Wydajność obliczeń równoległych

Wydajność obliczeń równoległych

- Cel zrównoleglenia – skrócenie czasu działania programów, zwiększenie wydajności obliczeń, uzyskanie szybszego przetwarzania
- Wyrażanie wydajności:
  - liczba operacji na sekundę (MIPS)
    - niemożliwość oszacowania jak szybko wykonywana jest pojedyncza instrukcja – zależność od kontekstu
    - trudność określenia z ilu i jakich instrukcji składa się program
  - miary zależne od dziedziny zastosowań
    - np. liczba operacji zmiennoprzecinkowych na sekundę (MFLOPS)
  - miary względne, w których porównuje się czasy wykonania programu (rozwiązania zadania) na jednym i wielu procesorach
    - Przyspieszenie obliczeń:
      - $S(p) = T_s / T_p(p)$ – *punkt widzenia użytkownika (przyspieszenie „bezwzględne”)*
      - $S(p) = T_p(1)/T_p(p)$ – *częste w badaniach skalowalności (przyspieszenie „względne”)*
    - Efektywność zrównoleglenia: $E(p) = S(p) / p$

Wydajność obliczeń równoległych

- Wydajność obliczeń równoległych
  - analiza Amdahla (1967)
  - prawo Amdahla:
    - przy liczbie procesorów zmierzającej do nieskończoności czas rozwiązania określonego zadania nie może zmaleć poniżej czasu wykonania części sekwencyjnej programu, przyspieszenie nie może przekroczyć wartości wyznaczonej przez udział części sekwencyjnej, a efektywność zrównoleglenia zmierza do zera
  - prawo Amdahla stwierdza istotne ograniczenie wydajności programów równoległych
  - co więcej, prawo Amdahla nie uwzględnia m.in.:
    - komunikacji, która dodatkowo spowalnia obliczenia równoległe
    - braku zrównoważenia obciążenia procesorów

Wydajność obliczeń równoległych

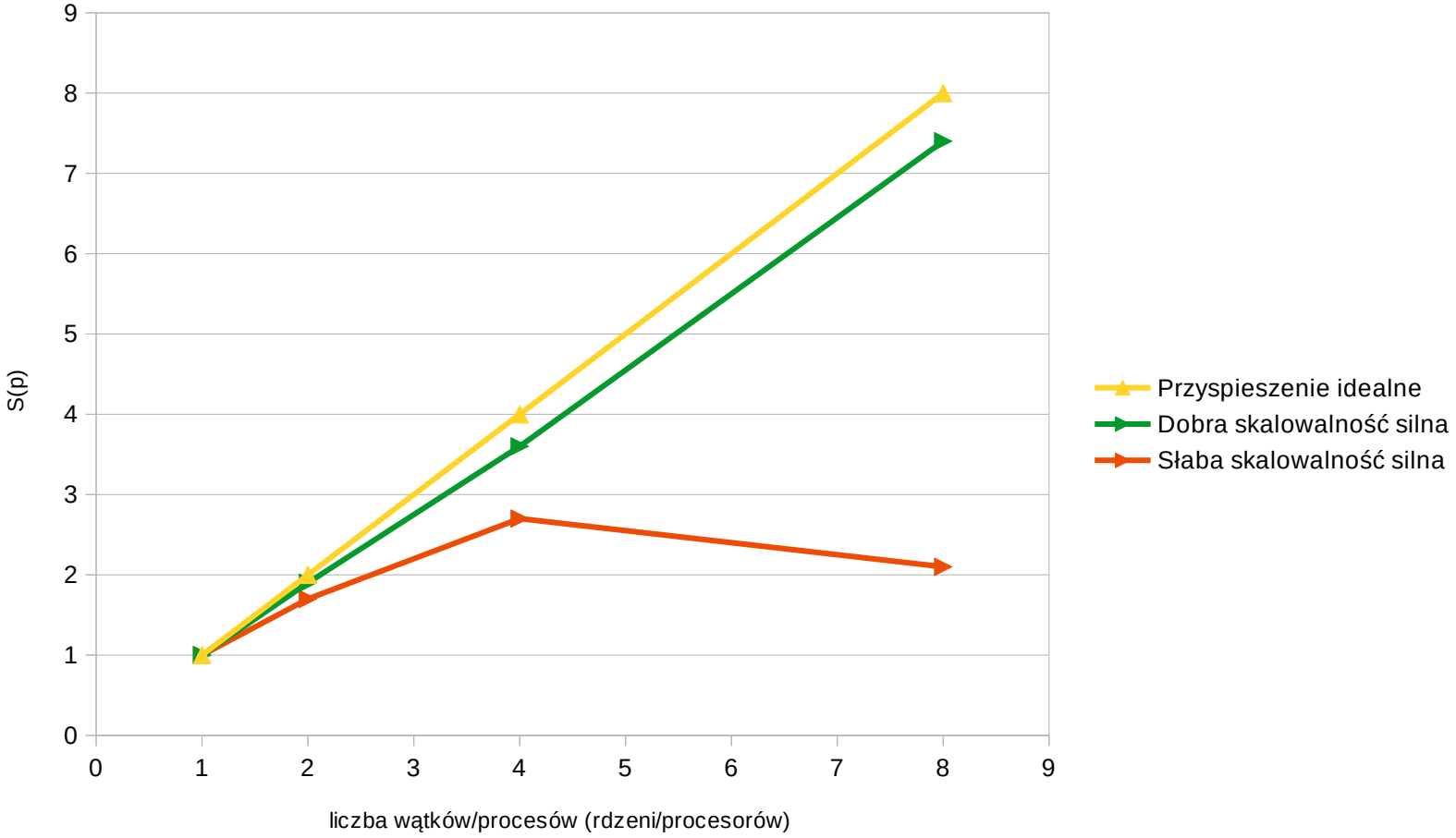
- Wydajność obliczeń równoległych:
  - prawo Amdahla dotyczy równoległego rozwiązania z coraz większą liczbą procesorów **tego samego zadania** (analiza przy stałym rozmiarze zadania)
  - można problem zrównoleglenia postawić jako problem rozwiązania sekwencji zadań przy stałym czasie wykonania
  - analiza Gustafsona (1988)
  - przyspieszenie przeskalowane: modyfikacja definicji przyspieszenia obliczeń i efektywności zrównoleglenia, tak żeby uwzględnić fakt zwiększania rozmiaru zadania wraz ze wzrostem liczby procesorów
  - przy zmodyfikowanych definicjach częściej możliwe jest uzyskanie przyspieszenia zmierzającego do nieskończoności i efektywności nie spadającej poniżej pewnej stałej wartości

