Návrh mikroprocesorov

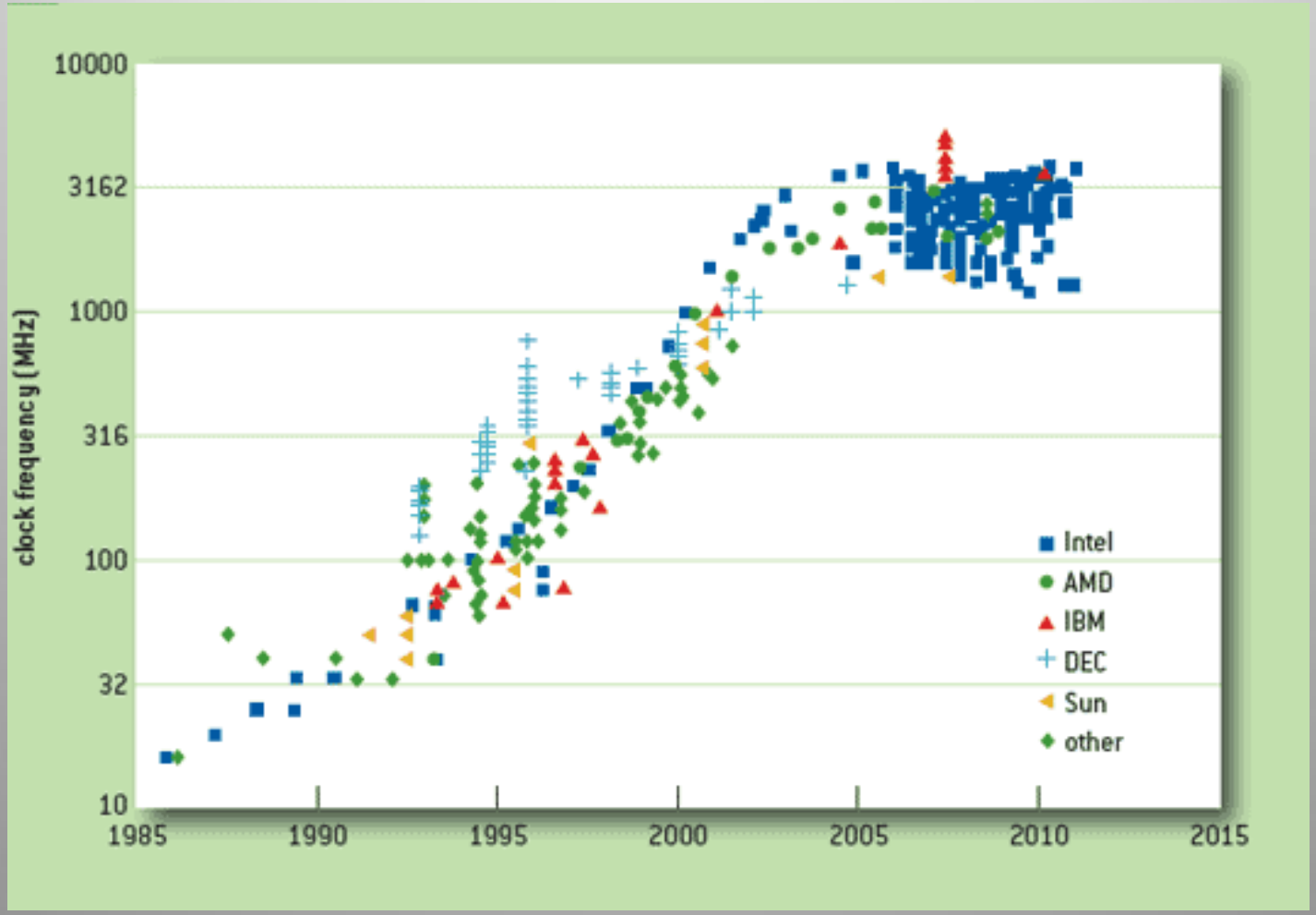
Od jedného jadra ku grafikám

STU
SLOVENSKÁ TECHNICKÁ
UNIVERZITA V BRATISLAVE

Obsah

- Prechod od sekvenčného na paralelný HW
- Vývoj tranzistorov a dosiahnutie niektorých limitov
- Viacjadrové CPU
- Využitie grafických procesorov

