



Fakulteta za
informacijske študije
Faculty of information studies

Univerza v Ljubljani
Fakulteta *za strojništvo*



Kreativno jedro:
Simulacije
Creative core: Simulations

Kampus šola HPC 2014

Predstavitev HPC

Matjaž Depolli
Odsek za komunikacijske sisteme
Inštitut „Jožef Stefan“



Naložba v vašo prihodnost
OPERACIJO DELNO FINANCIRA EVROPSKA UNIJA
Evropski sklad za regionalni razvoj



REPUBLIKA SLOVENIJA
**MINISTRSTVO ZA IZOBRAŽEVANJE,
ZNANOST IN ŠPORT**

Pregled predavanja

- Kaj pomeni HPC
- Superračunalniki
- Top 500
- Merila hitrosti
- Gradniki superračunalnikov
 - Pospeševalniki
 - Povezovalne mreže
- Uporaba HPC

Visoko zmogljivo računalništvo (HPC)

Visoko zmogljivo računalništvo

=

Združevanje računskih zmogljivosti na način, ki omogoča mnogo večje skupne zmogljivosti, kot bi jih zmogli tipični namizni računalniki ali delovne postaje, z namenom reševanja velikih problemov v znanosti, podjetništvu, industriji, inženirstvu

•Superračunalniki

•Računalniki, ki so trenutno med najhitrejšimi na svetu

•Računalniki, ki so namenjeni za HPC

•Vzporedni računalniki

•Porazdeljeni računalniki

•petaFLOPS

Super računalniki



Cray 1
Blue Gene
Tianhe-2
Beowulf cluster



Prvi superračunalnik

Cray 1

- 1976
- Temelji na integriranih vezjih
 - Okoli 200.000 logičnih vrat (NOR z 4 ali 5 vhodi)
- 64-bitni sistem
- 8 MB pomnilnika
- 80 MHz
- Max 250 MFLOPS
- 115 kW električne moči (hlajenje ni upoštevano)



Top 500

Projekt, ki rangira svetovne računalnike po hitrosti

- Od leta 1993
- Rangiranje 2x letno
- Štejejo le lokalni skupki
 - (CERN-ov WLCG torej ne šteje)
- LINPACK za merilo
 - Rezultat v FLOPS (Floating-point Operations Per Second)

Slovenija:

- Turboinštitut »Adria«
 - 72. mesto 2008
 - 49.2/36.8 teraFLOPS, 162 kW
- ?

Vrh iz Top 500



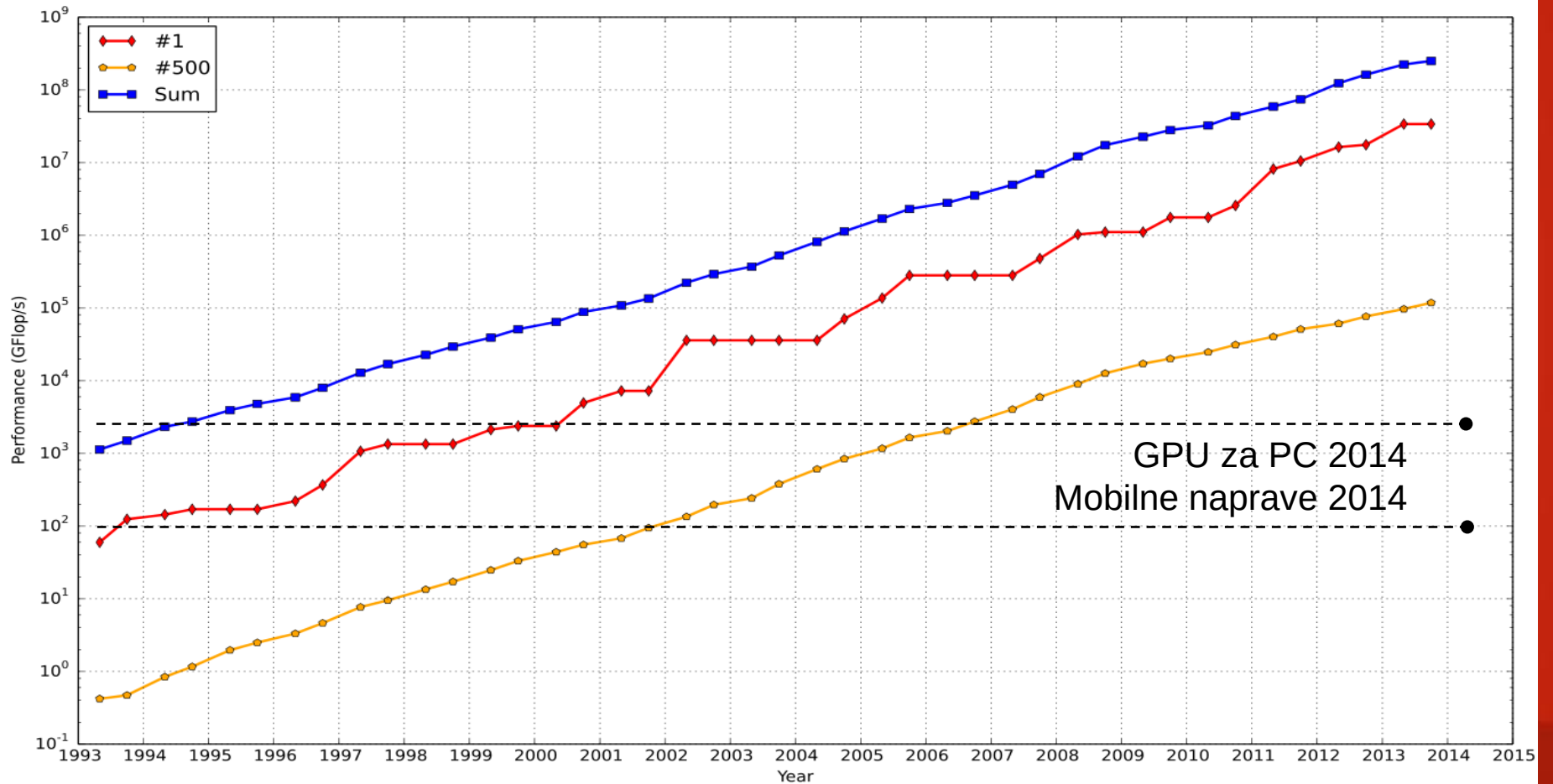
- Tianhe-2 (Kitajska)

