



Przełączniki i Routery

Co jest ważne i dlaczego



Łukasz Bromirski
lbromirski@cisco.com



Rafał Szarecki
rafal@juniper.net

PLNOG, Warszawa, marzec 2011

Zawartość (z grubsza)*

- Przełącznik Ethernet
- Router IP
- Rozważania o architekturze
- Q&A

* agenda może ulec zmianie bez ostrzeżenia, nawet w trakcie prezentacji

Przełącznik Ethernet czyli ballada o duplexach

Port dla użytkownika?

- Prędkość pracy – zakładamy symetryczną:
10Mbit/s, 100Mbit/s, 1Gbit/s, 10Gbit/s, 40Gbit/s, 100Gbit/s
- Duplex – czy port pracuje w obie strony jednocześnie czy wymaga pracy naprzemiennej
 - half – RX i TX naprzemiennie
 - full – RX i TX jednocześnie

Co łączy porty ze sobą?

- Porty fizyczne ze sobą łączą układy realizujące switching/routing Ethernetu/IP – jest ich na płycie jeden lub więcej
- Rzadko można już spotkać w sprzedaży urządzenia realizujące switching Ethernetu na zwykłym, ogólnodostępnym CPU (x86 386/486, PowerPC) – najtańsze rozwiązania bardzo często jednak tak właśnie wyglądały

...wolno ale działa!

Matryca przełączająca –
co to w ogóle jest?!

Matryca przełączająca

- Element realizujący obsługę przesyłania ruchu użytkownika pomiędzy wejściami i wyjściami z bardzo dużą wydajnością

Brak – najprostsze konstrukcje

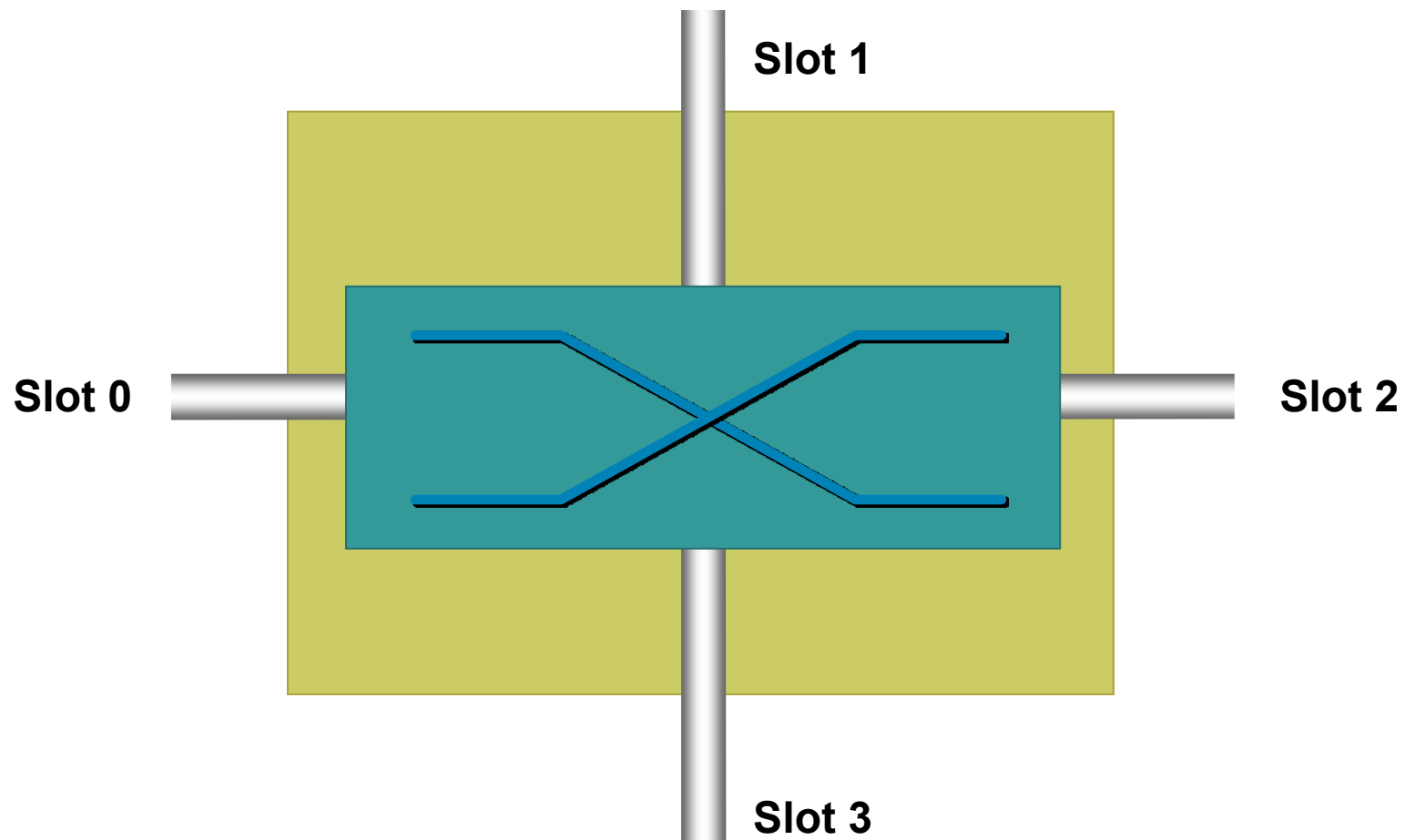
Centralna – małe przełączniki i część większych

Centralna, ale redundantna – urządzenia pracujące z podstawowym i zapasowym modułem przełączającym, zwykle przełączniki modułarne

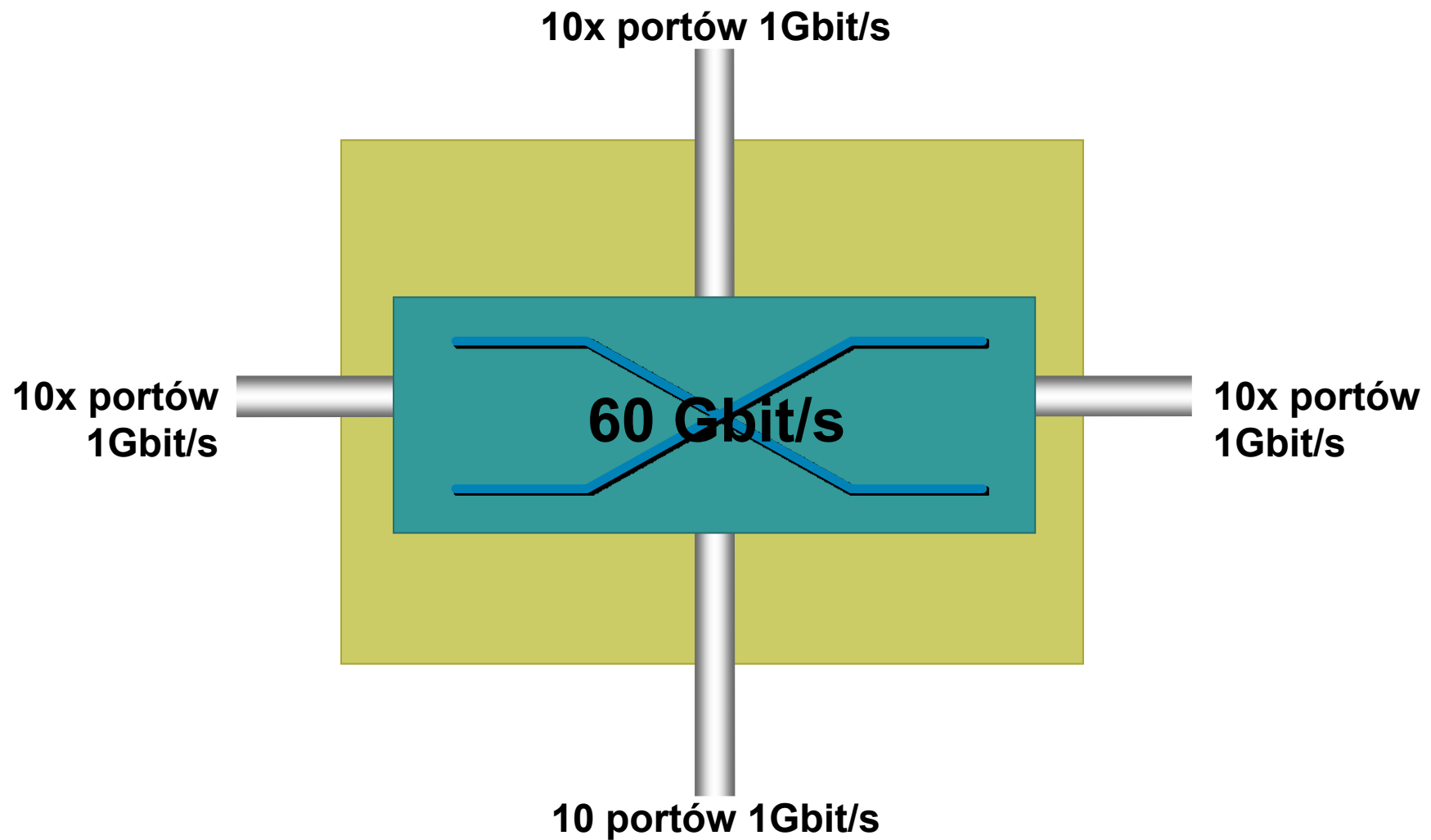
Nadmiarowa – $N+1$, gdzie $N > 1$ – zapewnia zapas przepustowości bez utraty wydajności w przypadku utraty jednej z matryc (np. $2+1$, $3+1$)

