Materiały do wykładu

10. Architektury specjalizowane

Marcin Peczarski

Instytut Informatyki
Uniwersytet Warszawski

27 maja 2014

Architektury do specyficznych zastosowań

- DSP
  - Digital Signal Processing – cyfrowe przetwarzanie sygnałów
  - Digital Signal Processor – procesor sygnałowy
- GPU
  - Graphics Processing Unit – procesor graficzny

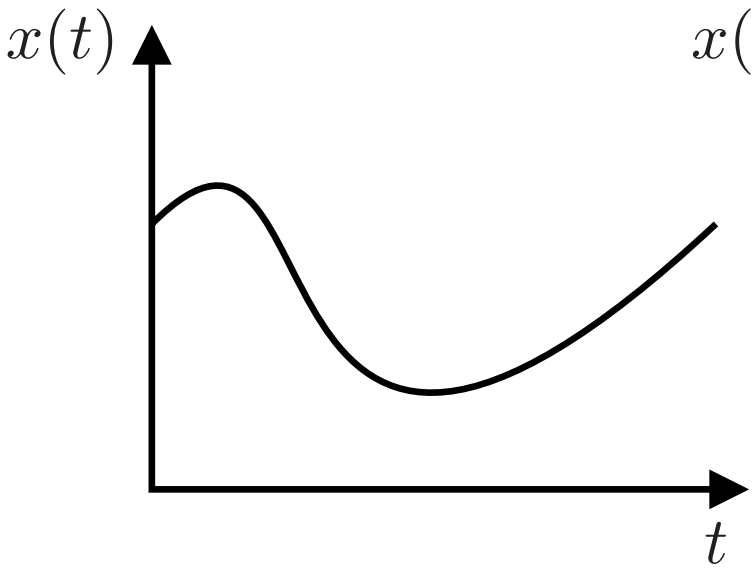
Zastosowania DSP

- Dziedziny
  - telekomunikacja
  - radiolokacja
  - sprzęt audio-wideo
  - medycyna
- Przykładowe aplikacje
  - kompresja i dekompresja sygnału mowy
  - wykrywanie rodzaju obiektu latającego
  - magnetyczny rezonans jądrowy
  - tomografia komputerowa

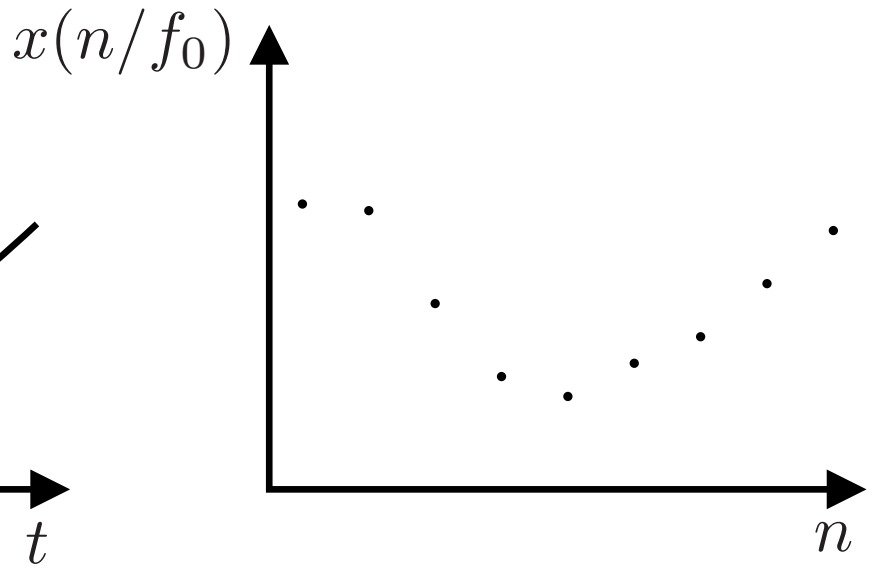
Dlaczego to działa?

