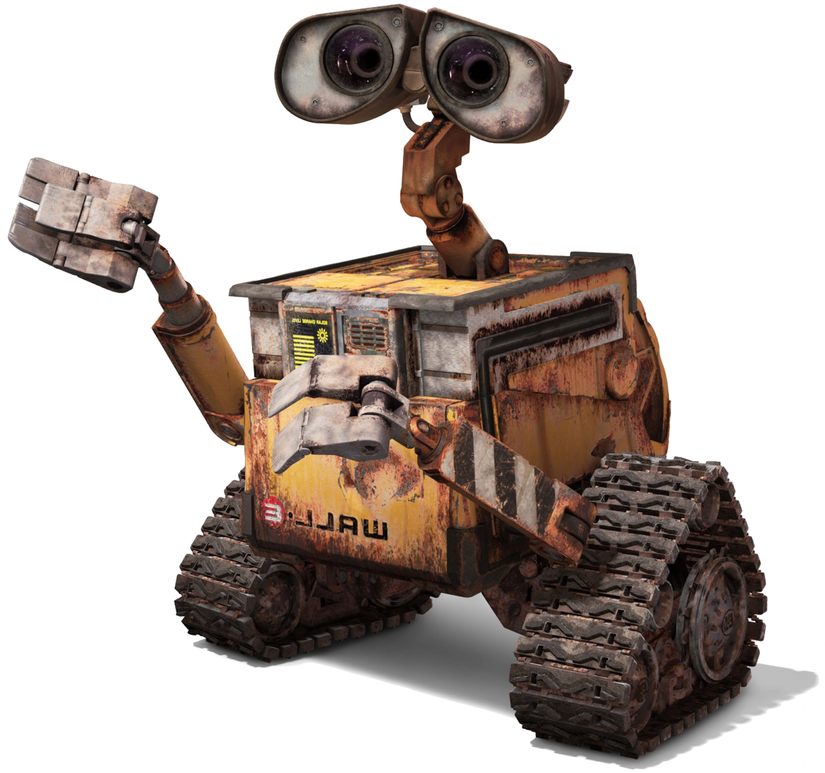
FILOZOFIA AUTOMATYZACJI DC

Skomplikowany system rozwiązujący **jeden problem**

Przy dużej skali **automatyzować** trzeba **wszystko**

Oceniamy potencjalne **skutki** automatyzacji

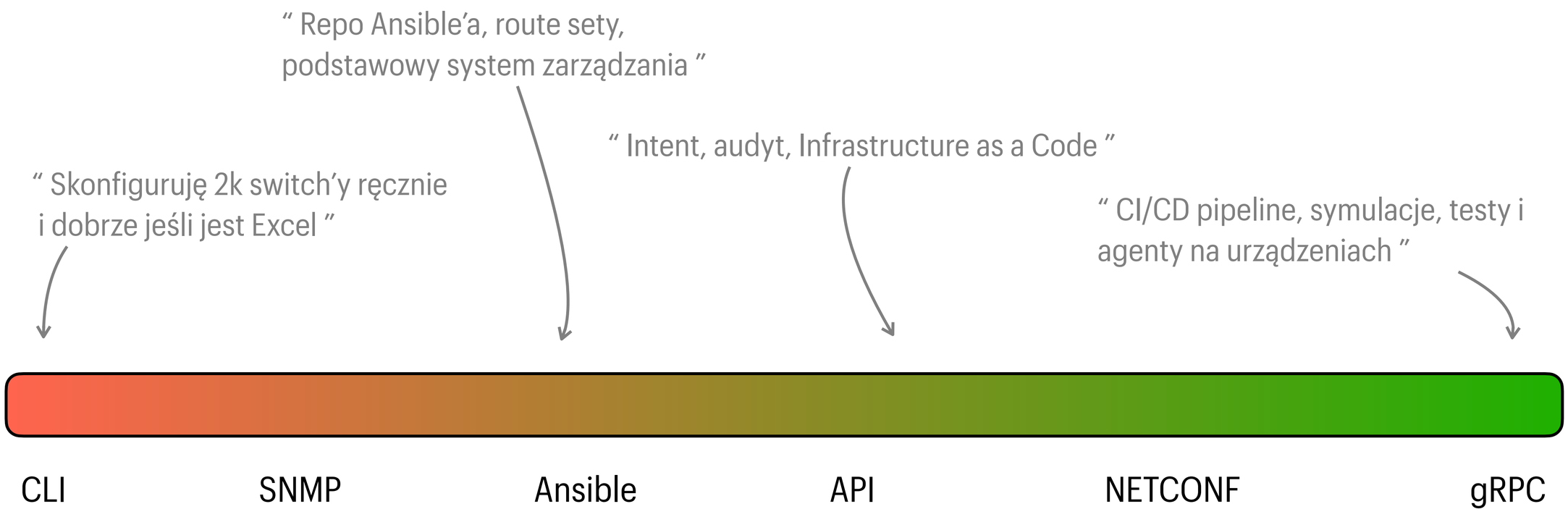
Obsługujemy przypadki skrajne?