Matryca przełączająca

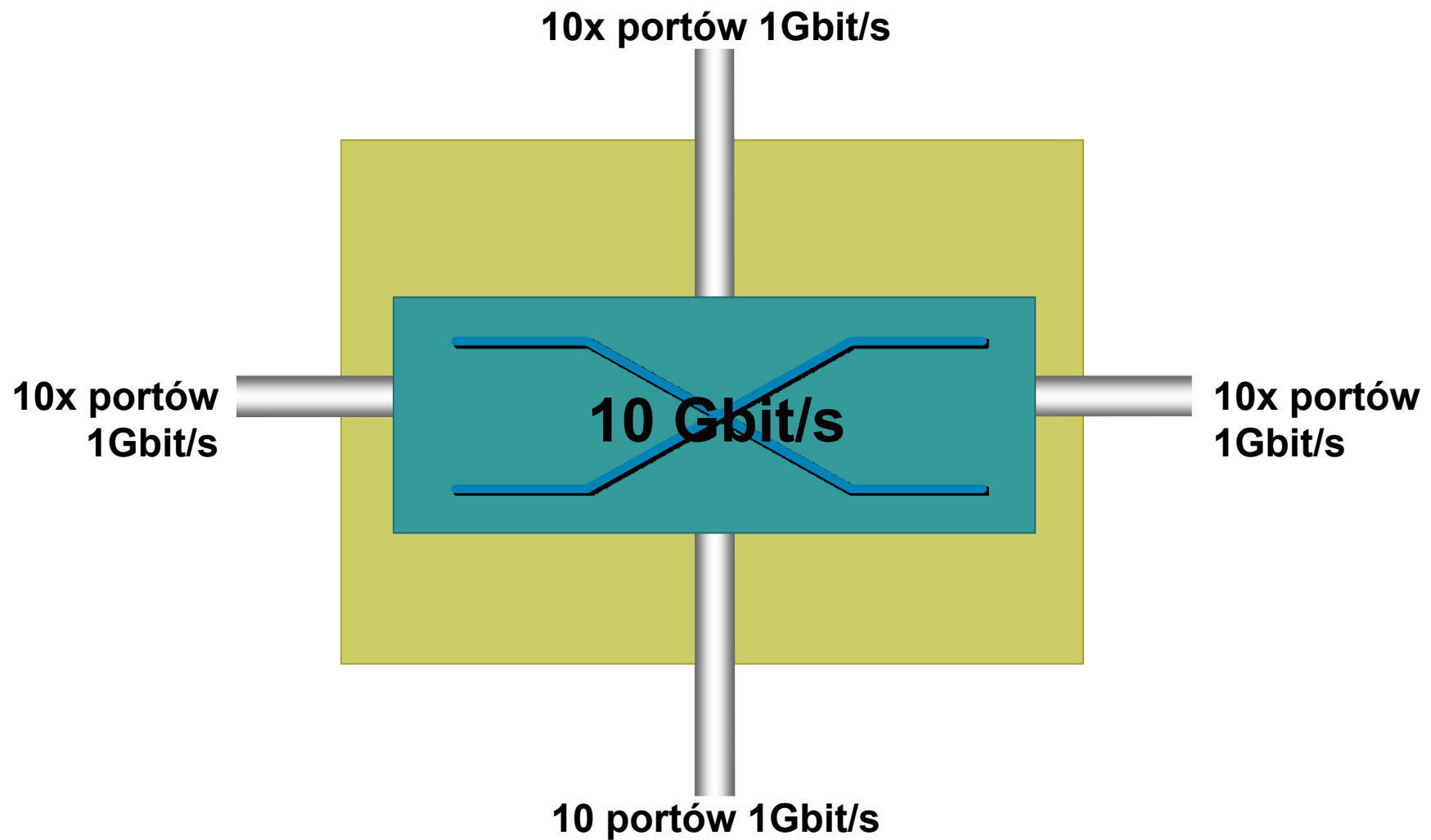
Połączenie pomiędzy portami/slotami w przełączniku



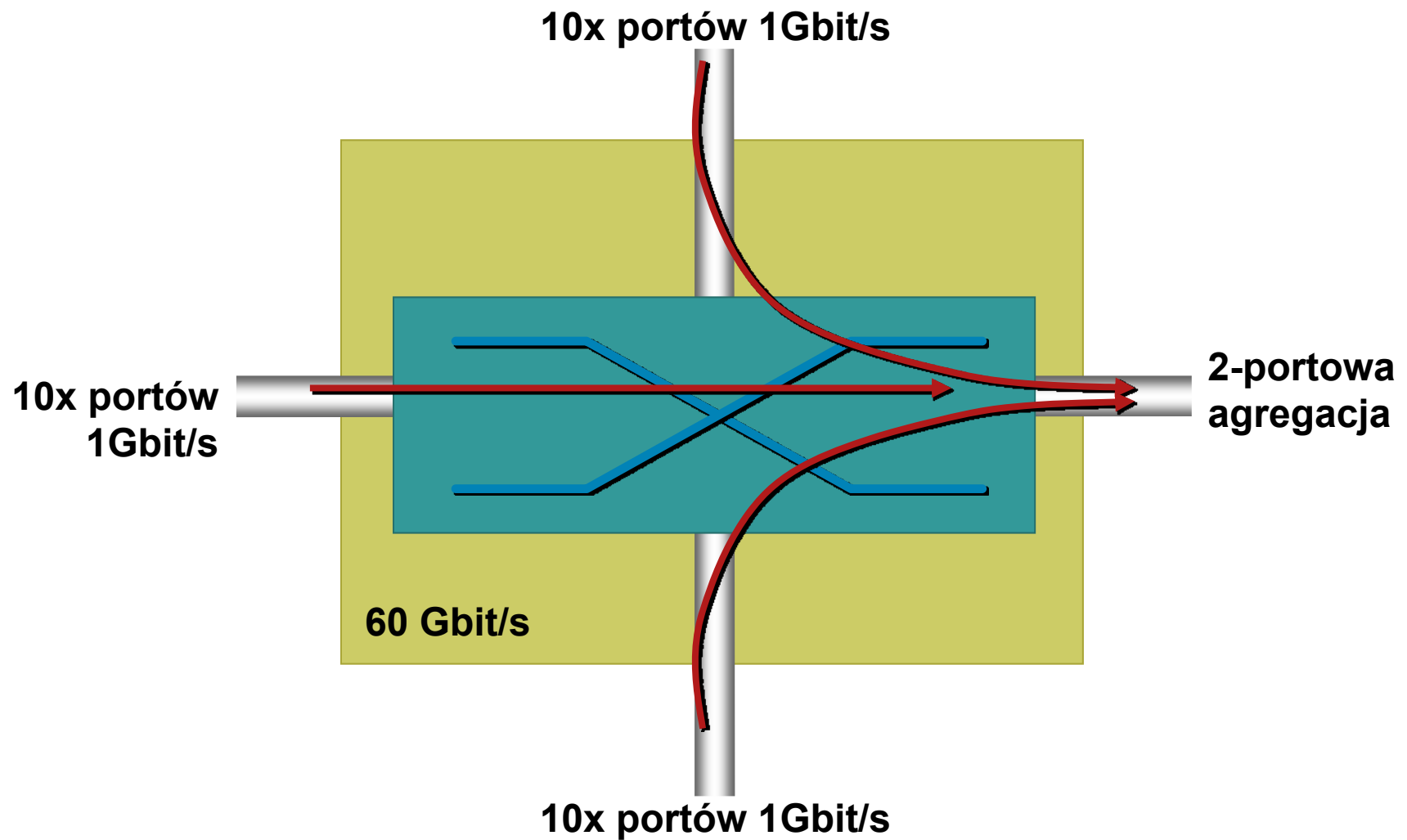
Nieblokująca matryca przełączająca



Blokująca matryca przełączająca

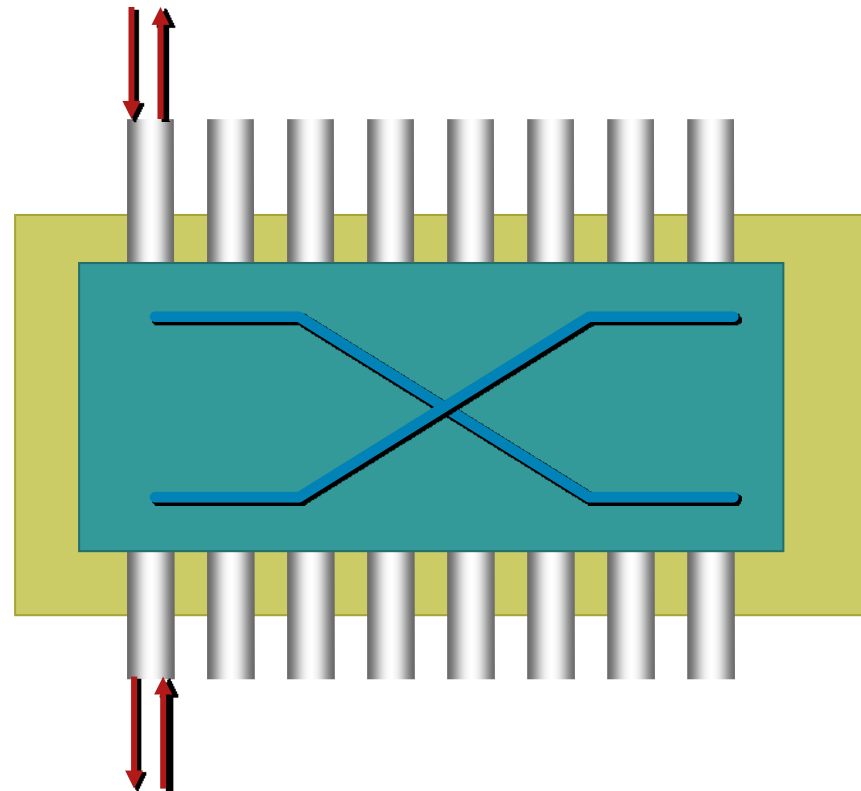


Jak ruch rozkłada się w sieciach?



Zrozumieć 'przepustowość'

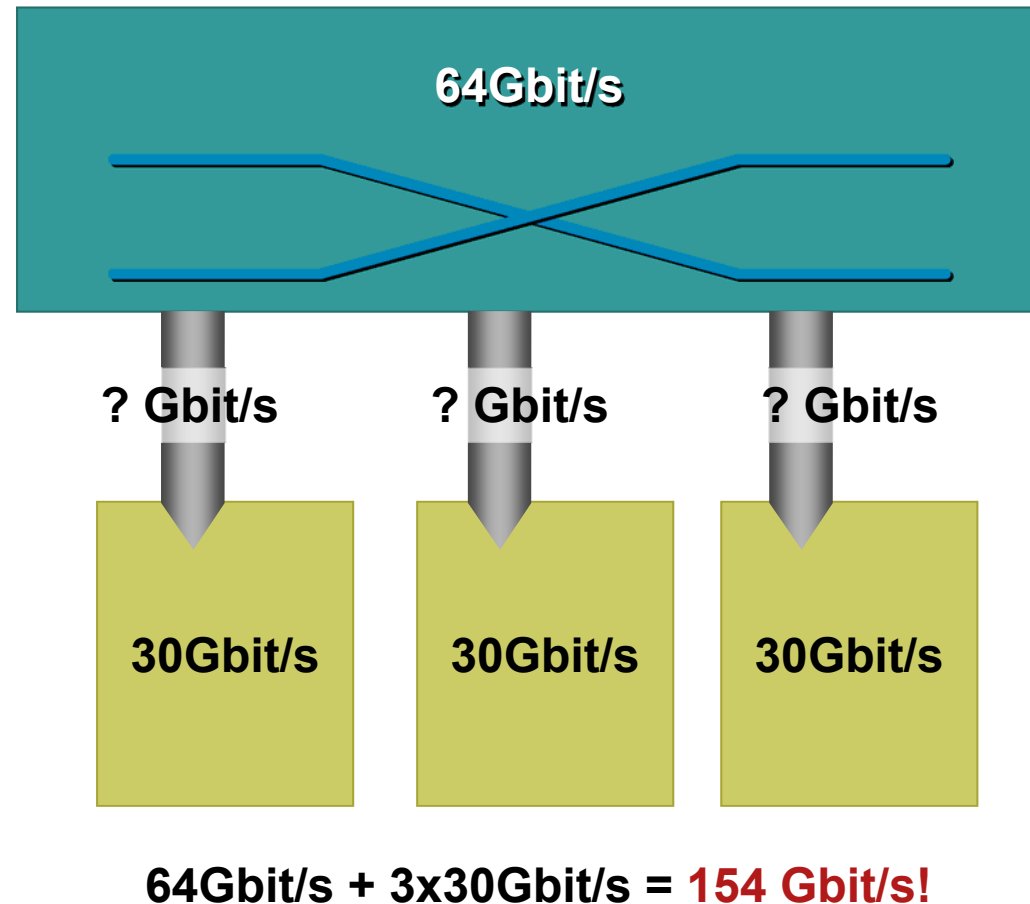
- Przepustowość to możliwości przesyłania ruchu przez matrycę przełączającą – wyrażane w Gbit/s
- Każdy z „interfejsów” matrycy jest w stanie zrealizować ruch nadający i odbierający – dla połączeń 1Gbit/s oznacza to wydajność na poziomie 2Gbit/s
- Wszystkie tego typu wydajności podawane są zwykle jako full-dupleks