Frekvencia CPU



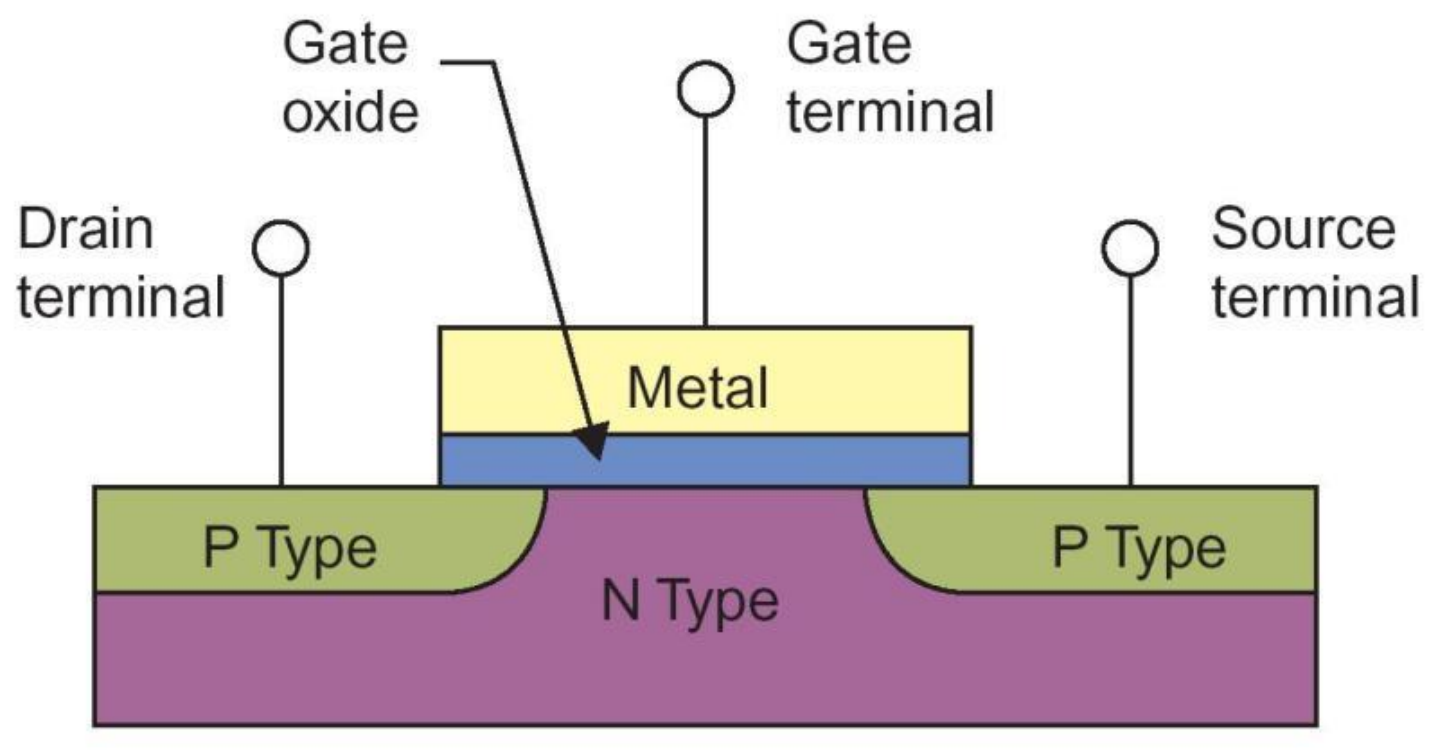
Sekvenčné vs paralelné vykonávanie

- Sekvenčné
  - Programátor nerieši, píše si ako chce
  - Kompilátor nerieši
  - Jednoduchý HW, programovanie aj debugovanie

- Paralelné
  - Programátor/kompilátor musí určiť, čo sa dá vykonať paralelne
  - Paralelný HW požaduje viacej/väčšiu pamäť – potrebuje často lokálne dáta, duplicitné
  - Paralelné časti navzájom komunikujú – časovo náročné
  - ALE! Paralelizmus je prirodzený a často vhodnejší.