- 16.000 računskih vozlišč
 - 2x Intel Xeon E5-2692 (12 jeder)
 - 3x Xeon Phi 31S1P (57 jeder)
 - 88 GB RAM
 - Povezave?
- 3.120.000 jeder
- Teoretično 54,9 petaFLOPS
- Doseženo 30.65 petaFLOPS
- 17.6/24 MW električne moči

.HE Dravograd



Napredek



Merilo hitrosti

- LINPACK test
 - Sistem linearnih enačb (gost)
 - Reševanje po principu LU dekompozicije
 - Operandi s plavajočo vejico (double)
 - Velikost sistema je poljubno nastavljiva
 - Problem se prilagodi na dan računalnik
 - Program se prevede na danem računalniku (sme se tudi spremeniti implementacijo)
 - Rezultat je razmeroma blizu teoretičnemu maksimumu

Drugi superračunalniki?

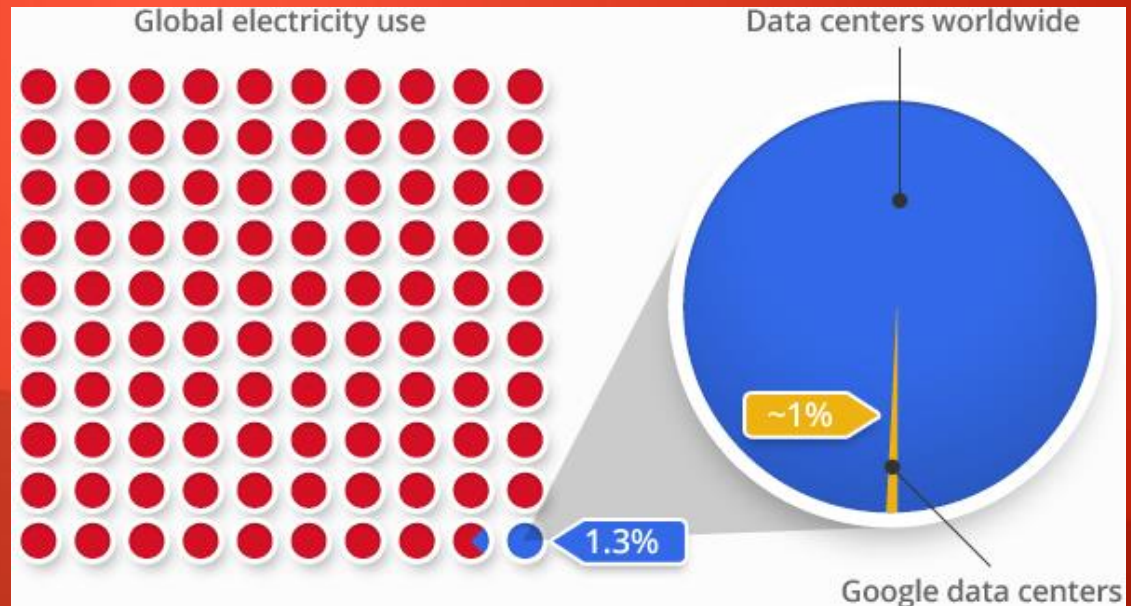
- Nekateri namenski superračunalniki v LINPACKu ne dosegajo dobrih rezultatov
- WLCG (Worldwide LHC Computing Grid) je narejen za shranjevanje, distribucijo in analizo 15 PB na leto (od ~2002 dalje) [15 PB/y = ~0.5 GB/s]
- Druga merila za računalnike:
 - Performanca na enoto električne moči
 - Performanca na prostorsko enoto
 - Performanca na denarno enoto
 - Količina obdelanih podatkov na sekundo