▷ Twierdzenie o próbkowaniu

Jeżeli sygnał ciągły w czasie $x(t)$ zawiera tylko składowe o częstotliwościach poniżej $f_0/2$, to jest jednoznacznie reprezentowany przez swoje wartości w dyskretnych chwilach czasu $x(n/f_0)$, gdzie n to liczba całkowita, a f_0 to częstotliwość próbkowania.



przed próbkowaniem



po próbkowaniu

Twierdzenie o próbkowaniu – odzyskiwanie sygnału

▷ Przy pewnych założeniach zachodzi

$$x(t) = \sum_{n=-\infty}^{\infty} x(n/f_0) S(t f_0 - n),$$

gdzie

$$S(\tau) = \begin{cases} \frac{\sin \pi\tau}{\pi\tau}, & \text{gdy } \tau \neq 0; \\ 1, & \text{gdy } \tau = 0. \end{cases}$$

Kwantowanie

- Zamiana wartości próbki na liczbę
- Rozdzielczość, liczba bitów reprezentacji n
- Proces nieodwracalny – wartości zaokrąglane do najbliższej reprezentowalnej wartości
- Szum kwantowania

$$SNR[\mathrm{dB}] = 20n \log 2 + 10 \log \tfrac{3}{4} \approx 6n - 1{,}25$$

- Równomierne, liniowe
- Nierównomierne, nieliniowe: A-law, μ-law

Dlaczego to działa?

- *Wrażenie widzenia przez ludzkie oko* prawie dowolnej barwy można wywołać przez sumowanie z wagami trzech kolorów podstawowych:
  - czerwony 700,1 nm,
  - zielony 546,1 nm,
  - niebieski 435,8 nm.
- Metoda sumowania przestrzennego:
  - oko ludzkie nie rozróżnia blisko leżących punktów.
- Metoda sumowania czasowego:
  - oko ludzkie nie rozróżnia zmian o częstotliwości powyżej kilkunastu Hz.

Arytmetyka procesorów sygnałowych

- Arytmetyka stałopozycyjna U2

$$b_0{,}b_1 \ldots b_{n-2}b_{n-1}$$

$$-b_0 + \sum_{i=1}^{n-1} b_i 2^{-i}$$

- Reprezentuje wartości z przedziału $[-1, 1)$.
- Mnożenie liczb z przedziału $(-1, 1)$ nie powoduje przepełnienia.
- Mnożenie -1 razy -1 powoduje przepełnienie.
- Dodawanie z nasyceniem

Typowe operacje DSP (1)

- Filtr SOI (skończona odpowiedź impulsowa, ang. *FIR – finite impulse response*)

$$y_t = \sum_{i=0}^{n} b_i x_{t-i}$$

- Filtr NOI (nieskończona odpowiedź impulsowa, ang. *IIR – infinite impulse response*)

$$y_t = \sum_{i=1}^{m} a_i y_{t-i} + \sum_{i=0}^{n} b_i x_{t-i}$$

Typowe operacje DSP (2)

- Splot

$$z_n = \sum_{i=0}^{n} x_i y_{n-i}$$

- Dyskretne przekształcenie Fouriera (DFT)