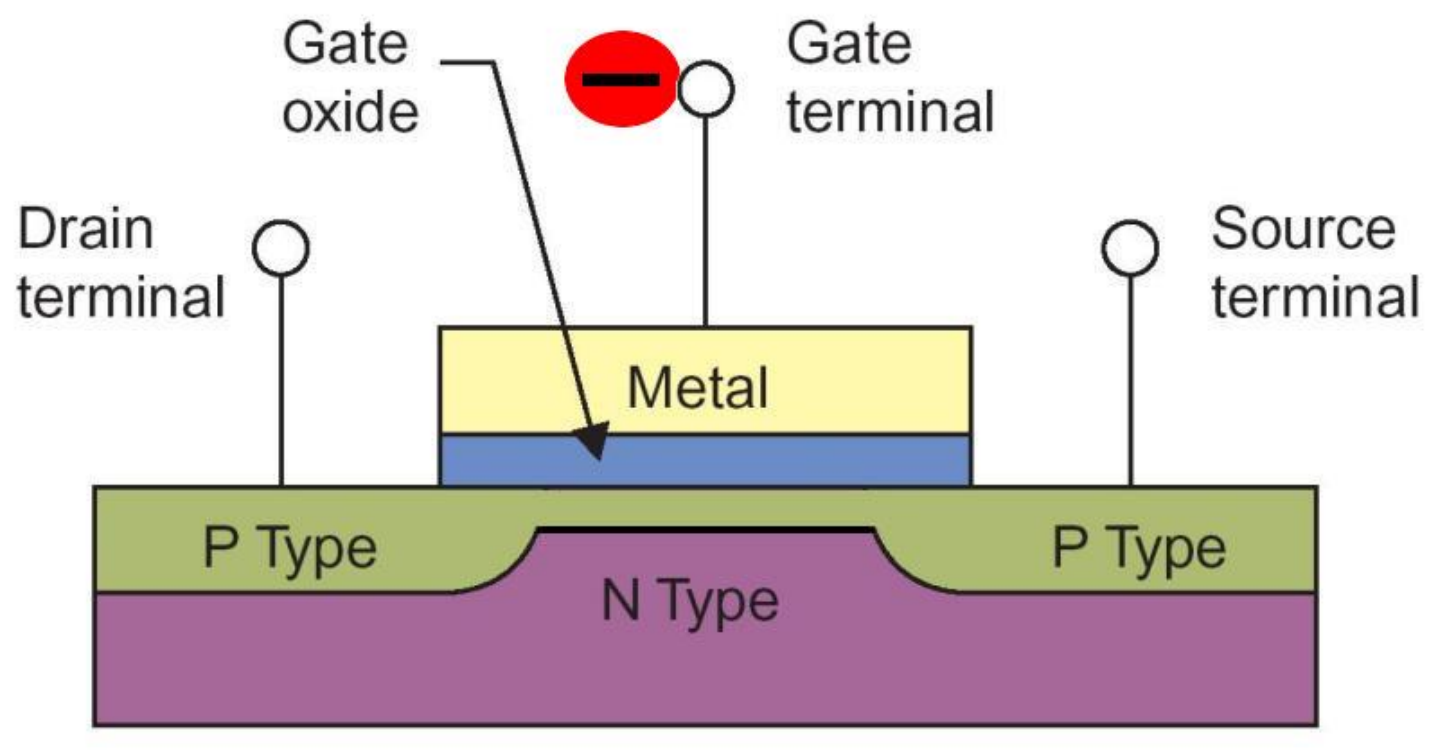
Tranzistor

- MOSFET, **M**etal-**O**xide-**S**emiconductor **F**ield-**E**ffect **T**ransistor