48x 1GE + 2x10GE = 48Gbit/s + 20Gbit/s
Przepustowość: 68Gbit/s FD (136Gbit/s HD)

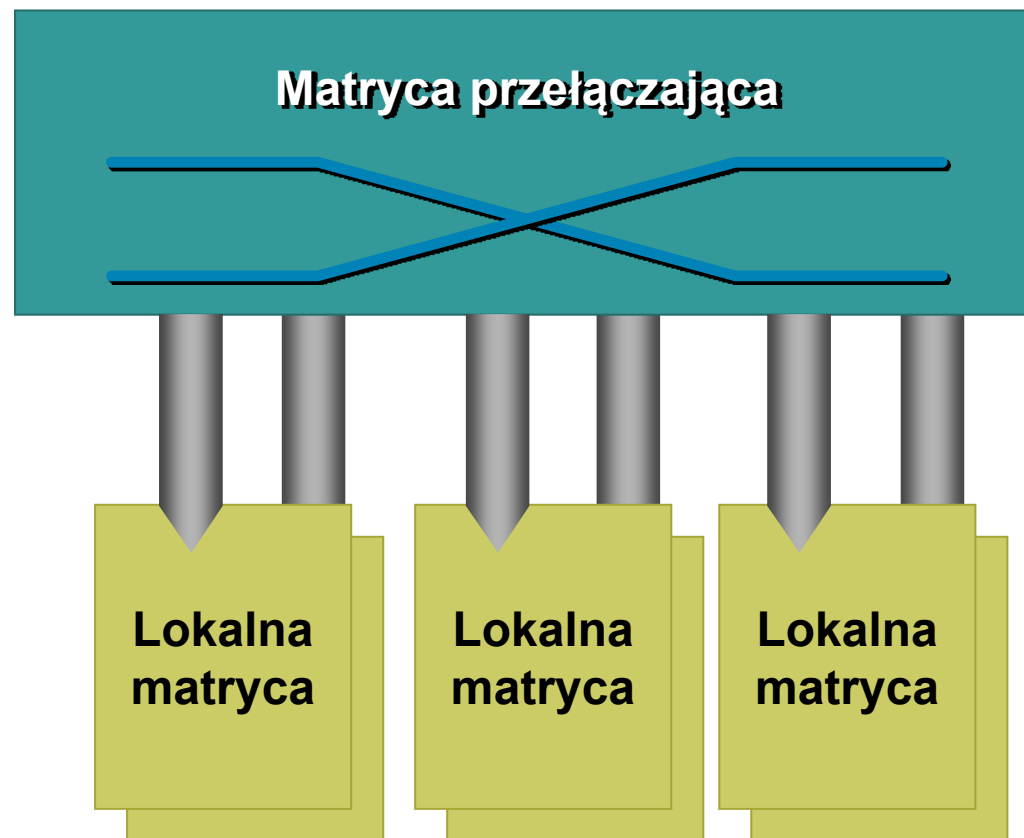
'Marketingowa' przepustowość...

- Lokalne przetwarzanie pakietów powoduje, że w teorii cały system ma większą wydajność
- Jednak nawet pojedynczy strumień ruchowy może zablokować całą matrycę jeśli nie jest ona odpowiednio wydajna
- Dodawanie/doliczanie wydajności poszczególnych kart pozwala osiągnąć atrakcyjnie wyglądające, ale mało rzeczywiste (zwykle) wartości



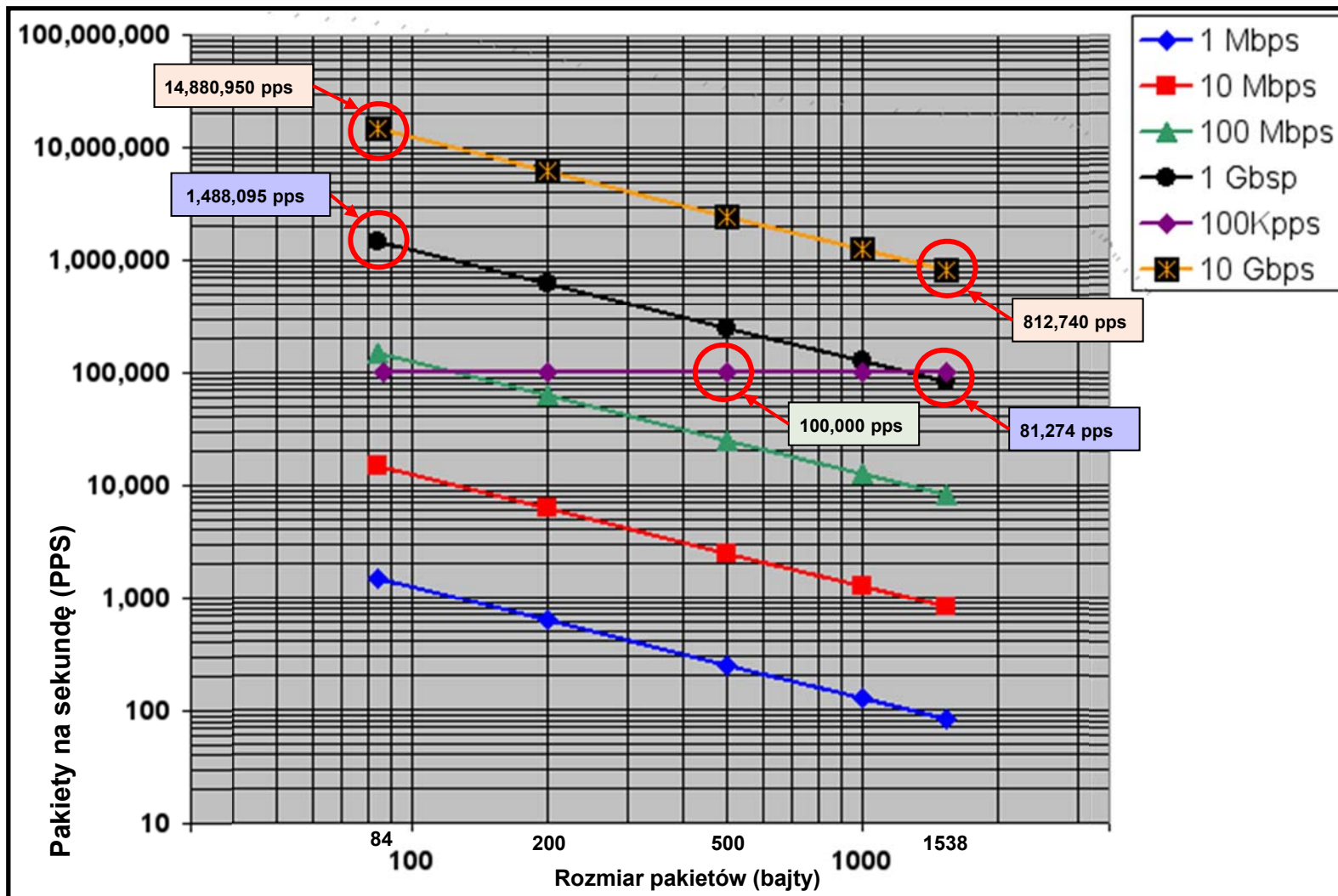
A wydajność?

- Ponieważ przepustowość daje się łatwo „zmanipulować”, wskaźniki wydajności mogą być bardziej rzeczywiste
- Wydajność to najwyższa częstotliwość z jaką można wysłać pakiety bez utraty części z nich
- Wydajność wskazuje na sprawność układu FE (forwarding engine)
- Wydajność mierzy się w pakietach na sekundę



Prosty test: wydajność x 1000 \geq przepustowość HD

Forwarding engine – wydajność?

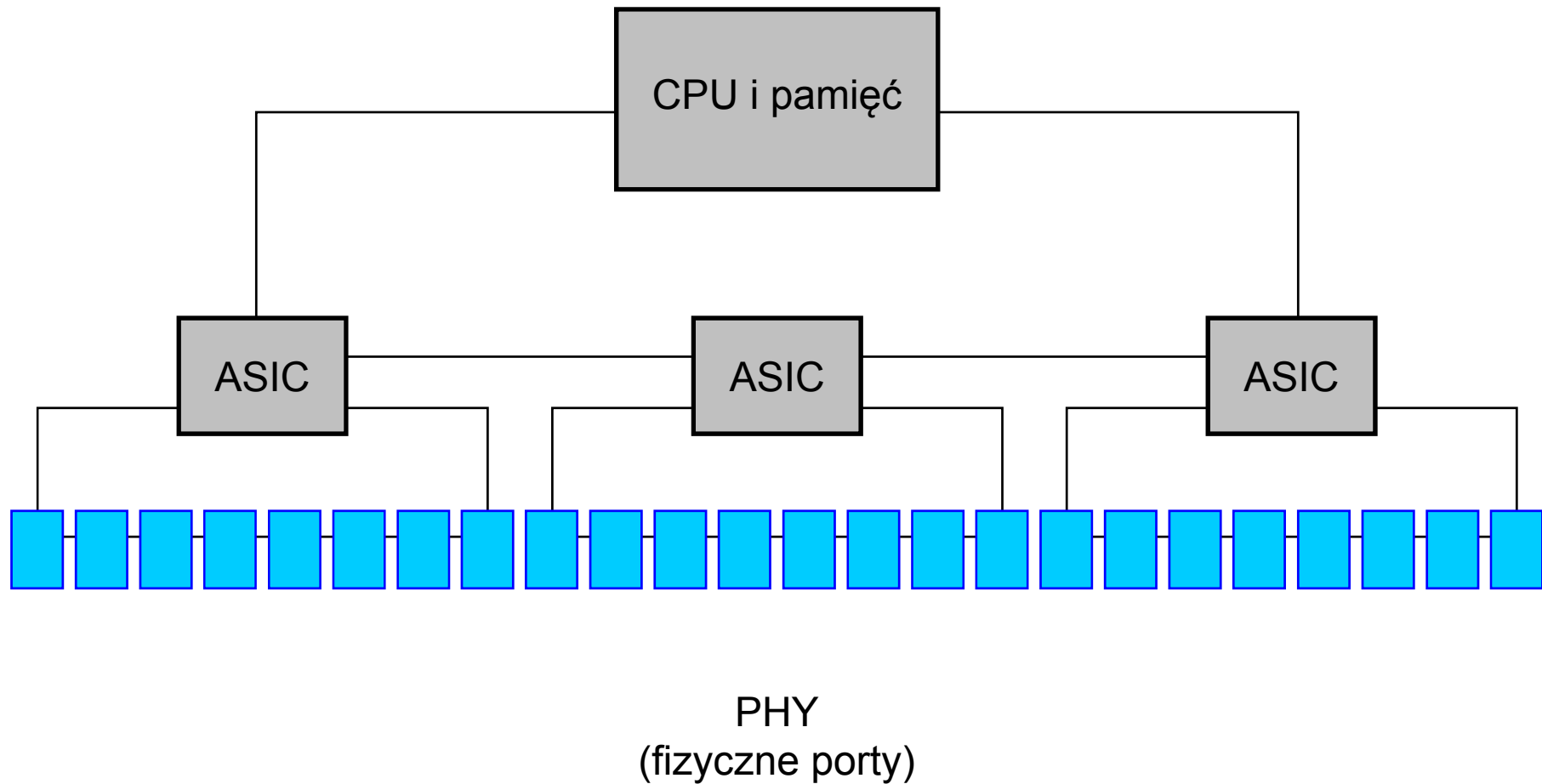


Switch matrix? Forwarding engine?