$$A_m = \sum_{n=0}^{N-1} a_n w^{mn}$$

Opearacja MAC

- Typowe operacje DSP pelegają na mnożeniu z akumulacją wyniku

$$a := a + x_i y_j$$

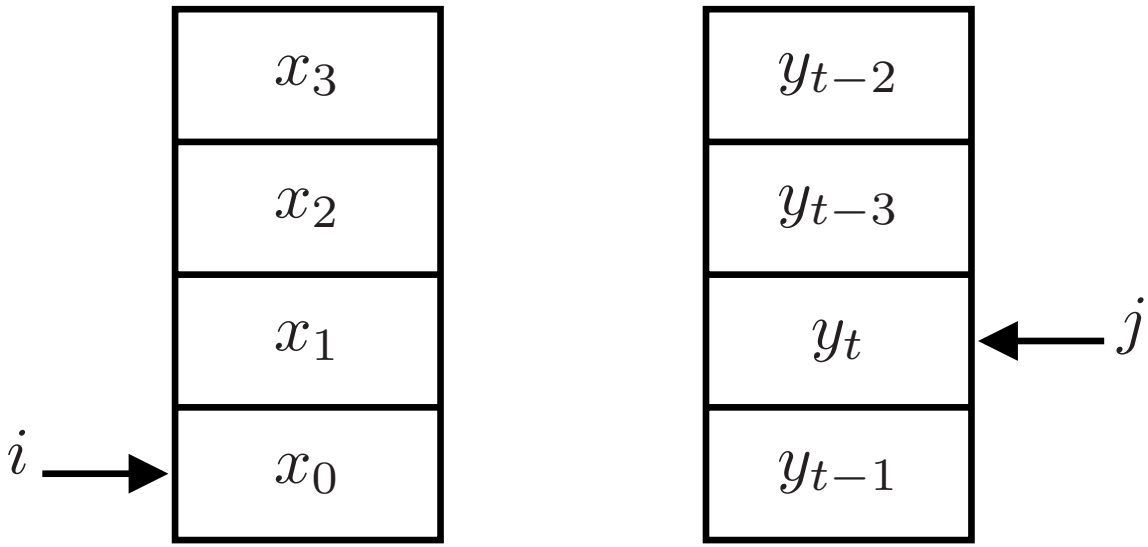
- MAC – multiply and accumulate

$$\begin{aligned} a &:= a + x[i] * y[j] \\ i &:= (i + 1) \bmod n \\ j &:= (j - 1) \bmod n \end{aligned}$$

- Wyniki pośrednie w rejestrach tymczasowych o zwiększonej precyzji

MAC (1)

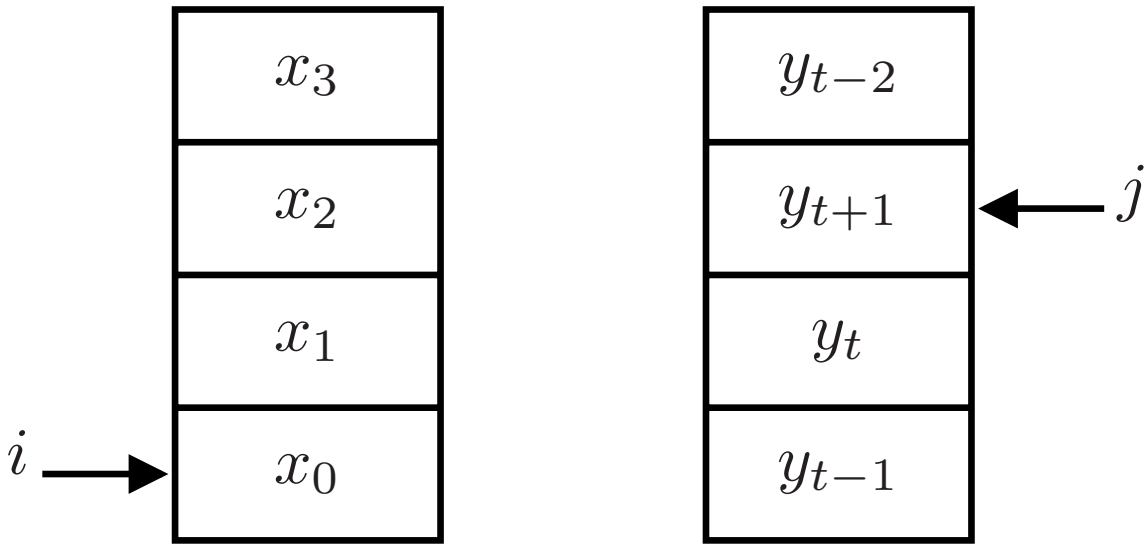
- Liczymy $a_t = \sum_{i=0}^{3} x_i y_{t-i}$ dla kolejnych indeksów t.
- Inicjujemy $a_t := 0$ i indeksy $i := 0$, $j := t \bmod 4$.



- Po czterech operacjach MAC dostajemy $a_t = \sum_{i=0}^{3} x_i y_{t-i}$.
- Indeksy wskazują na pierwotne pozycje.

MAC (2)

- Inicjujemy $a_{t+1} := 0$ i przesuwamy indeks $j := (j+1) \bmod 4$.
- Zamieniamy niepotrzebną już próbkę y_{t-3} na nową y_{t+1}.



- Po kolejnych czterech operacjach MAC $a_{t+1} = \sum_{i=0}^{3} x_i y_{t+1-i}$.
- Powtarzamy powyższy krok dla $t+2, t+3, \ldots$.