Metal

Tranzistor



$$P = C \times V^2 \times F$$

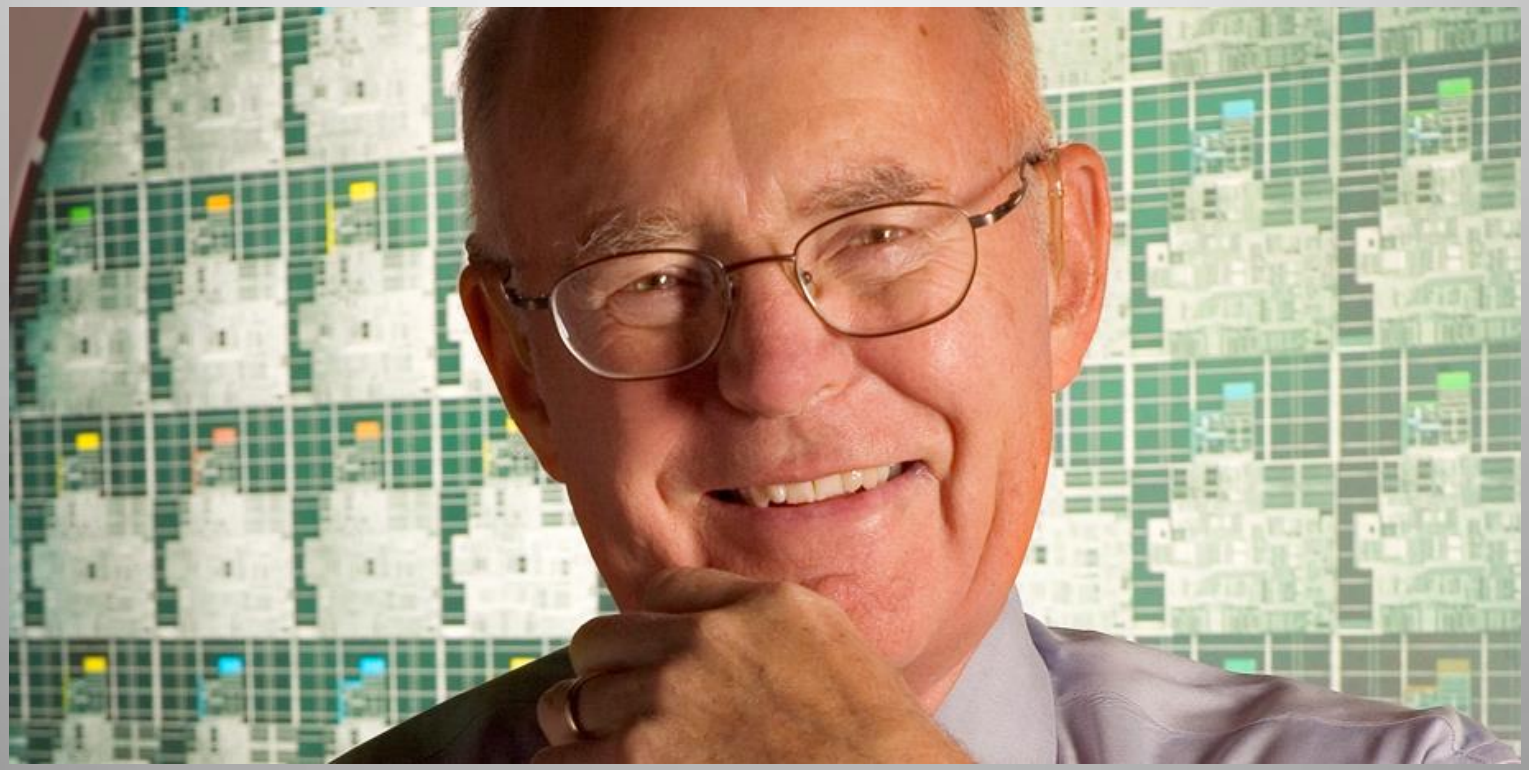
Intel Tejas & Jayhawk

- 2004
- Založené na Pentium 4 (Tejas) a Xeon (Jayhawk)
- 65nm výrobný proces
- 7 GHz+ operačná frekvencia
- Zrušené kvôli vysokej spotrebe, nahradené Dual Core

Vývoj tranzistorov - Scaling

- Dennardovo škálovanie
  - Klasické, lacné
  - Hrúbka oxidu sa zmenšuje spolu s ostatnými rozmermi
- Limity zmenšovania tranzistorov
  - 2005: Intel 65nm – hrúbka oxidu 1.2nm ~ 1 molekula SiO_2
  - Tunelovanie cez vrstvu oxidu – úniky spotreby
  - Ďalšie úniky spotreby
  - Zvyšovanie teploty – teplý vodič = zlý vodič
- Moore-ov Zákon