Green 500

- Seznam 'zelenih' superračunalnikov
- Enota: MFLOPS/W
- Trenutno visoko uvrščeni (1/2 2014) tipično vsebujejo grafične pospeševalnike NVIDIA Tesla K20X
- TSUBAME-KFC (Tokyo Institute of Technology)
 - 4389.82 MFLOPS/W
 - 1901.54 MFLOPS/W (za primerjavo Tianhe-2, 49. mesto)

Primer: Google

- Google uporablja strojno opremo z najboljšo performanco na denarno enoto (vključuje ceno nakupa, vzdrževanja in električne energije)
- Merjenje njihove opreme s faktorjem efektivne uporabe energije
- Dosegali naj bi faktor 1.12 (le 12% energije ni izkoriščene za računanje)



Gradniki superračunalnikov

- Samostojne enote z večjedrnimi mikroprocesorji in pomnilnikom
- Komunikacijska oprema – omrežje
- Pospeševalniki – dodatne procesorske enote (splošne, namenske, poenostavljene)
- Enote za hranjenje podatkov (diskovne)
- Napajanje in hlajenje
 - Hlajenje s klimatskimi napravami doda ~30% k porabi energije
 - Že Cray 1 je bil zgrajen okoli hladilne naprave

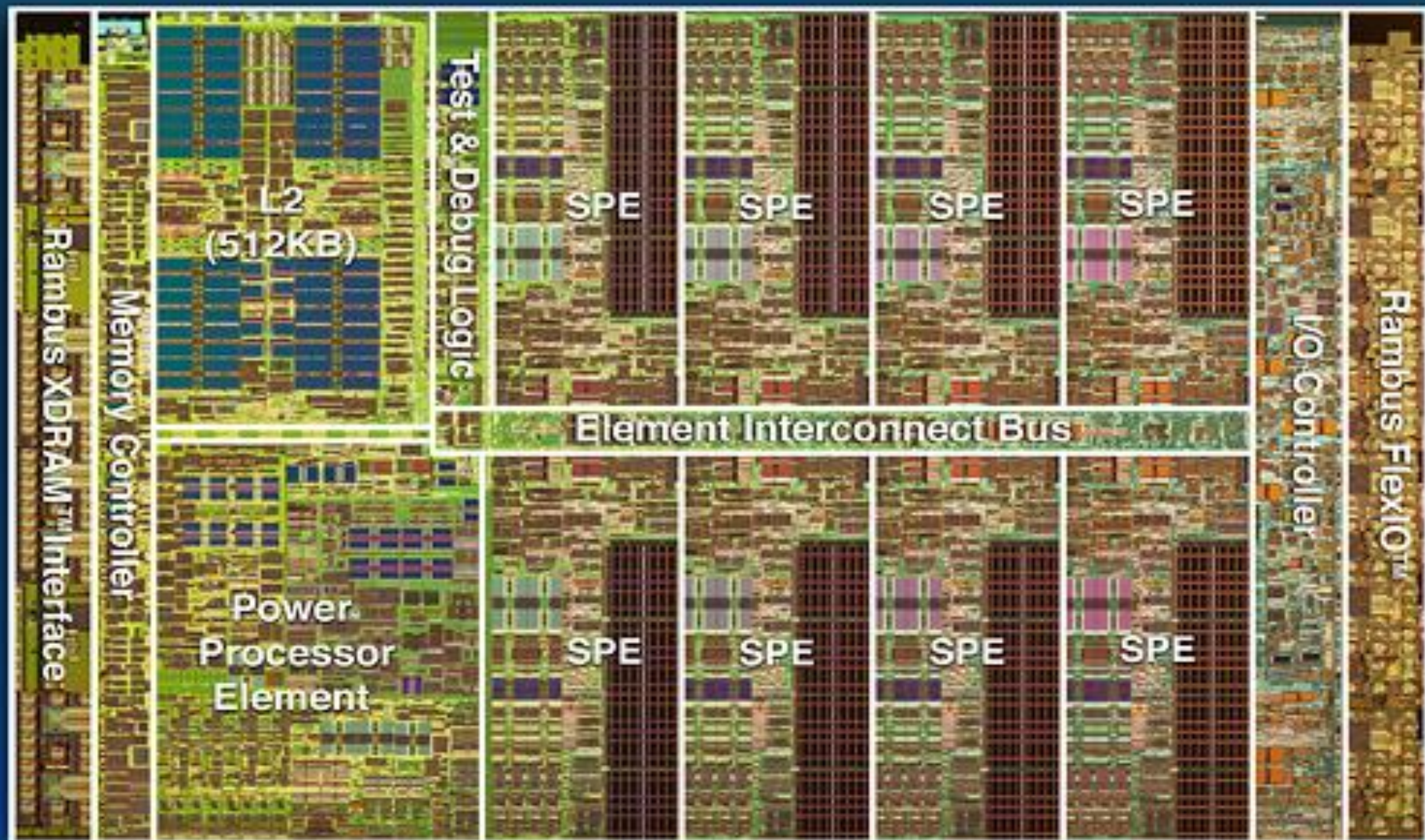
Dva pomembna tipa računalnikov

- SMP (Symmetric MultiProcessing)
 - Tipično večjedrni procesorji
 - Vsi procesorji so enaki (tipično jih je malo)
 - Skupen pomnilnik
 - služi tudi za komunikacijo
 - Lahko postane ozko grlo
- NUMA (NonUniform Memory Access)
 - Več lokalnih pomnilnikov
 - Lokalen dostop do pomnilnika je hitrejši od oddaljenega
 - Primer so pospeševalniki z lokalnim pomnilnikom (GPU)
 - Počasna globalna komunikacija