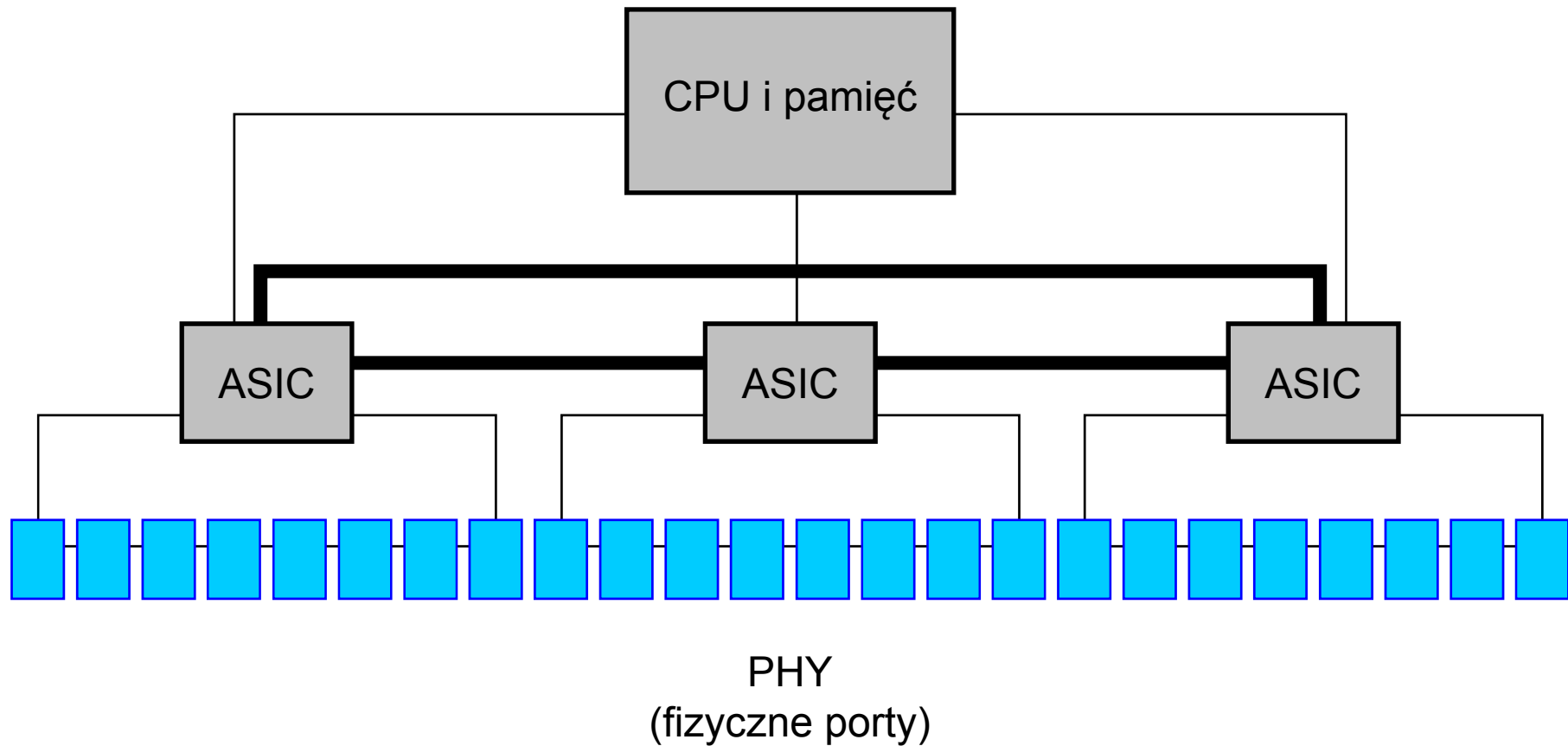
- Switch matrix – matryca przełączająca – ma za zadanie transportować ruch między portem wejściowym a portem (portami) wyjściowymi
- Forwarding engine – ma zidentyfikować i podjąć skonfigurowane akcje na ruchu otrzymanym z portu wejściowego:
 - wybór ścieżki przez matrycę
 - sklasyfikować
 - odfiltrować
 - nałożyć politykę QoS
 - dodatkowo otagować lub zdjąć tag
 - zrealizować forwarding (L2) lub switching/routing (L3)
 - ...

Jak wygląda przełącznik
Ethernet?

Jak wygląda przełącznik?



Jak wygląda przełącznik?

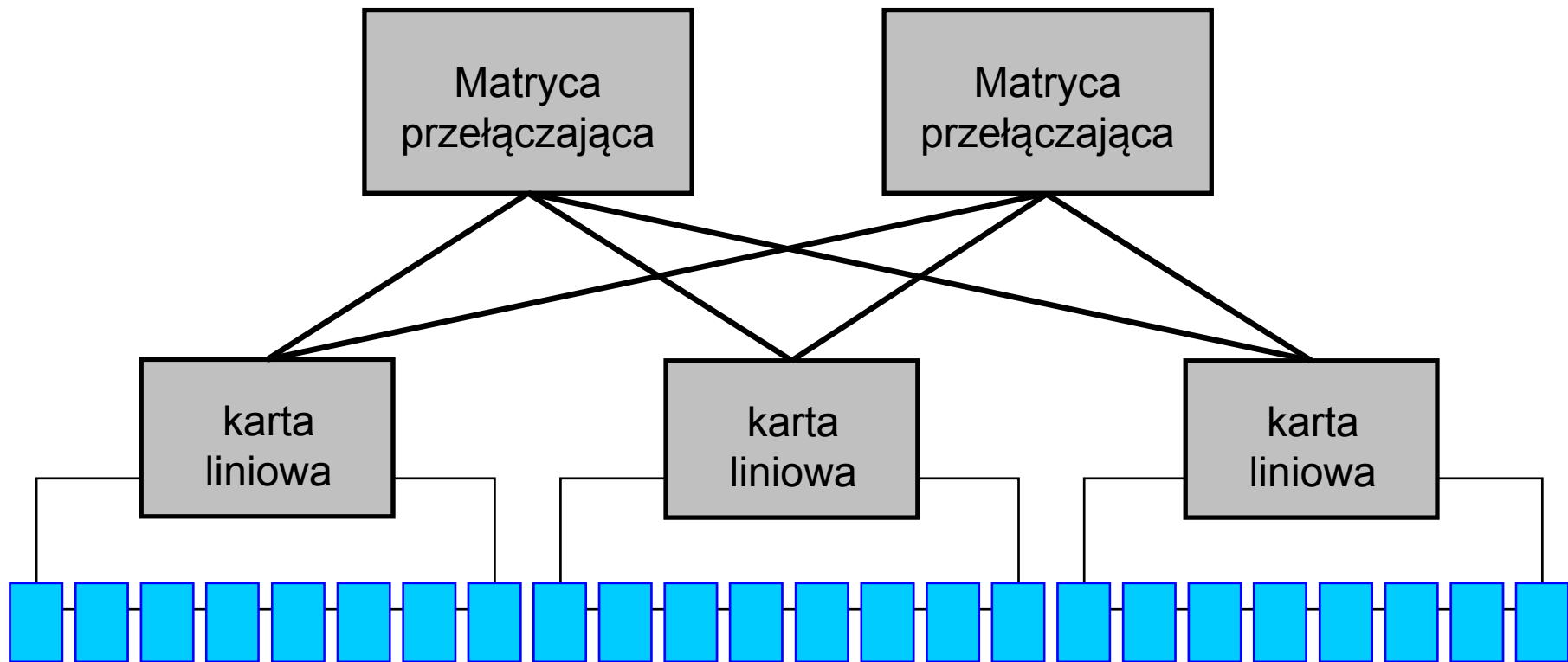


Tutaj matryca – bezpośrednie połączenia full mesh pomiędzy Forwarding Enginami

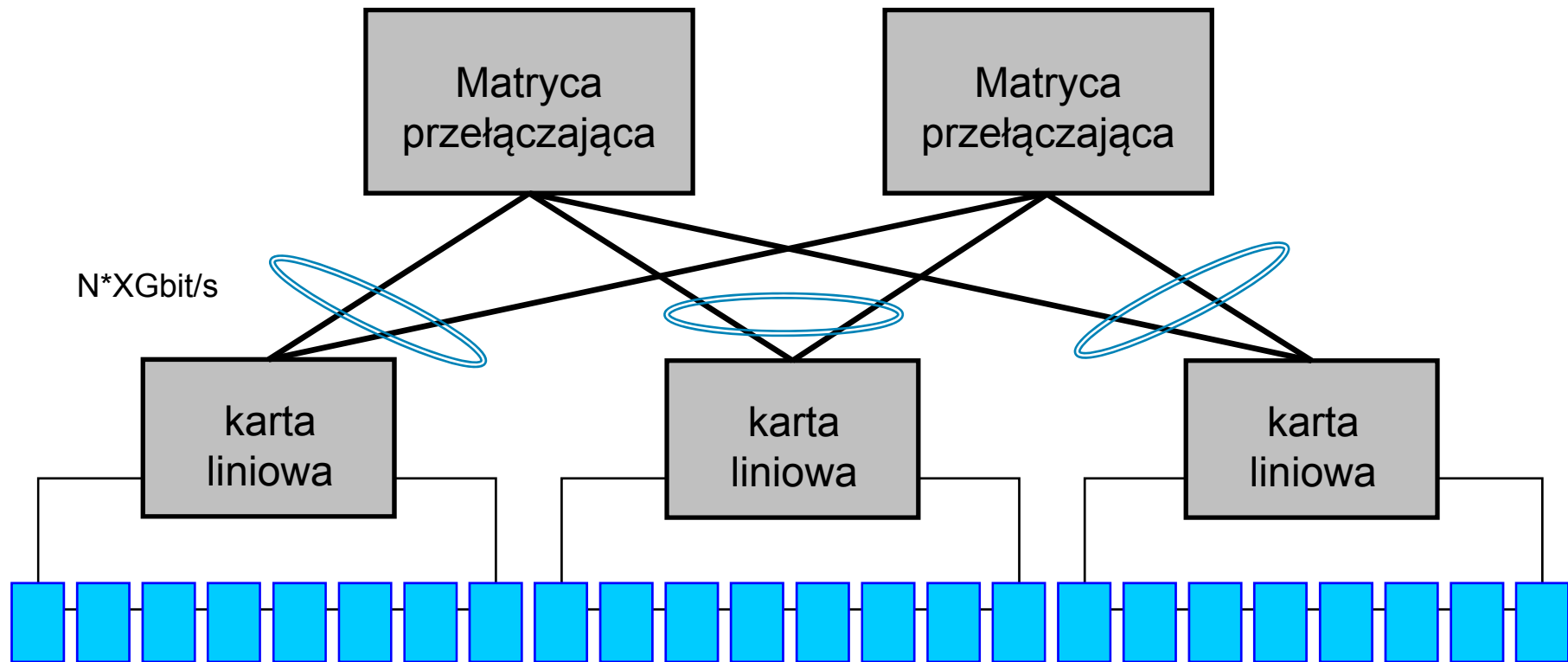
„Stakowalne” przełączniki dostępowe?