NIKITA MURAVYTSKYI AGH

O ROBOCIE ŚMIECIARCE WALL-E

POZIOMY AUTOMATYZACJI



Protokoły interakcji z urządzeniami sieciowymi

O ROBOCIE ŚMIECIARCE WALL-E

ZERO TOUCH PROVISIONING

1 Przyjeżdża szafa do DC

Uzupełniamy bazę danych
o specyfikację urządzeń

2 Podpięcie zasilania i podłączenie
do reszty sieci

DHCP i LLDP urządzenia i
weryfikują topologię

3 Provisioning przełącznika

Konfiguracja w zależności od
- lokalizacja
- roli
- wersji

4 Provisioning serwera

- instalacja OS
- instalacja aplikacji
- test
- dołączenie do puli zasobów

Chłodzenie

NIKITA MURAVYTSKYI AGH

O TYM ŻE CHŁODZENIE TO RACZEJ O WODĘ NIŻ O POWIETRZE

CO SPOCZYWA NA DNIE OCEANU?

W 2018 Microsoft testował swoje pierwsze podwodne centrum danych

Bezpośrednie połączenie do wiatraków i kabli podwodnych

Azot zamiast tlenu

12 szaf, 864 serwery

O TYM ŻE CHŁODZENIE TO RACZEJ O WODĘ NIŻ O POWIETRZE

NIE SERWERY DO MORZA TYLKO MORZE DO SERWERÓW

Serwery są zanurzane w nieszkodliwej dla sprzętu cieczy

Większa gęstość zasobów

Wydajniejsze chłodzenie

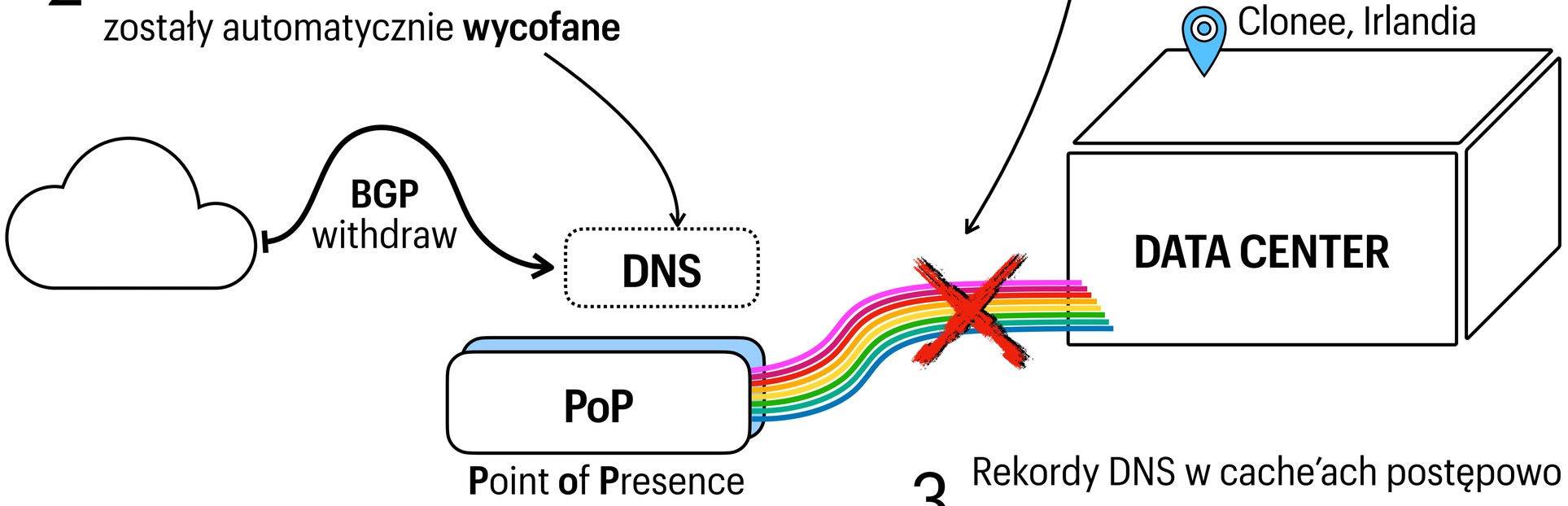
Dłuższy czas życia sprzętu?

O TYM JAK NIE ZOSTAWIAĆ KLUCZYKÓW W AUCIE

CO FACEBOOK ZROBIŁ NIE TAK?