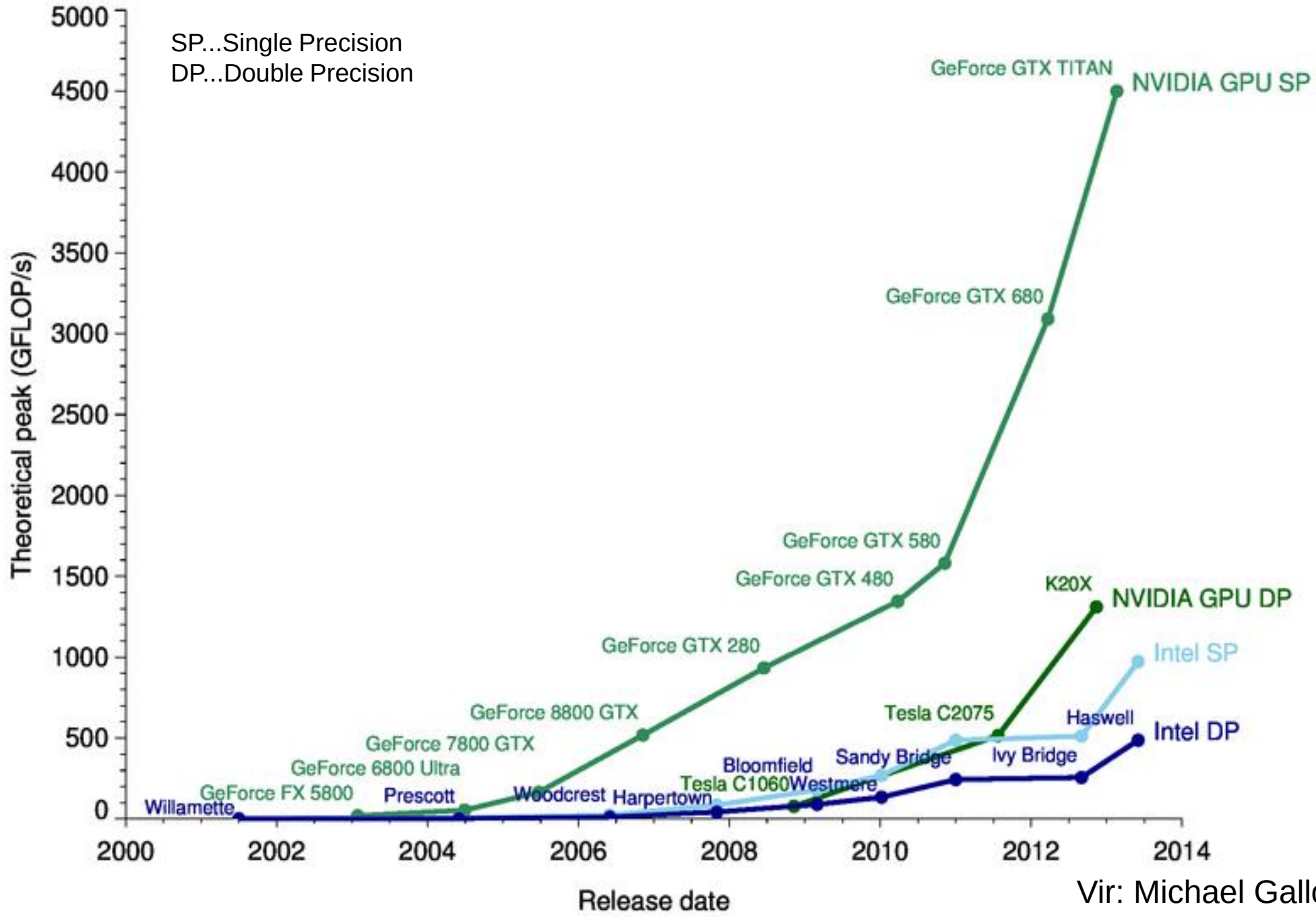
Pospeševalniki

- Arhetip NUMA arhitekture
- Začelo se je s IBM Cell procesorji (PS 3)
 - Začetki leta 2008
 - PS3
 - Glavno jedro kontrolira 8 'sinergijskih' poenostavljenih jeder
- Nadaljevanje z GPU karticami
 - Sprva single-precision, omejene možnosti programiranja
 - CUDA, OpenCL, OpenACC
 - Primer: Tesla K20X – 2688 jeder, 732 MHz, ~6 GB RAMa, 235 W TDP

Cell Broadband Engine Processor



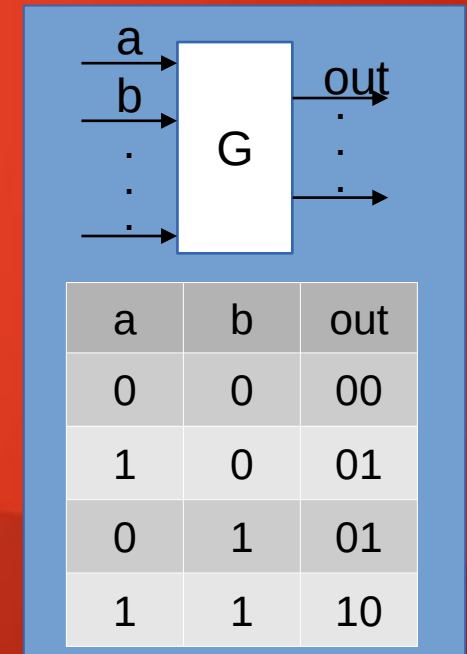
SP...Single Precision
DP...Double Precision