- Matryca przełączająca przełącznika jest rozsunięta na połączenia/ring pomiędzy przełącznikami
- Teoretycznie powinna zatem zapewniać skalowalność bez zmian w funkcjonalności
 - oprócz wydajności na ringu pomiędzy przełącznikami
 - wiele FE – matryca to już nie full mesh
- Część rozwiązań ‘stakowalnych’ przełączników nie rozsuwa matrycy przełączającej w stos, a same matryce przełączające połączone są zwykłymi portami (inaczej zaterminowanymi)
 - Spanning Tree blokuje jeden z portów (ring) lub buduje się automatycznie agregacja portów

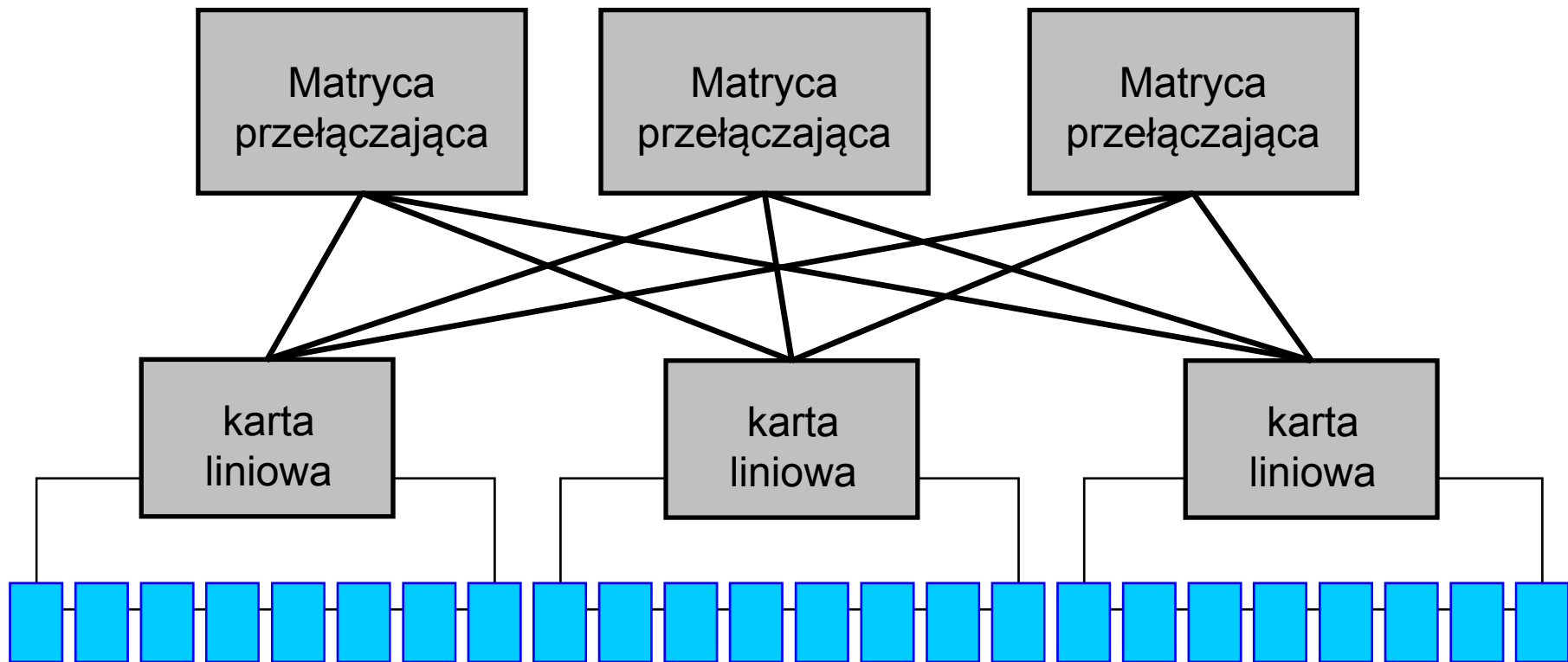
„Tradycyjny” przełącznik modułarny



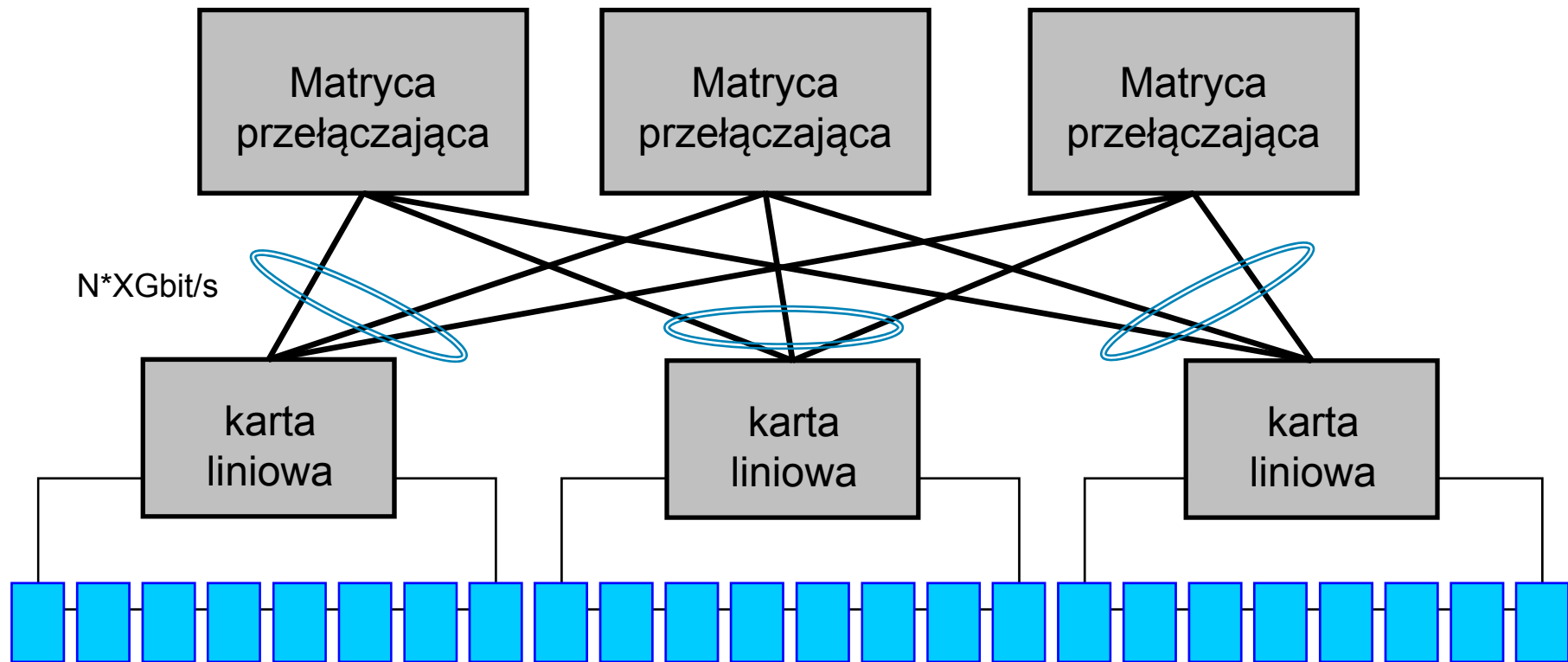
„Tradycyjny” przełącznik modułarny



„Większy” przełącznik modułarny



„Większy” przełącznik modułarny



Architektury routerów

Czym jest router IP?

- Urządzenie realizujące przekazywanie ruchu na podstawie docelowego adresu IP

RIB (Routing Information Base) tworzony jest przez protokoły routingu

Z RIB tworzony jest FIB, FIB wykorzystywany jest w procesie routingu

IPv4 – do 32 ,pasujących’ wpisów, dla IPv6 – 64/128

FIB musi być zoptymalizowany do szukania najdokładniejszego wpisu – rozwiązania producentów (J-Tree, CEF)

Możliwe inne tryby działania – switching Ethernetu, MPLS, OTN, ATM, FR

- ...ale to nie wszystko

Czym jest router IP?

- Router jest niezależny od technologii L2 – musi wspierać wiele rodzajów interfejsów

Ostatnie lata pokazują, że pojawiła się pod-klasa routerów zoptymalizowanych dla Ethernet – MX, ASR 9k, 7750

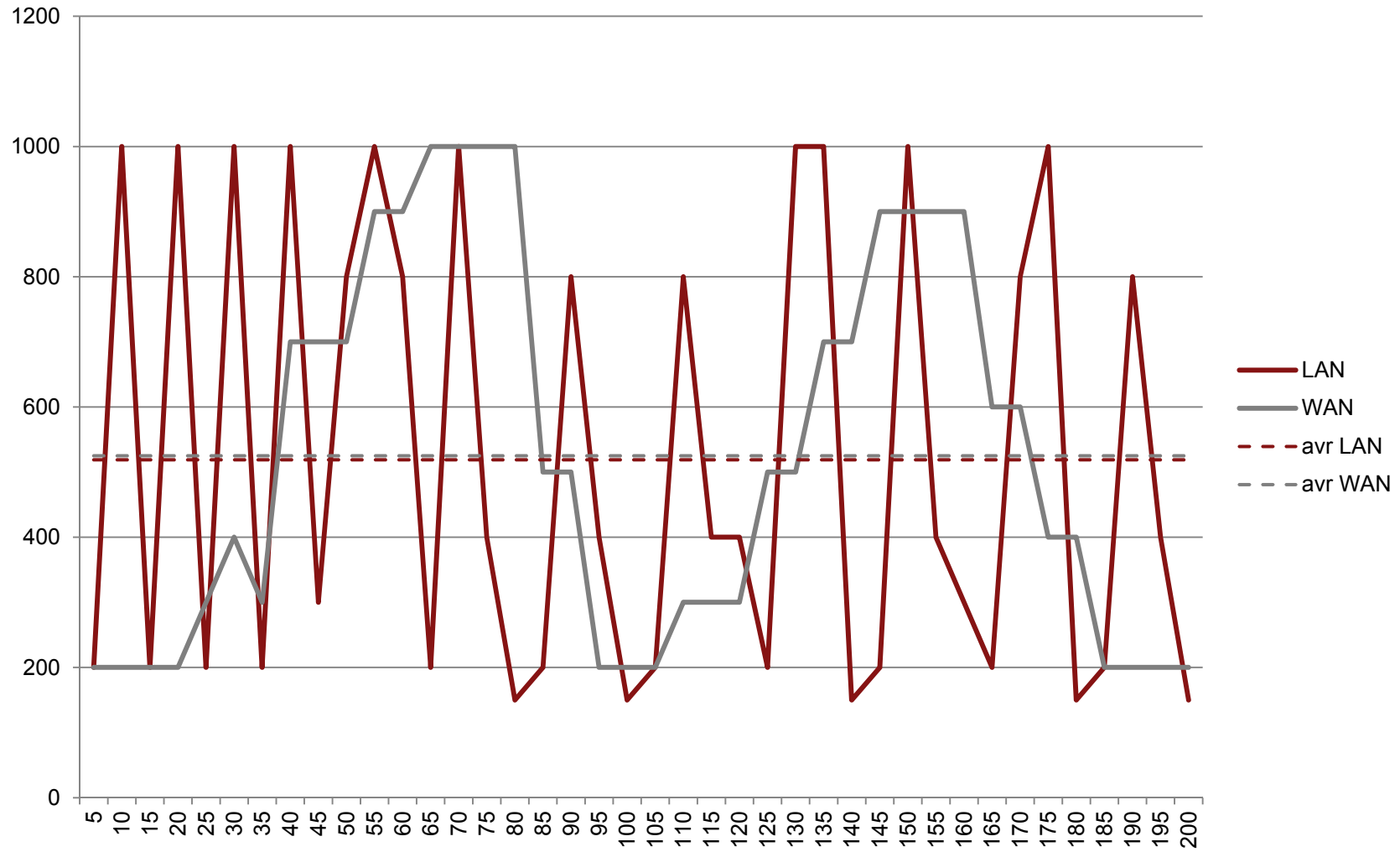
Routery te wspierają zwykle nie-ethernetowe interfejsy, ale są to rozwiązania relatywnie drogie w stosunku do przepustowości oferowanych przez interfejsy Ethernet