1 Pomyłkowe wprowadzenie stanu **maintenance** wszystkich plane'ów na raz sprawiło, że sieć szkieletowa przestała działać

2 Wykrywszy niedostępność DC, wpisy **BGP** rozgłaszające serwery DNS zostały automatycznie **wycofane**



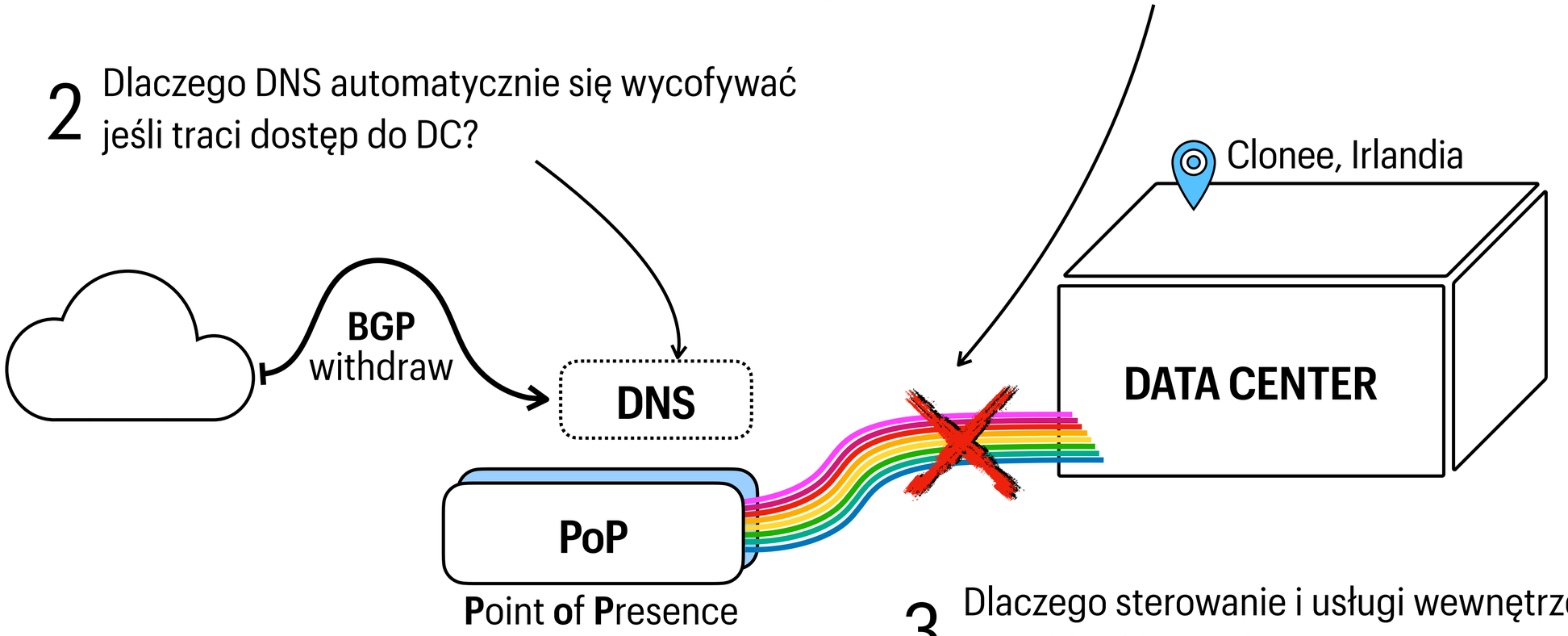
3 Rekordy DNS w cache'ach postępowo znikały, odcinając kolejne usługi - również **wewnętrzne**

O TYM JAK NIE ZOSTAWIAĆ KLUCZYKÓW W AUCIE

CO FACEBOOK ZROBIŁ NIE TAK?

1 SD-WAN injector wycofał prefiksy ze **wszystkich** plane'ów.
Dlaczego to było możliwe?

2 Dlaczego DNS automatycznie się wycofywać
jeśli traci dostęp do DC?



3 Dlaczego sterowanie i usługi wewnętrze
współdzieliły zasoby sieciowe?

Dziękuję za uwagę!

O TYM CO SIĘ NIE ZMIEŚCIŁO NA SLAJDACH

ABANDON ALL HOPE, YE WHO ENTER HERE

Alternatywne topologie DC

- Application of Butterfly Clos-Network in Network-on-Chip
- Flattened butterfly: a cost-efficient topology for high-radix networks
- Dragonfly+: Low Cost Topology for Scaling Datacenters
- Jellyfish: Networking Data Centers Randomly

Inne

- Load balancing, what you should really know
- Wiring the Planet: Scaling Meta's Global Optical Network
- Jak przełączniki optyczne pozwalają lepiej skalować DC w Google
- RIFT - alternatywa BGP w DC