Wydajność obliczeń równoległych

- Skalowalność obliczeń
  - skalowalność obliczeń równoległych oznacza właściwe zachowanie programu w przypadku rosnących zasobów obliczeniowych (i ewentualnie rosnącego obciążenia obliczeniowego)
  - rozróżnia się dwa podstawowe pojęcia związane ze skalowalnością obliczeń:
    - skalowalność w sensie silnym i skalowalność w sensie słabym
  - skalowalność w sensie silnym oznacza liniowe (lub pozostające blisko liniowego) przyspieszenie programu równoległego zgodnie z klasyczną analizą ze stałym rozmiarem zadania
    - poza rzadkimi przypadkami programów z minimalnym udziałem narzutów wykonania równoległego (tzw. programy "żenująco" równoległe, *embarassingly parallel*) zdecydowana większość programów nie wykazuje skalowalności w sensie silnym
    - standardowo, przyspieszenie programów nie tylko nie jest liniowe, ale wręcz z rosnącą liczbą użytych wątków/procesów (rdzeni/procesorów) często spada do zera

Skalowalność w sensie silnym

- Wykres skalowalności w sensie silnym
  - wykres przyspieszenia $S(p) = T_p(1) / T_p(p)$ dla tego samego zadania



Wydajność obliczeń równoległych

- Skalowalność obliczeń (cd.)
  - skalowalność w sensie słabym odnosi się do badania zachowania programów równoległych, gdy rozmiar jest stały na pojedynczy wątek/proces (całkowity rozmiar zadania W, najczęściej wyrażany liczbą wykonywanych operacji, rośnie proporcjonalnie do liczby wątków/procesów)
    - jednym z możliwych wskaźników dobrej skalowalności jest utrzymywanie stałego (lub nieznacznie rosnącego) czasu wykonania – czas wykonania zadania p-razy większego przy użyciu p wątków/procesów pozostaje niezmienny
      - $\mathbf{T_p(p, W(p)) \approx const}$ dla $\mathbf{W(p) = p*W(1)}$
      - przy całkowitym braku skalowalności taki czas mógłby rosnąć p-krotnie, tak jak czas wykonania sekwencyjnego (z użyciem jednego wątku/procesu)
    - drugą miarą skalowalności w sensie słabym jest tzw. przyspieszenie przeskalowane (*scaled speed-up*), gdzie przyspieszenie dla każdej liczby wątków/procesów p oblicza się indywidualnie, badając każdorazowo zadanie o rozmiarze dostosowanym do liczby wątków/procesów
      - $\mathbf{S^S(p) = S^S(\ p, W(p)\ ) = T_p(1, W(p)) / T_p(p, W(p))}$ dla $\mathbf{W(p) = p*W(1)}$