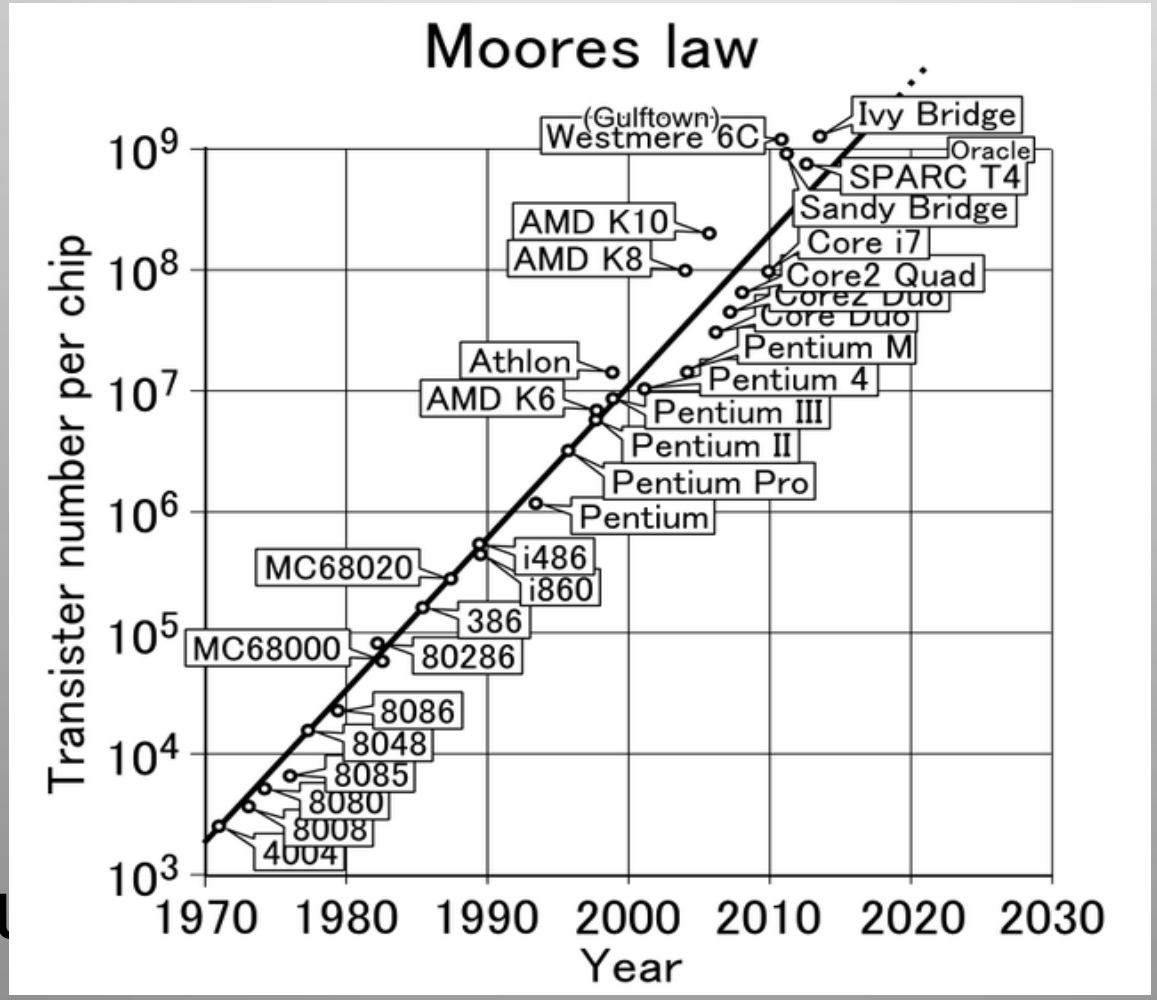
Gordon Moore (tiež)