FFT

- Szybkie przekształcenie Fouriera
- Sprzętowe wsparcie dla operacji występujących w algorytmie FFT
- Operacja motylkowa

$$x := x + \omega y$$
$$y := x - \omega y$$

- Operacja odwracania bitów

$$01011100 \rightarrow 00111010$$

Cechy procesorów sygnałowych

- Systemy czasu rzeczywistego
  - ściśle oganiczony czas przetwarzania jednej próbki sygnału
  - szybka obsługa przerwań
- Równoległy dostęp do kodu instrukcji i argumentów
  - architektura harwardzka
- Brak sprzętowego wsparcia dla wielozadaniowości
- Wykonywanie pętli bez przerywania potoku
- Wbudowane interfejsy szeregowe dla wprowadzania i wyprowadzania przetwarzanego ciągu danych

Wyświetlanie obrazów

- CRT (ang. *cathode ray tube*) – kineskop, lampa katodowa
- LCD (ang. *liquid crystal display*) – wyświetlacz ciekłokrystaliczny
- DLP (ang. *digital light processing*), DMD (ang. *digital micromirror device*)
- Ekran plazmowy (ang. *plasma display panel*)
- LED (ang. *light-emitting diode display*) – wyświetlacz z diodami elektroluminescencyjnymi
- OLED (ang. *organic light-emitting diode displays*) – wyświetlacz z organicznymi diodami elektroluminescencyjnymi
- e-papier, nanorurki, ...

Parametry monitorów

- Rozdzielczość
- Częstotliwość odchylania pionowego
  - liczba obrazów (półobrazów) na sekundę
- Sposób wyświetlania
  - przeplot (ang. *interlace*)
  - skanowanie progresywne (ang. *progressive, noninterlaced scanning*)