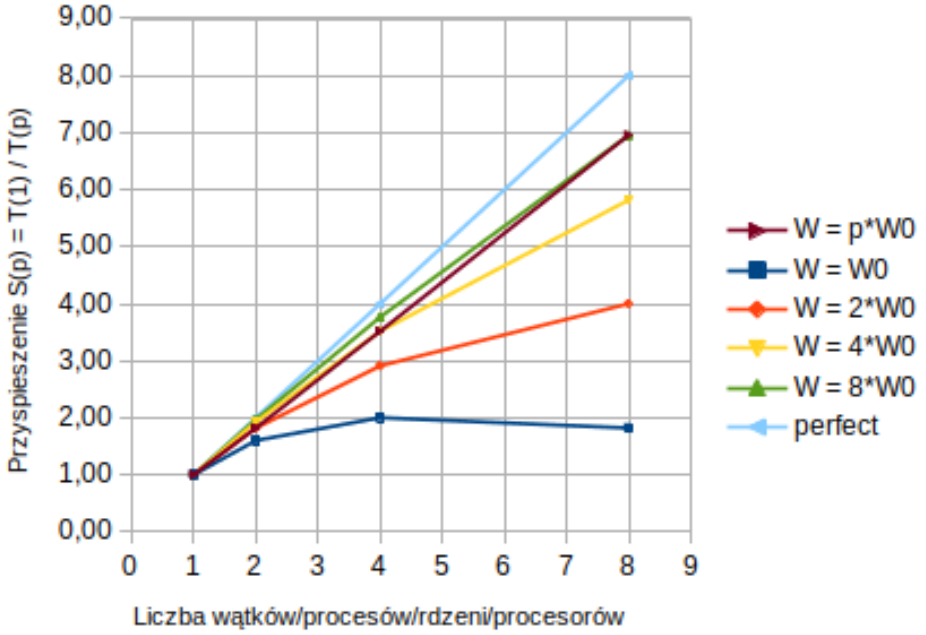
Skalowalność w sensie słabym

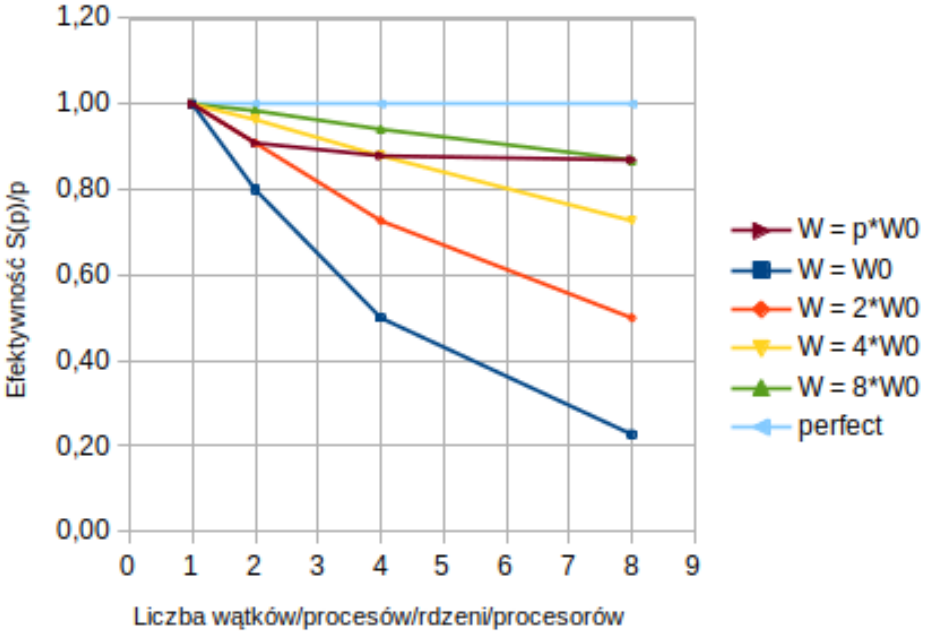
- Tabela wyników do badania skalowalności w sensie słabym
  - czasy wykonania dla różnych liczb wątków/procesów/rdzeni/procesorów, p oraz **różnych rozmiarów zadania W (wyrażanych w liczbie operacji zmiennoprzecinkowych, nie w rozmiarze danych wejściowych)**
  - czasy na przekątnej głównej oznaczają czasy wykonania programu dla p wątków/.../.../... i zadania p-razy większego niż dla 1 wątku/.../.../...
  - stosunek czasów w pierwszym wierszu do czasów na przekątnej głównej służy do skonstruowania wykresu przyspieszenia przeskalowanego, $S^S(p)$
  - w podanym w tabeli przykładzie dla liczby wątków p=4 wyczerpywały się zasoby maszyny uniemożliwiając dalszy wzrost wydajności i spadek czasu wykonania