Cramming More Components onto Integrated Circuits, 1965

Moore-ov zákon

- Počet tranzistorov na čipe sa zdvojnásobí každých 24 (18) mesiacov



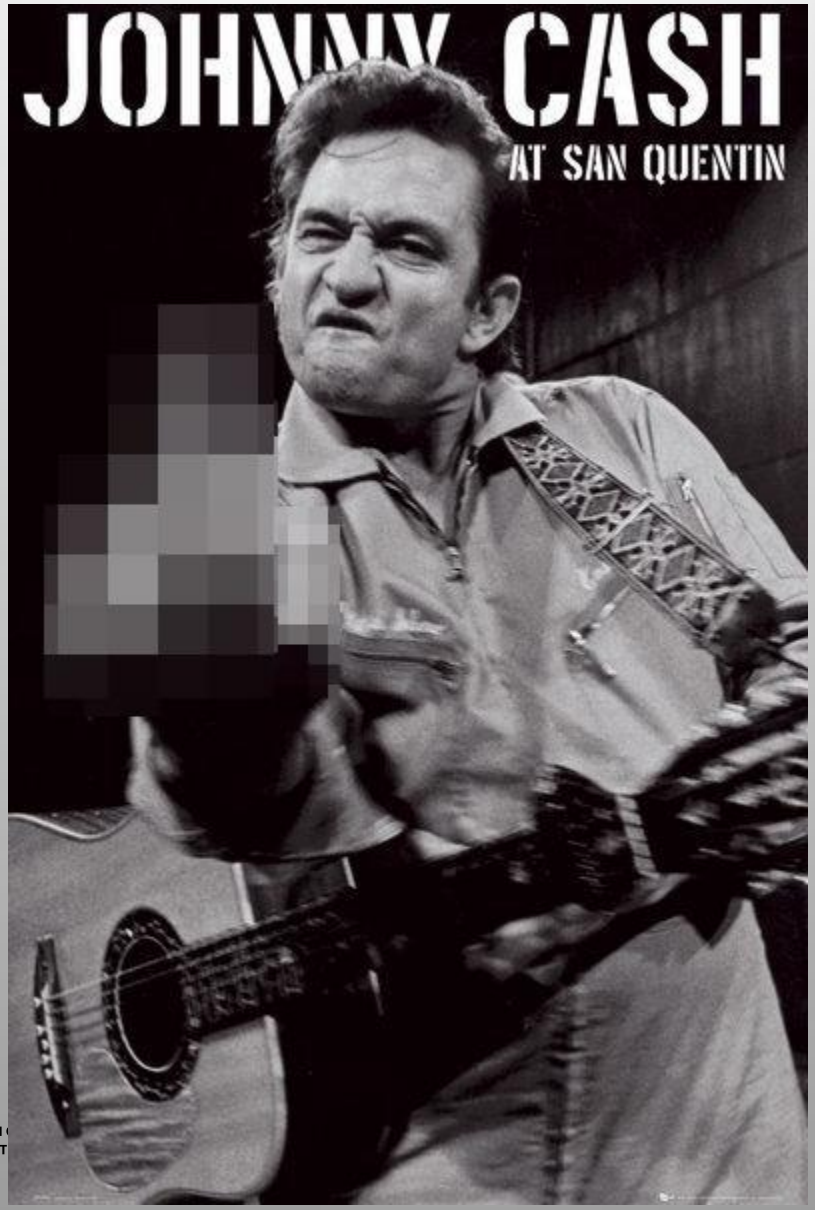
Zhrnutie

- Frekvencia limitovaná spotrebou, fyzikálnymi limitmi
- Zmenšovanie tranzistorov už nezvyšuje frekvenciu
- Počet tranzistorov na čipe ale stále rastie
- Ideálne pre viacero CPU jadier
- Zabezpečenie zvyšovania výkonu voči rastúcim požiadavkám trhu

Viacjadrové CPU

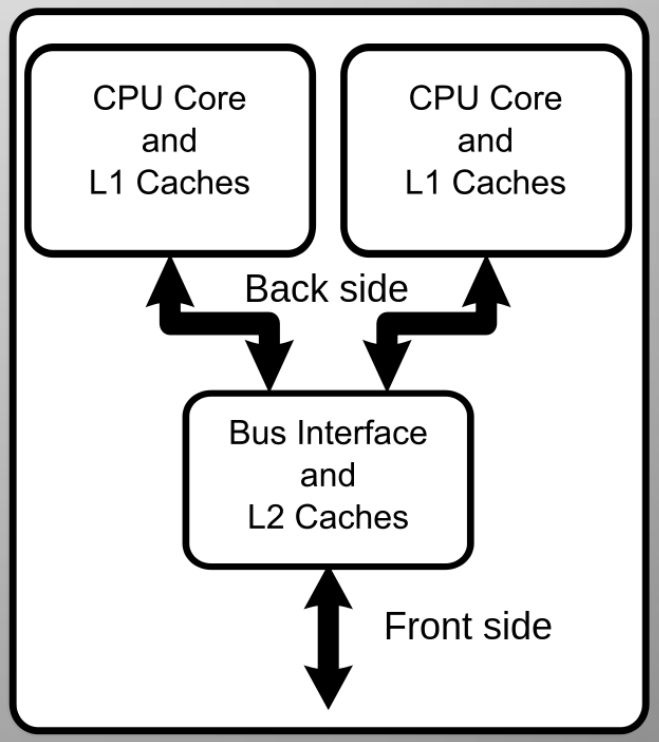
- Paralelizmus existoval už v sériových CPU!
  - Pipelining
  - Superskalárne CPU
- V multicore jadrá fungujú ako samostatné procesory
  - Oneskorenie pamäte kompenzujú vlastnými L1 Cache pamäťami

Cache



Multicore architektúry

- Jadrá fungujú ako samostatné procesory
  - Oneskorenie pamäte kompenzujú vlastnými L1 Cache pamäťami
- Ale zdieľajú niektoré časti
  - Napr. Spoločná L2 Cache



Limity paralelného spracovania

- Hlavný limit – sekvenčná časť kódu
  - Napr.: Kód sa vykonáva 10s pričom sekvenčná časť 5s a paralelná tiež 5s.
    - Minimálna doba vykonávania = 5s
    - Teda maximálne zrýchlenie x2 :(!!