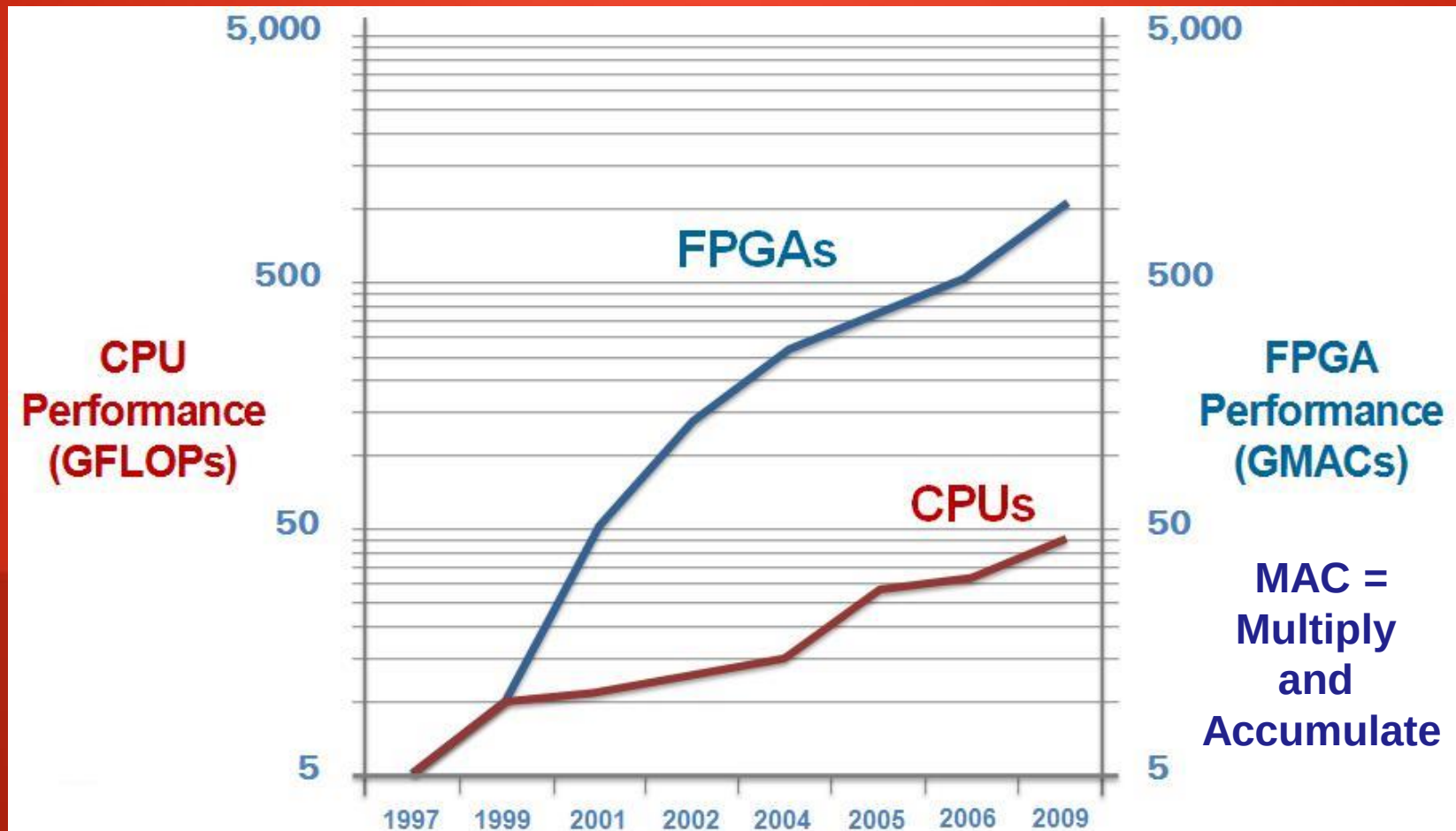
Vir: Michael Galloy

Pospeševalniki

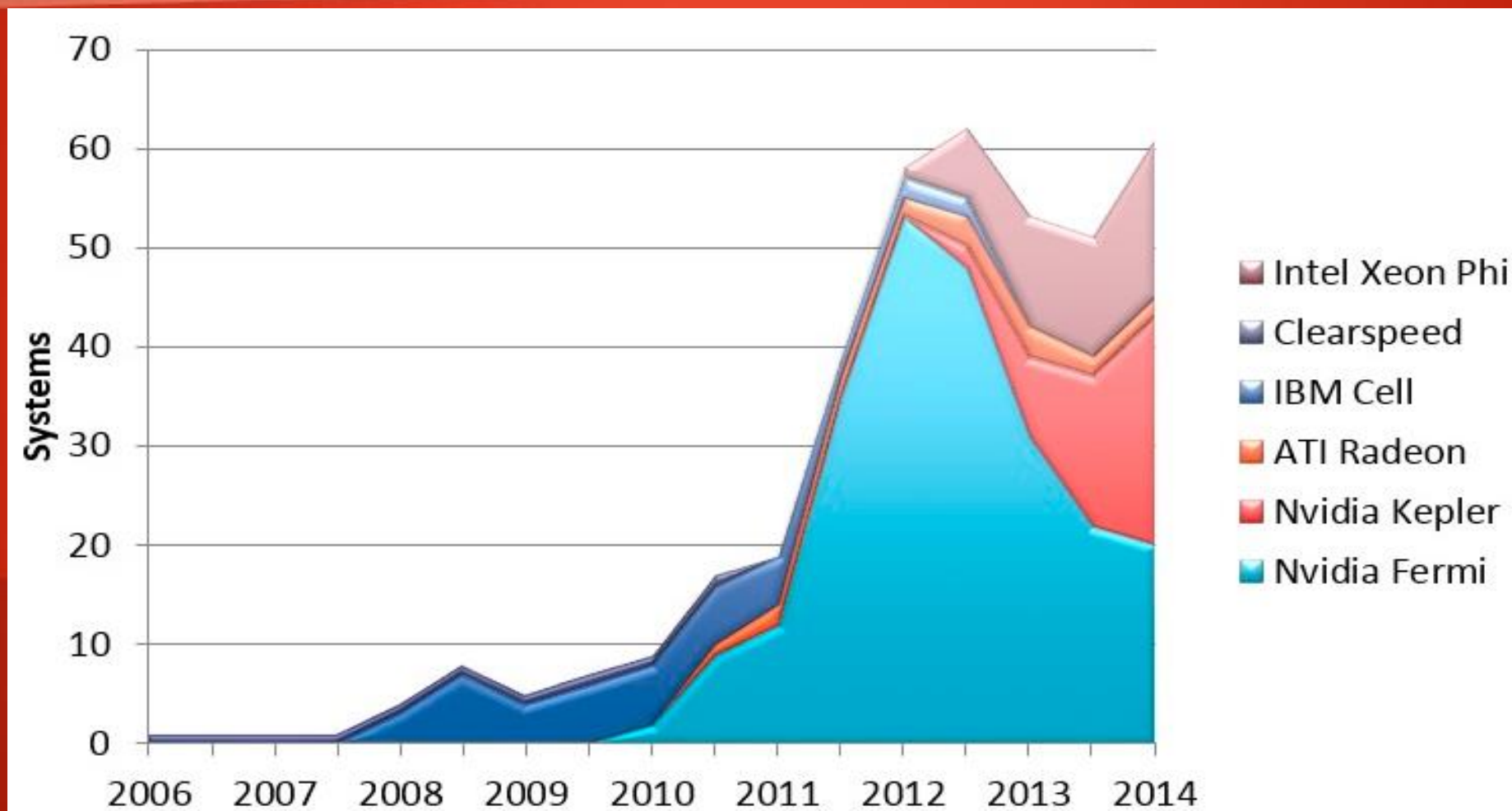
- Razvoj okoli FPGA
 - Field-Programmable Gate Array
 - Vsebujejo fiksne DSPje za hitre operacije tipa $a=a+b*c$
 - Lasten RAM
 - Nizke frekvence
- Moderni koprocesorski sistemi
 - Intel Xeon Phi:
 - ~60 x86-64 jeder (many-core)
 - Do 32 GB RAMa