	$W = W_0$	$W = 2*W_0$	$W = 4*W_0$	$W = 8*W_0$
p = 1	0,008148	0,016521	0,033008	0,066044
p = 2	0,004187	0,008405	0,016775	0,033659
p = 4	0,002671	0,004964	0,009839	0,019458
p = 8	0,002492	0,004745	0,008975	0,017806

Skalowalność

- Przyspieszenie i efektywność zrównoleglenia w badaniu skalowalności
  - dla rosnącego rozmiaru zadania krzywe przyspieszenia często zbliżają się do przyspieszenia idealnego, a efektywność do 100%
  - w badaniu skalowalności, przy konstrukcji przyspieszenia przeskalowanego, wybierane są punkty dla rozmiaru rosnącego proporcjonalnie do liczby wątków/procesów/rdzeni/procesorów





Wydajność obliczeń równoległych

- Narzut obliczeń równoległych to wszystkie czynniki wpływające na odbieganie czasu wykonania programu równoległego od czasu idealnego (p razy krótszego dla p wątków/procesów)
  - narzut całkowity: $N_p(p) = p*T_p(p) – T_p(1)$ [=> $T_p(p) = T_p(1)/p + N_p(p)/p$]
    - narzut na pojedynczy procesor/rdzeń $N_1(p) = N_p(p)/p = T_p(p) - T_p(1)/p$
  - w analizie Amdahla (przy stałym rozmiarze zadania) narzut na pojedynczy procesor/rdzeń (≈czas wykonania części niedającej się zrównoleglić) pozostaje stały
    - w efekcie jego proporcja w miarę wzrostu liczby procesorów/rdzeni w stosunku do malejącego czasu wykonania na pojedynczym procesorze/rdzeniu rośnie
  - w analizie Gustafsona narzut na pojedynczy procesor/rdzeń (≈czas wykonania części niedającej się zrównoleglić) także pozostaje stały
    - jednak jego proporcja w stosunku do stałego czasu wykonania na pojedynczym procesorze/rdzeniu pozostaje stała (dzięki rozważaniu zadań o rozmiarze rosnącym wraz z liczbą procesorów/rdzeni)
  - ogólnie:
    - do uzyskania skalowalności w sensie silnym konieczne jest, żeby wraz z rosnącą liczbą procesorów/rdzeni narzut na jeden procesor/rdzeń był zerowy lub malał (i to odwrotnie proporcjonalnie do liczby wątków/procesów, czyli jak 1/p)
    - do uzyskania skalowalności w sensie słabym wystarcza, żeby narzut na jeden procesor/rdzeń pozostawał stały albo malał (lub tylko nieznacznie rósł)

Wydajność obliczeń równoległych

- Czynniki wpływające na wydajność obliczeń równoległych (decydujące o narzucie obliczeń równoległych):
  - Czas wykonywania części niedającej się zrównoleglić (część sekwencyjna, *serial fraction*)
  - (Nie)zrównoważenie obciążenia
    - różnica między maksymalnym czasem obliczeń jednego z wątków/procesów, w stosunku do średniego czasu obliczeń wszystkich wątków/procesów
  - Czas komunikacji/synchronizacji
  - Czas realizacji dodatkowych obliczeń
  - Czas wykonywania operacji systemowych (uruchomienie wątków, procesów, alokacja pamięci itp.)
  - Inne czynniki (np. skalowanie częstotliwości pracy rdzeni)
- Jednym z podstawowych powodów ograniczenia wydajności obliczeń dla maszyn z pamięcią wspólną jest wyczerpanie zasobów:
  - liczby rdzeni w maszynie
  - przepustowości magistrali z pamięcią (dla programów wykonujących relatywnie mało obliczeń w stosunku do rozmiaru pobieranych danych z pamięci)

Wydajność obliczeń równoległych

- Wnioski z przeprowadzonej analizy:
  - możliwe jest osiągnięcie zadowalających parametrów wykonania równoległego jeżeli:
    - nie dążymy do uzyskania zerowego czasu działania dla pewnego konkretnego zadania o stałym rozmiarze
    - staramy się optymalnie zrównoleglić zadania o rozmiarze rosnącym wraz z liczbą używanych procesorów
  - inaczej:
    - **nie po to stosujemy obliczenia (masowo) równoległe, żeby coraz szybciej rozwiązywać te same zadania, ale po to, żeby efektywnie rozwiązywać zadania coraz większe, problemy coraz bardziej złożone**