- Router pracuje zwykle na brzegu sieci, w WAN – wymaga mniej lub bardziej zaawansowanych mechanizmów QoS

Środowiska – WAN i LAN

WAN	LAN
Droga transmisja <ul style="list-style-type: none">• Laser• DWDM• Złożone – OPEX	Tania transmisja <ul style="list-style-type: none">• MMF• Małe odległości
Kompromis pomiędzy przepustowością a kosztami.	Łatwo można zaprojektować sieć przewymiarowaną <ul style="list-style-type: none">• W każdym ułamku sekundy (np. 10ms)• W przypadku awarii (STP ma wbudowane)
Duża ilość ruchu i sesji – dziesiątki i setki tysięcy sesji	Mniejsza ilość hostów i sesji – setki-tysiące hostów.
Zauważalne opóźnienie <ul style="list-style-type: none">• Utrata pakietów zauważalna po czasie RTT.• BDP do wzięcia pod uwagę	Pomijalne opóźnienie <ul style="list-style-type: none">• Utrata pakietów od razu zauważalna dla nadawcy• Mały BDP
Potencjalnie duże i skomplikowane wymagania dotyczące QoS	Minimalne bufory
Większa maksymalna częstotliwość pakietów, mniejszy średni rozmiar pakietów	Mniejsza maksymalna częstotliwość pakietów, większy rozmiar pakietów (w tym często ruch ramek jumbo)

Charakterystyka ruchu



IP router vs. L3 switch i LSR

	IP Router	L3 switch	LSR
Decyzja w oparciu o sprawdzenie IP	TAK	TAK	NIE
Niezależny od L2	TAK	NIE	TAK
Bufory wyjściowe dla QoS	~50ms lub więcej	~5ms	~50ms lub więcej
Poziomy priorytetów QoS	3-5	2-3	3-5
Zagnieżdżone klasy usługowe QoS	TAK	NIE	TAK
RED	TAK	NIE	TAK
CIR/PIR dla klasy, dodatkowe wagi/etc.	TAK	NIE	TAK

„Router klasy operatorskiej” – co to jest?

- Przewidywalna wydajność i skalowalność

Control i Management plane

Data plane

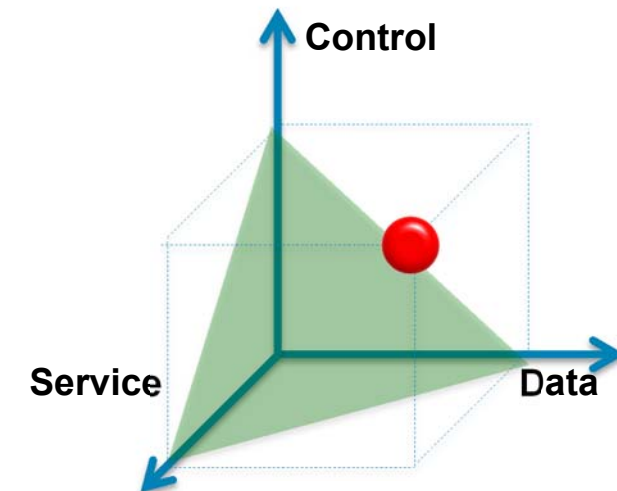
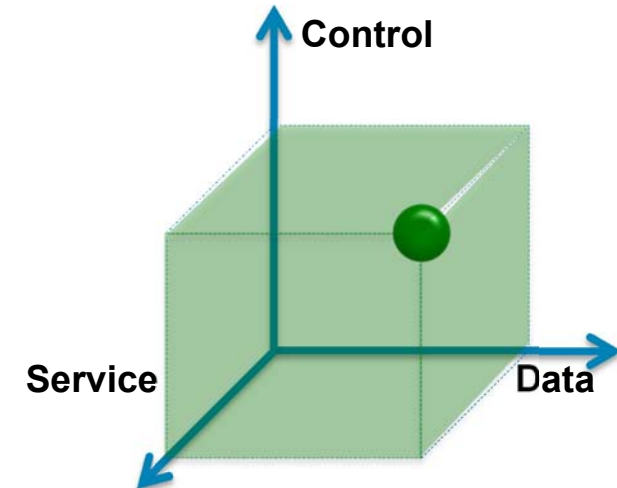
Service plane

Przewidywalna != nieograniczona lub nieskończenie rozbudowywalna

- Wykrywanie i „naprawa” awarii

Sprzęt jest podstawą

Oprogramowanie jest krytyczne do wykorzystania możliwości sprzętu

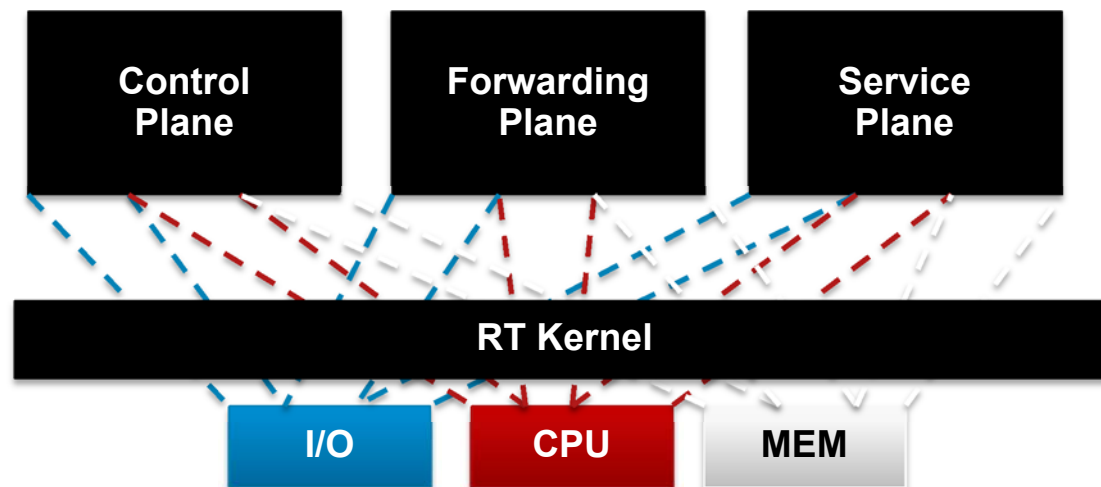


Przewidywalny router

- Dedykowane zasoby dla każdego z obszarów działania routera.
 - CPU – cykle CPU
 - NPU, ASIC
 - Pamięć operacyjna, buforów, etc.
- Przewidywalny != ASIC
- Przewidywalna wydajność == NPU/ASIC
- Przewidywalna wydajność przy bardzo dużych przepustowościach == ASIC

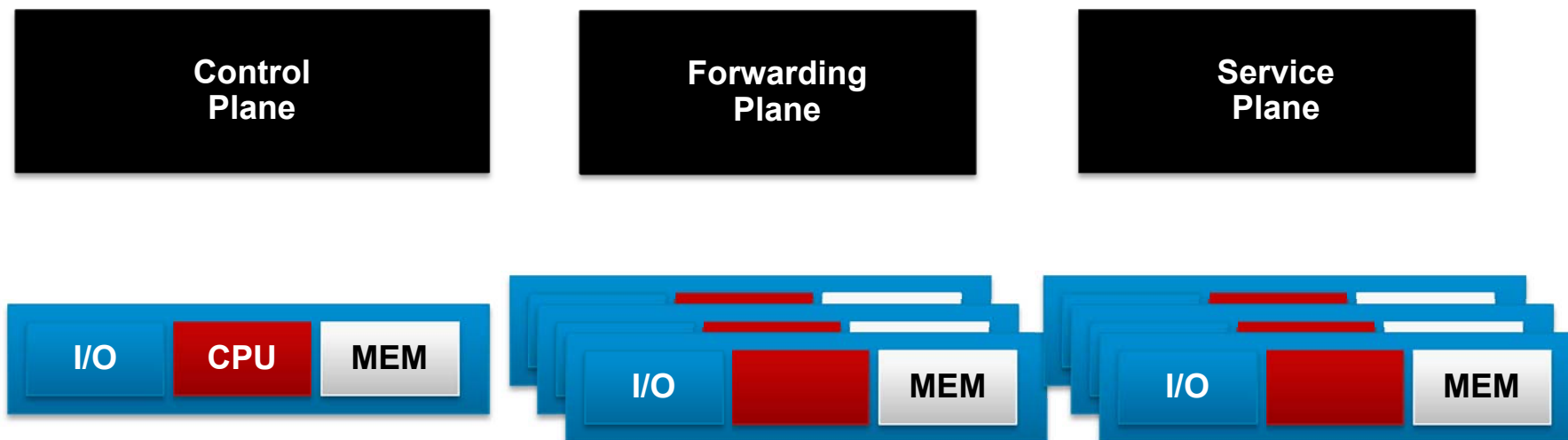
Przewidywalny router programowy?

- Router programowy – jeden kompleks wykonujący wszystkie funkcje
- Po co? Do zastosowań typu CPE, mały brzeg, agregacja
- Tańszy, ale ograniczony wydajnościowo
- Routery serii J, Cisco ISR G1/G2 oraz 7200



Przewidywalny router sprzętowy?

- Dedykowany kompleks sprzętowy do obsługi pakietów
- Zwykle dystrybuowany – wiele elementów realizujących te same funkcje rozproszonych w architekturze routera
- Przykłady - ASR 1000, M10i/7i, MX80, ASR 9000, MX-960/480/240, M-series, T-series, CRS-1/3



Co jest tym „sprzętem”?