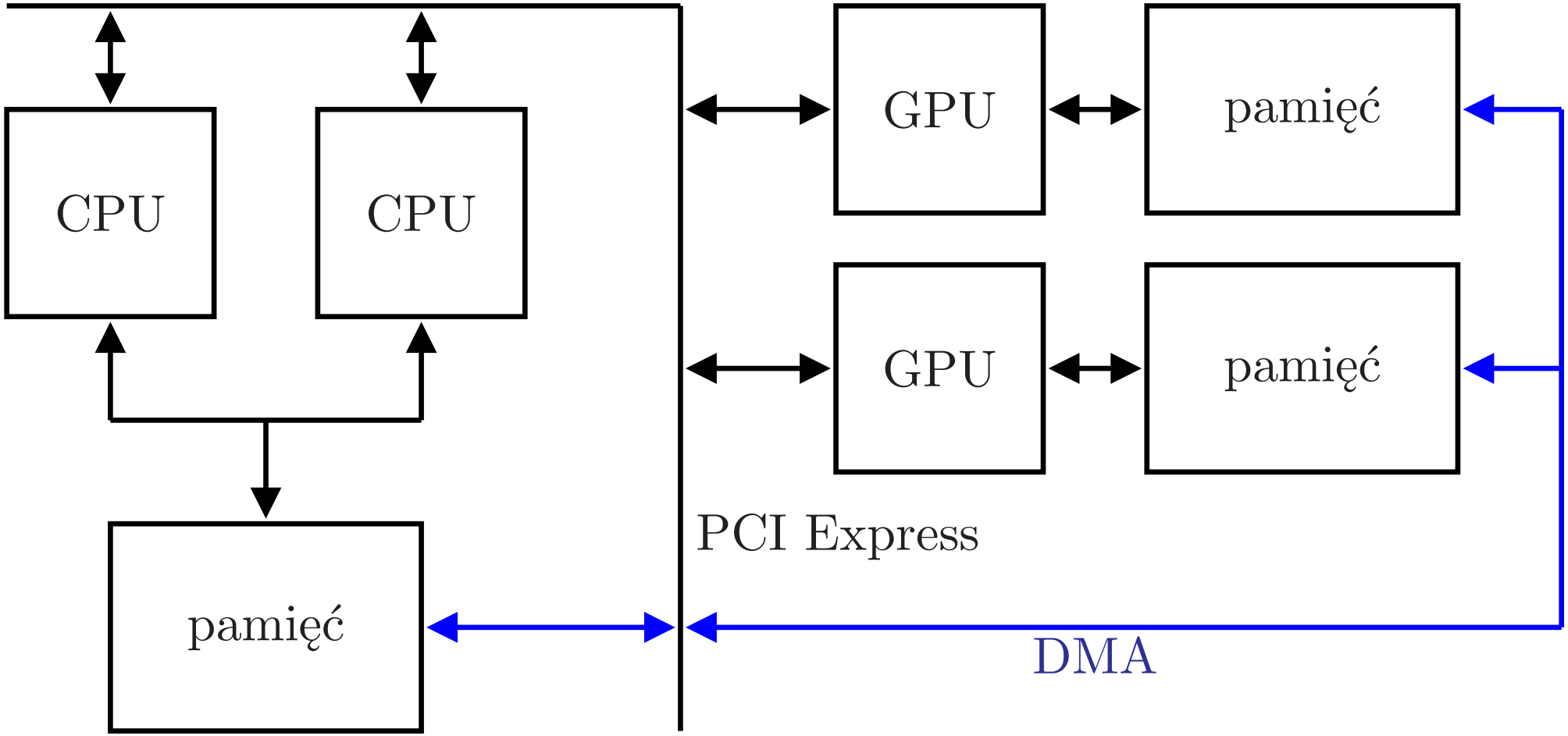
Głębia kolorów

- Obraz czarno-biały
  - 1 bit na piksel
- Paleta kolorów
  - 16 kolorów, 4 bity
  - 256 kolorów, 8 bitów
- Kolor bezpośrednio
  - 8 bitów – czerwony i zielony 3 bity, niebieski 2 bity
  - 15 bitów – 5 bitów na kolor
  - 16 bitów – czerwony i niebieski 5 bitów, zielony 6 bitów
  - 24 bity – 8 bitów na kolor

Rozwój procesorów graficznych

- Specjalizowane układy scalone generujące sygnały dla monitora na podstawie zawartości pamięci obrazu
- Sprzętowe wspomaganie animacji w grach komputerowych – duszki
- Wsparcie dla operacji graficznych
  - tłumaczenie wysokopoziomowego opisu na postać pikseli
  - nadawanie faktury figurom geometrycznym
  - transformacje układu współrzędnych
  - operacje wektorowe i macierzowe
- Użycie procesorów graficznych do obliczeń niezwiązanych z grafiką
- Karty graficzne bez złącza do monitora!
- Integrowanie GPU z CPU

Ogólna architektura



Uproszczenia arytmetyki procesorów graficznych

- Głównie operacje zmiennoprzecinkowe pojedynczej precyzji
- Brak wsparcia dla liczb zdenormalizowanych
- Dzielenie implementowane jako mnożenie przez odwrotność
- Tylko tryby zaokrąglania do najbliższej parzystej i obcięcie
- Ograniczone wsparcie dla NaN
- Pamięć tekstury indeksowana wartościami zmiennoprzecinkowymi, różne tryby interpolacji

Architektura procesorów graficznych (1)

- Procesor ma wiele jednostek wykonawczych
- Zadanie dzieli się na wątki
- W danej chwili jedna jednostka wykonawcza wykonuje jeden wątek
- Wątki mogą być ułożone w jedno-, dwu- lub trójwymiarową tablicę
- Wszystkie wątki wykonują ten sam kod
- Wątek ma dostęp do identyfikatora, składającego się z indeksów w tablicy, dzięki czemu wątki mogą wykonywać różne obliczenia

Architektura procesorów graficznych (2)

- Wątki wykonywane są symultanicznie w paczkach, aby ukryć opóźnienia w dostępie do pamięci graficznej
- Szybkie przełączanie paczek wątków
- Wątki w obrębie jednej paczki wykonują równocześnie tę samą instrukcję
- Jeśli warunek w kodzie powoduje, że pewna instrukcja ma być wykonana tylko przez niektóre wątki w paczce, to pozostałe wątki muszą czekać

Architektura procesorów graficznych (3)

- Wątki dzielą się zasobami: rejestry i pamięć
- Wątki komunikują się przez pamięć
- Odczyty z pamięci mogą odbywać się równocześnie, jeśli dotyczą różnych banków albo wszystkie odwołują się do tego samego adresu
- Jawne zarządzanie hierarchią pamięci