Zmogljivost FPGA pospeševalnikov

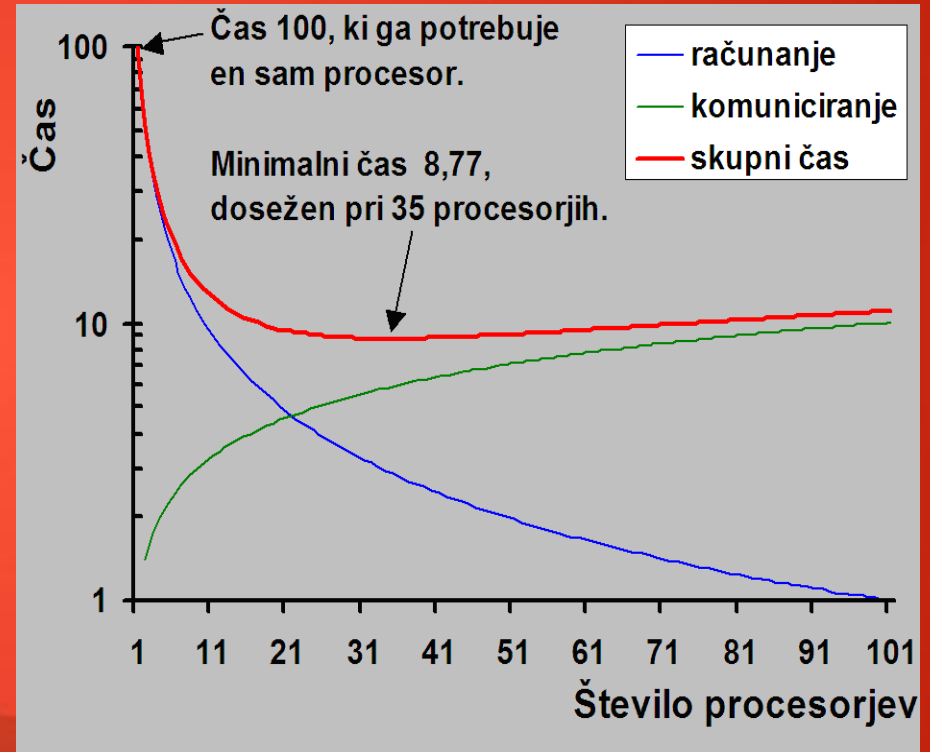


Pospeševalniki v Top 500



HPC ni le računanje

- Računanje + komunikacija
- Komunikacije ni pri zaporednih programih; pri vzporednih predstavlja 'overhead'
- Poleg komunikacije se prikrade še čakanje (na komunikacijo, na zadnje podatke, na proste vire, ...)