- Programátor musí určiť paralelnú časť kódu, náročnejšie programovanie a debugging

- Takmer každý (bežný) kód obsahuje nejakú sekvenčnú časť

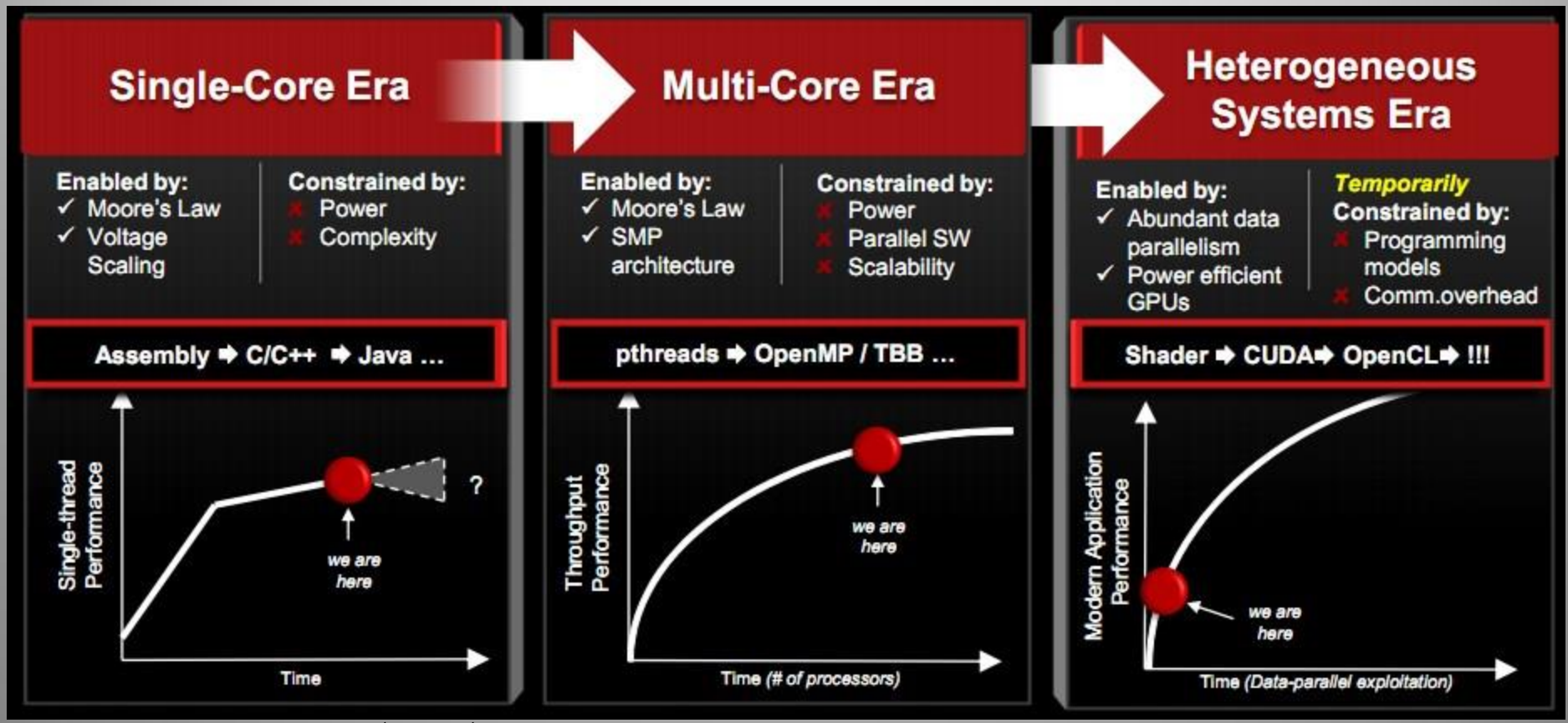
“Tmavý kremík” (Dark Silicon)

- Časť čipu ktorá musí byť trvale neaktívna – inak by sa prehrial CPU
- Zvyšovanie počtu tranzistorov s takmer nemennou spotrebou a frekvenciou
- Odhaduje sa že pri 8nm:
  - bude 50% čipu tmavý kremík
  - bude stačit 35 CPU jadier na dosiahnutie 90% zrýchlenia
  - Bude maximálny využitený počet CPU jadier 448
- Viac jadier ≠ väčší výkon (AMD Bulldozer)