	CPU	NPU	ASIC
Control Plane	Zwykle stosowany, coraz częściej wykorzystuje się SMP i wirtualizację <ul style="list-style-type: none">• PSD (JNPR)• SDR (CSCO)	Nieefektywne – bardzo trudno zaimplementować protokoły kontrolne	Niedostępne
Data Plane	Routery serii J, ISR G1/G2/7200	Stosowany w routerach brzegowych	Routery szkieletowe – seria T, CRS-1/3
Service Plane	Routery serii J, ISR G1/G2/7200	Zwykle stosowane.	--

NP vs ASIC

- Granice coraz bardziej się zacierają

- W przeszłości

NPU – duży zestaw prostych operacji typowych dla przetwarzania ruchu w sieci zaimplementowany w krzemie; niezbyt szybkie, ale bardzo elastyczne

ASIC – mały zestaw złożonych operacji ale bardzo szybko realizowanych; bardzo szybkie, ale bez elastyczności

- Dzisiaj

NPU – więcej zaawansowanych funkcji

ASIC – większy zestaw operacji, „programowalne” ASICi

NP vs ASIC

- NPU

Zwykle kości dostępne od producentów na rynku (Broadcom, Marvell, EZTech)

Zwykle również bardzo uniwersalne – wiele odbiorców końcowych

...ale wiecie do czego jest „klej do wszystkiego...”

- ASIC

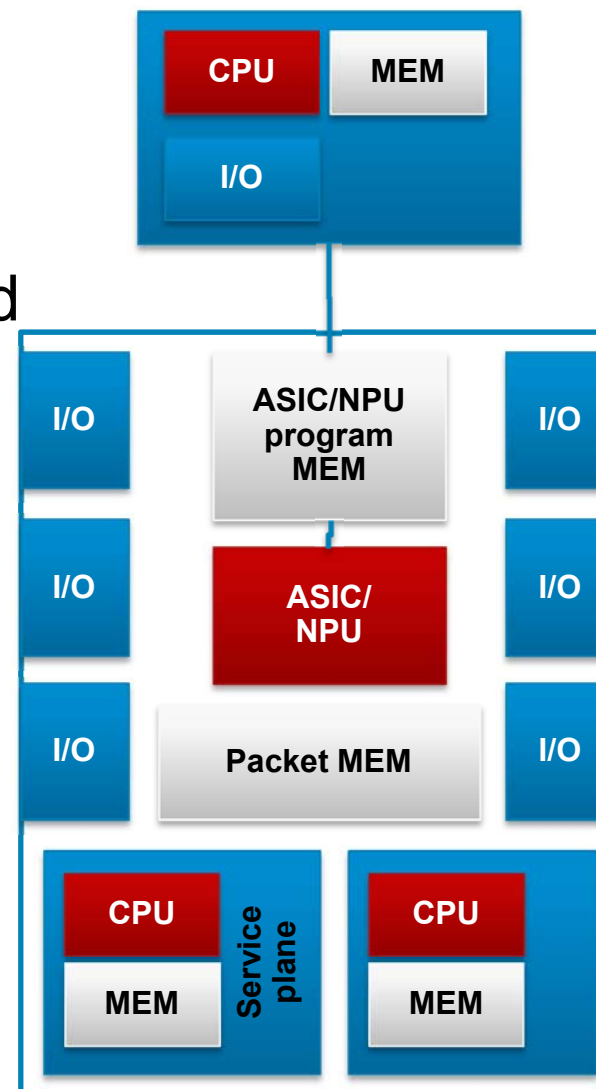
Duże koszty produkcji – bardzo specjalizowane zadania

Produkowane przez producentów sprzętu (Juniper, Cisco)

Od razu konkretnie dostosowane do pełnionej funkcji, pozycjonowania

Zcentralizowany router sprzętowy

- Pojedynczy kompleks odpowiedzialny za data plane
- Zwykle architektura ze współdzieleniem pamięci (shared memory)
- Zwykle do 10GE i mała wielokrotność 10GE
- Przykłady
 - MX80, M7i/10i
 - ASR1k



Rozproszony router sprzętowy

- Wiele kompleksów data plane

Każdy w architekturze ze współdzieloną pamięcią

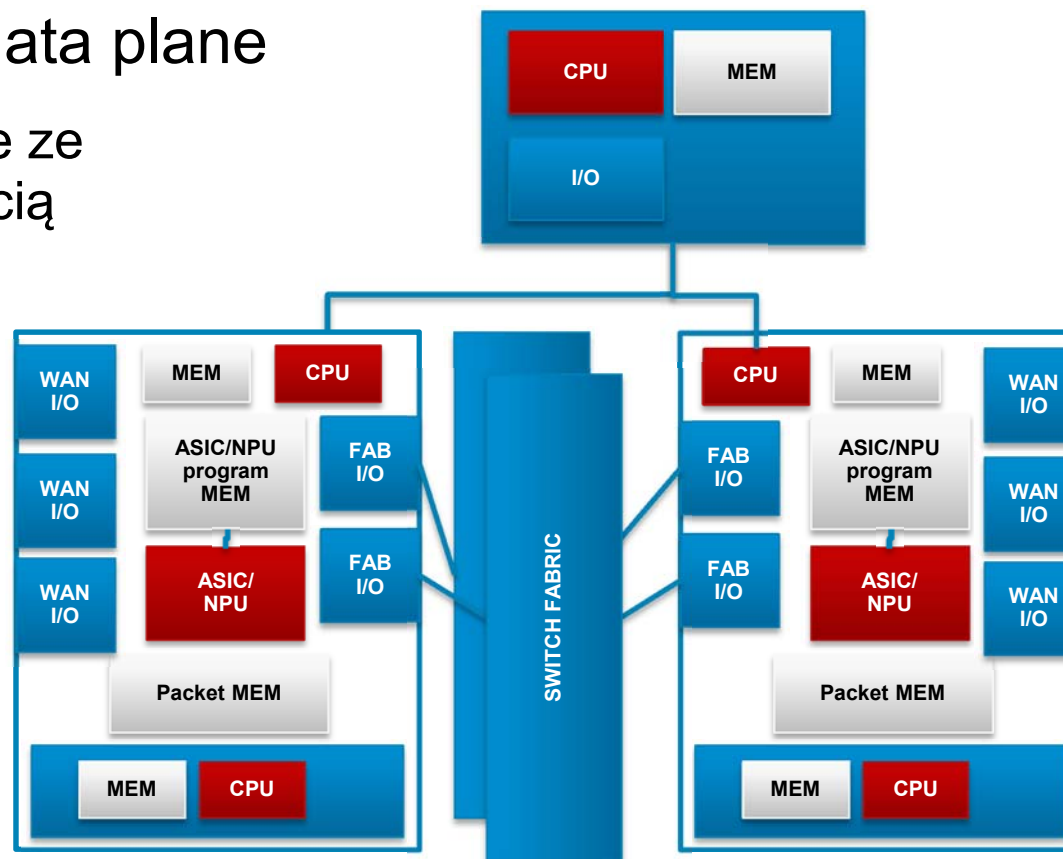
Połączone matrycą

- Duże systemy

- Przykłady

Seria T, MX

ASR 9000



Gdzie są ograniczenia?

- Ograniczenia systemu są ograniczeniami jego komponentów i połączeń między nimi

Interakcja ASICów z pamięcią programu

ilość wpisów per funkcjonalność (np. ACL, QoS)

typy pamięci - SRAM vs. RLDRAM vs. SDRAM vs TCAM...

Matryca – i jej połączenia

wejście i wyjście, buforowanie

gwarancje – ruch unicast, multicast

poziom overspeed

HoL blocking, mechanizmy flow-control

Inne łącza/szyny

Gdzie są ograniczenia? C.d.

ASIC/NPU
program
MEM

ASIC/
NPU

- Pamięć gdzie ASIC/NPU znajduje informacje jaką akcję należy podjąć dla danego pakietu

FIB

Filtry ruchowe (ACL), rate-limiters, klasyfikatory

Scheduling, modyfikacje nagłówka, enkapsulacja L2

- Wielokrotny odczyt/dostęp dla każdego pakietu.

Przy projektowaniu systemów przyjmuje się jako parametr – budżet dostępu do pamięci per pakiet.

Budżet może być przekroczony dla niektórych pakietów

Jeśli średnia ilość dostępu per pakiet przekroczy budżet, wtedy następuje degradacja wydajności (pps)

Gdzie są ograniczenia? C.d.

ASIC/NPU
program
MEM

ASIC/
NPU

- Szybsza pamięć – większy budżet

TCAM -> \$\$\$\$ + wysokie pobór mocy i wydzielane ciepło.

SRAM -> 10ns czas dostępu -> \$\$\$

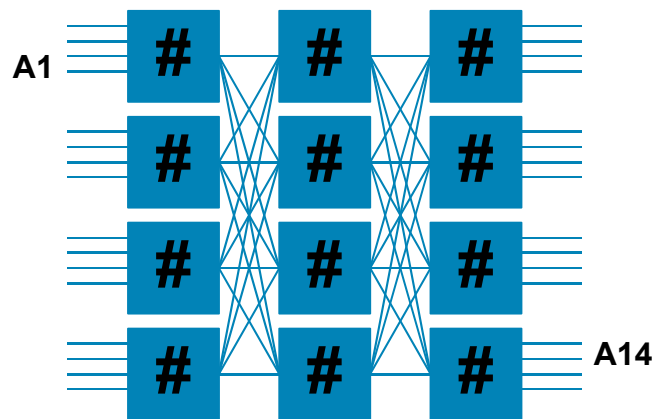
RLDRAM -> 20ns -> \$\$

DRAM -> 60ns czas dostępu-> \$

20ns -> 10ns mała różnica? Ale to 10Mpps -> 20Mpps. @64B
przepustowość 5Gbps -> 10Gbps.