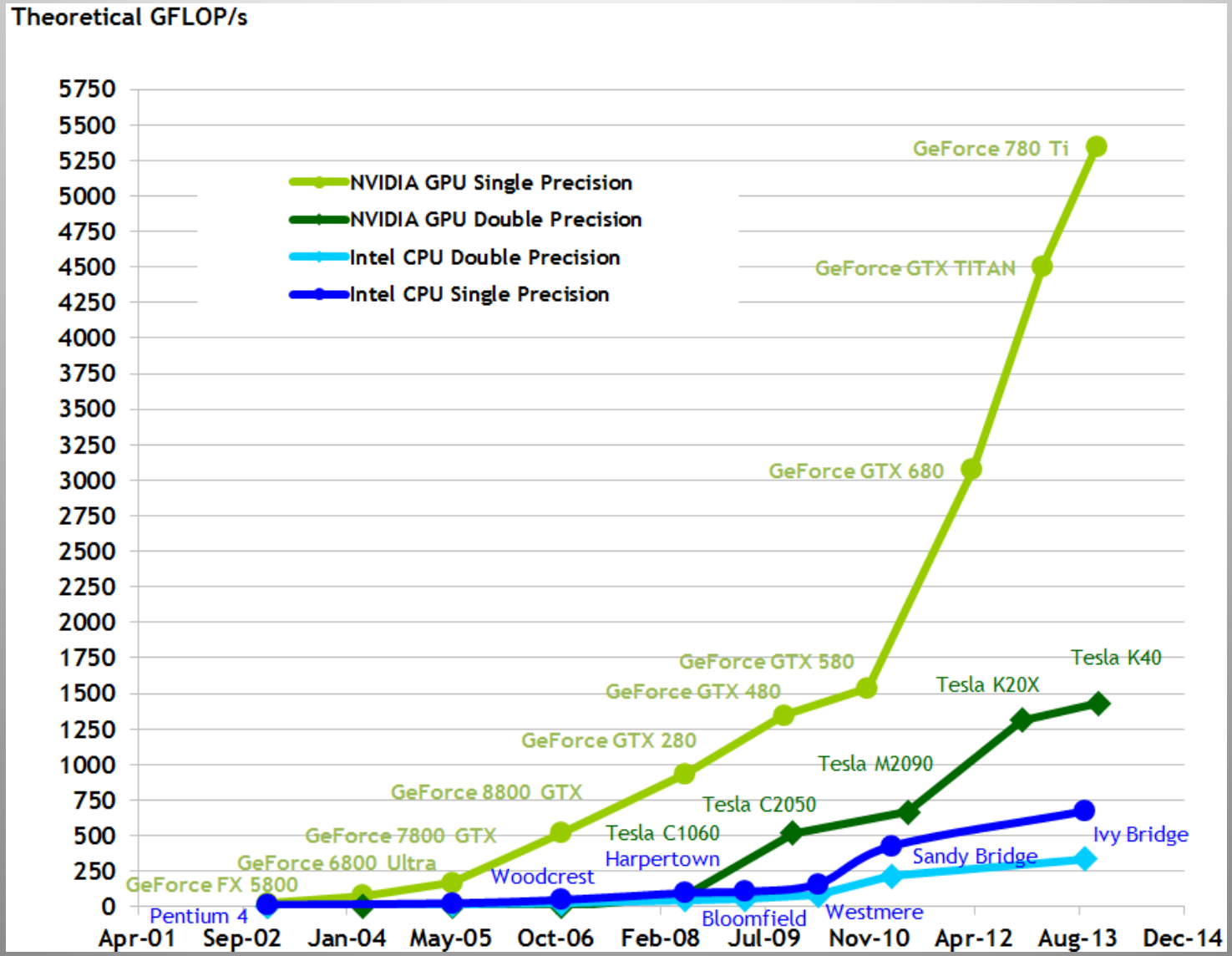
Distribuované počítanie

- Úloha sa rozdelí medzi viacero uzlov v sieti
  - Cluster – lokálna sieť, často symetrické uzly (Beowulf)
  - Grid – internet, rôzne uzly
- Vhodné pre nízku údajovú závislosť, komunikácia medzi uzlami je limitujúcim faktorom
- MPI – Message Passing Interface (C, C++, Java, Fortran)
- PVM – Paralell Virtual Machine (C, C++, Fortran)

Čo teraz a čo ďalej



NVIDIA Performance



Záver

- Koniec zvyšovania frekvencie CPU
- Rozšírenie paralelizmu, na úkor programátora
- Aj multicore má svoje limity
- Úzko špecializovaný a paralelný HW?
- More than Moore
  - Svet je predsa analógový
- Je procesor vlastne pomalý?

Ďakujem za pozornosť

peter.pistek@stuba.sk