Architektura Switch Fabric

- Cross-Bar, CLOS/BENES, torus – wymagają adresowania



- Pakiet/ramka pakowana w wewnętrzną enkapsulację.
Często celki (ale nie ATM)
 - efekt „piły” (ang. saw-tooth)
 - stosuje się overspeeding matrycy i łącz do niej

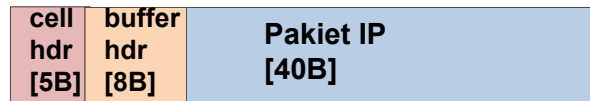
Efekt „piły” w rzeczywistości

Przykład
Formatu
ramki



Stały narzut [nagłówek, ~10%]
Nagłówek matrycy
Wyrównanie

40B pakiet IP:



Najlepsza efektywność

1Mpps = 1Mcps
1Gb/s → 1.33Gb/s

41B pakiet IP:



Słabo

1Mpps = 2Mcps
1Gb/s → 2.6Gb/s

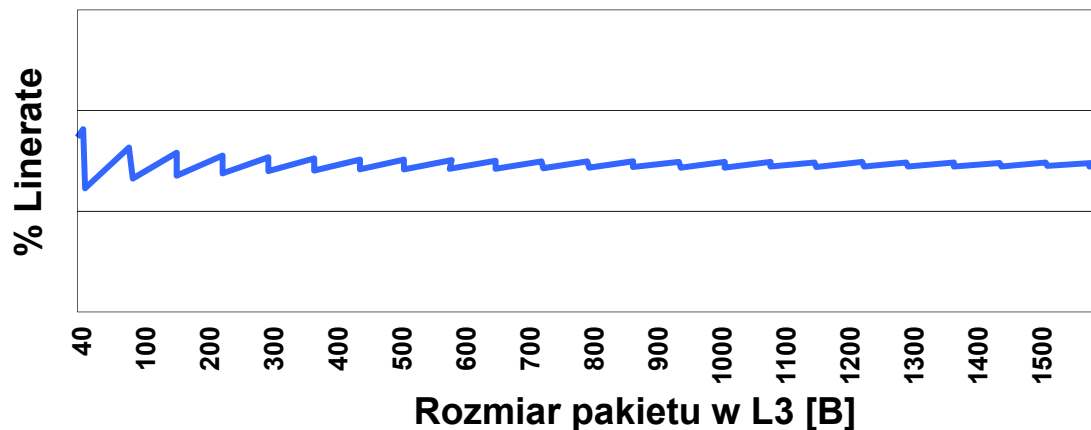
64B pakiet IP:



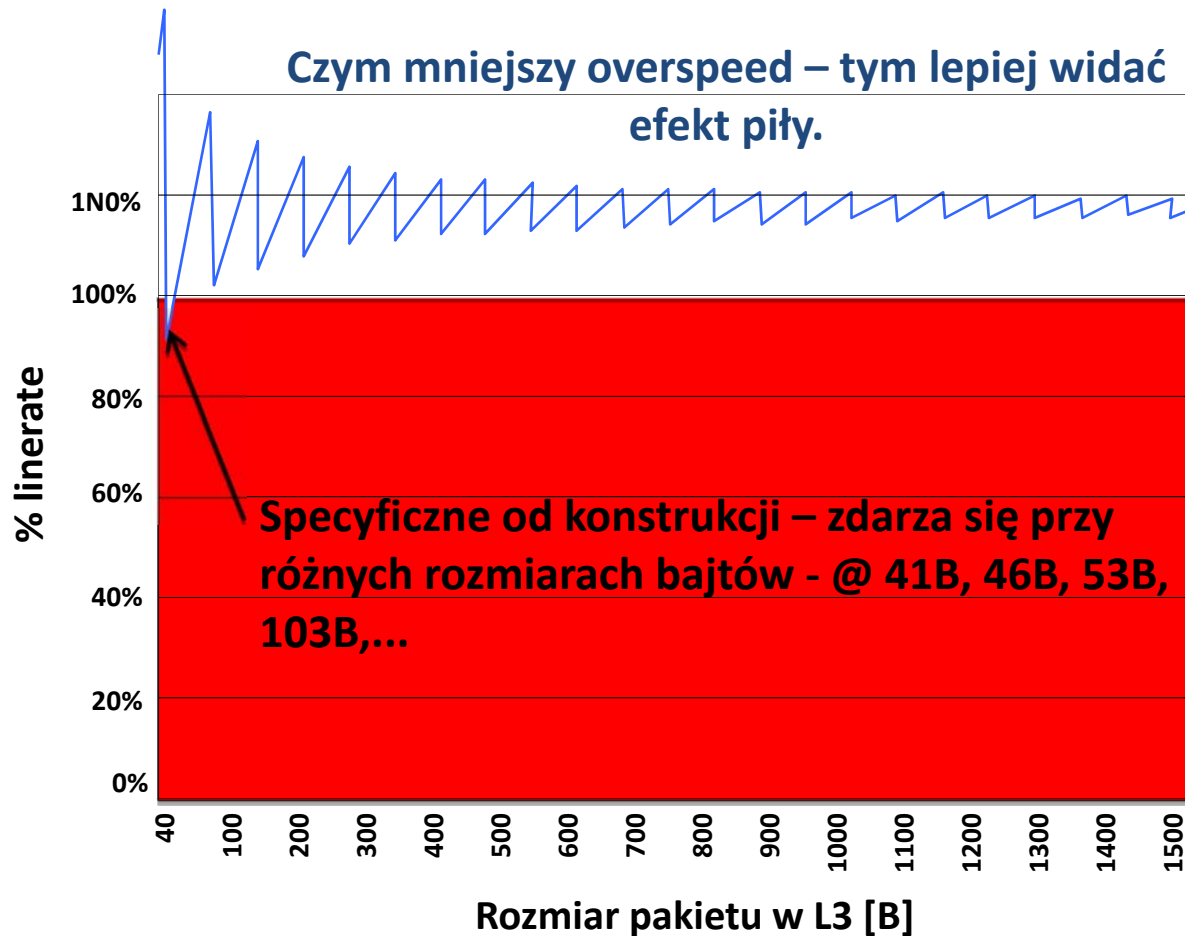
Lepiej:

1Mpps = 2Mcps
1Gb/s → 1.7Gb/s

Efekt na wykresie pomiaru ruchu

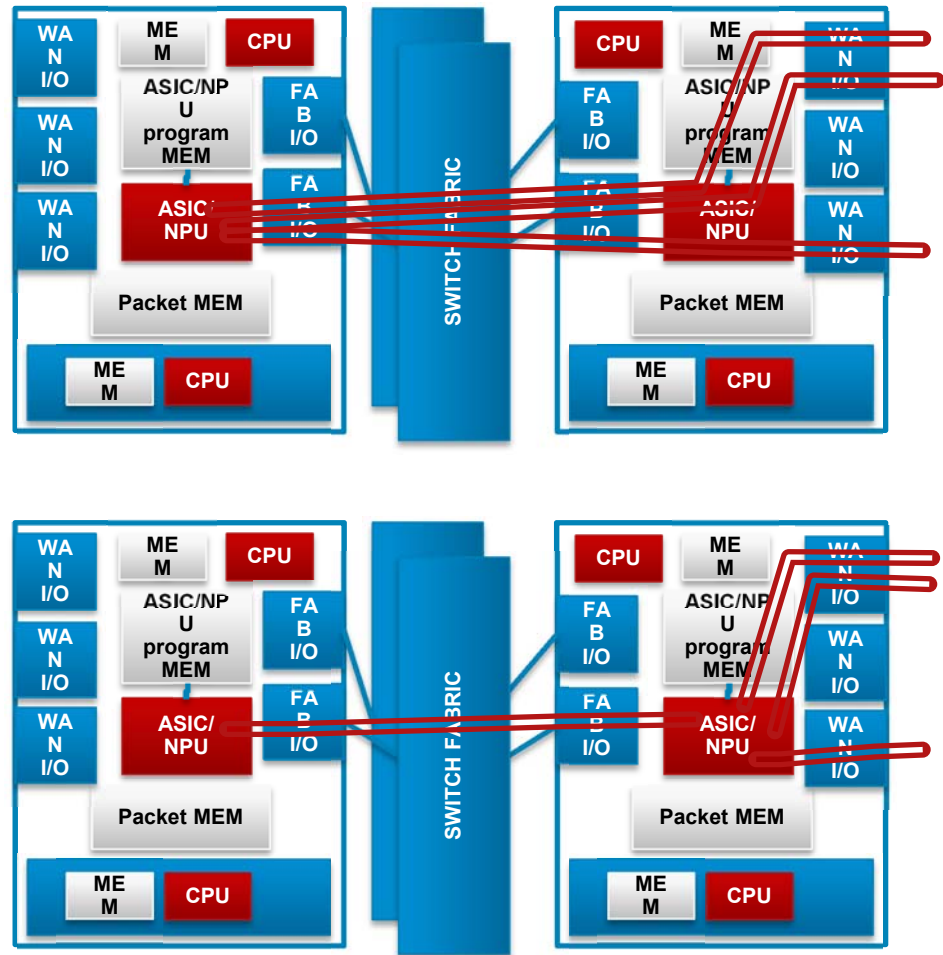


Overspeed – „żeby nie było widać”



Architektura Switch Fabric

- Address docelowy == identyfikator interfejsu wyjsciowego (VC/VLAN)
 - nie ma potrzeby wykonywania lookup na karcie wyjsciowej
 - ograniczona skalowalnosc (tysiące interfejsów)
 - VoQ – N kolejek to-Fabric per interfejs
 - GRANT/back-pressure realizowany in-band
- Address docelowy == identyfikator wyjsciowego Forwarding Engine
 - niezbędny lookup na karcie wyjsciowej
 - wysoka skalowalnosc
 - N kolejek to-Fabric queues per Forwarding Engine
 - GRANT/back-pressure realizowany out-of-band (linie sterujace)
 - Wspólczesnie stosowane.



Switch Fabric – kontrola przepływu

- Nawet dla architektury nieblokującej i przy zastosowaniu overspeed, wyjście ze Switch Fabric jest ograniczone
- SF zwykle nie jest w stanie radzić sobie z natłokami i HoLB
- Nadawcy muszą zwolnić:

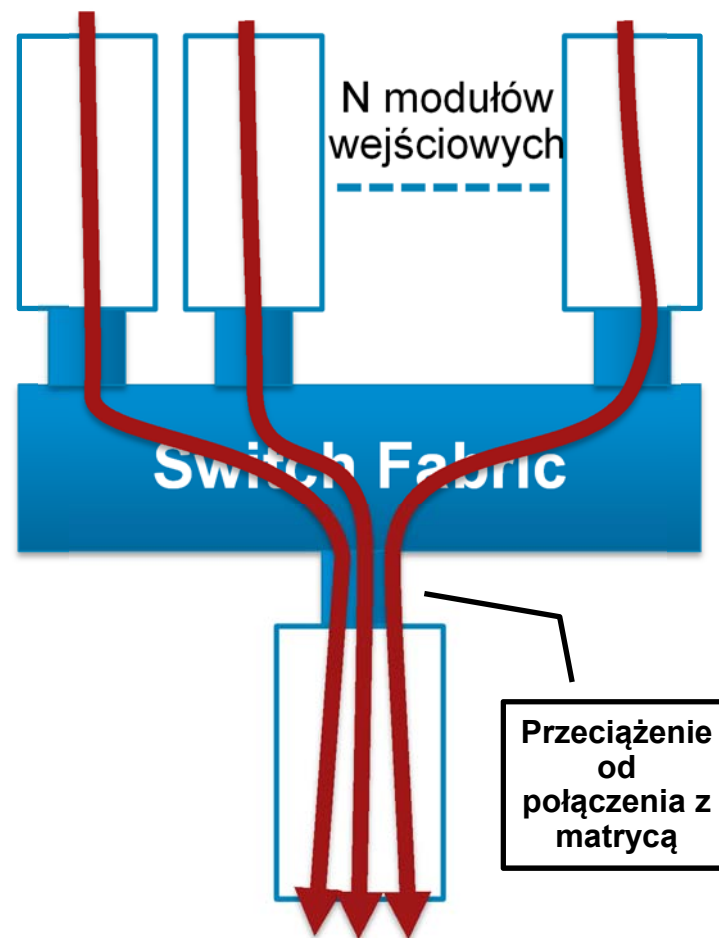
Buforowanie pakietów w kolejkach QoS przed wejściem do SF

Proaktywne request-grant - przed wysłaniem każdego pakietu do SF

brak grantu – brak możliwości transmisji

Reaktywne back-pressure

Inicjowany przez wyjściowy FE kiedy 'widzi' przeciążenie.



Pytania?