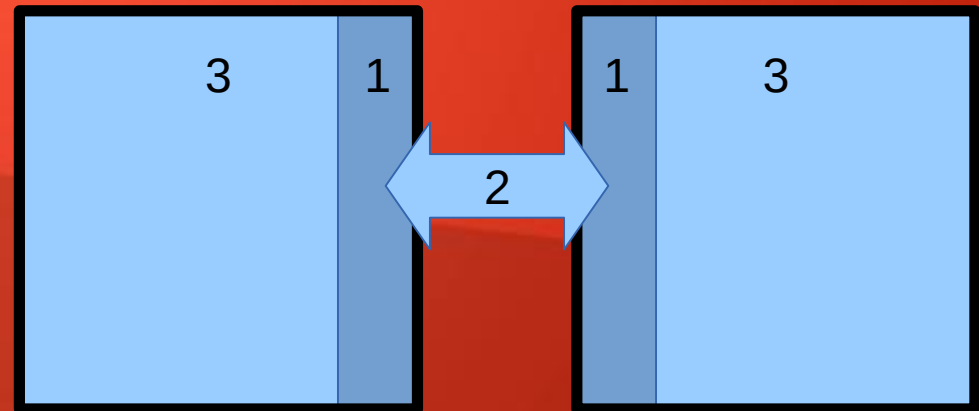
Prekrivanje računanja in komunikacije

- Delna samostojnost perifernih enot
 - Pospeševalniki
 - Mrežna oprema
- Organizacija računanja in komunikacije
 - Procesor si zagotovi podatke od sosedov še preden jih potrebuje
 - Procesor najprej izračuna tisti del podproblema, ki ga mora poslati sosedu
- Procesor samo odda naročilo za pošiljanje ali prejemanje
- Računanje in komunikacija se odvijata hkrati (sta časovno prekrita)

Prekrivanje računanja in komunikacije

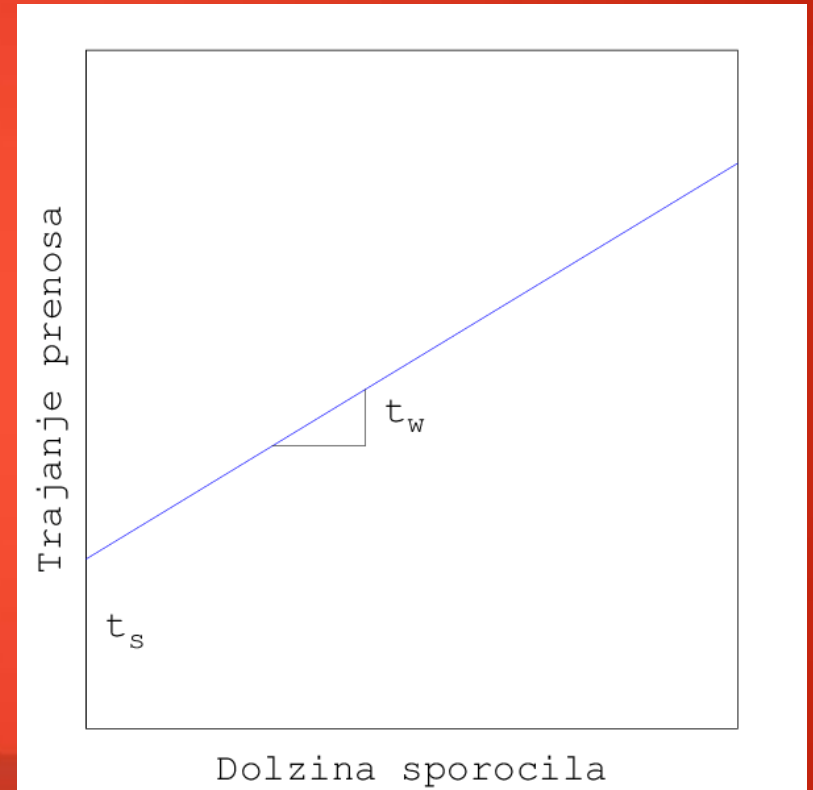
- Dva dela domene
 - Neodvisno računanje vsako iteracijo
 - Pred naslednjo iteracijo je treba del domene poslati drugemu procesorju
- Naivno iteriranje:
 - Računanje celotne domene
 - Pošiljanje/sprejemanje dela domene

- Hitrejše iteriranje:
 - Računanje dela domene
 - Asinhrono pošiljanje in sprejemanje dela domene
 - Računanje ostale domene
 - Zaključek komunikacije



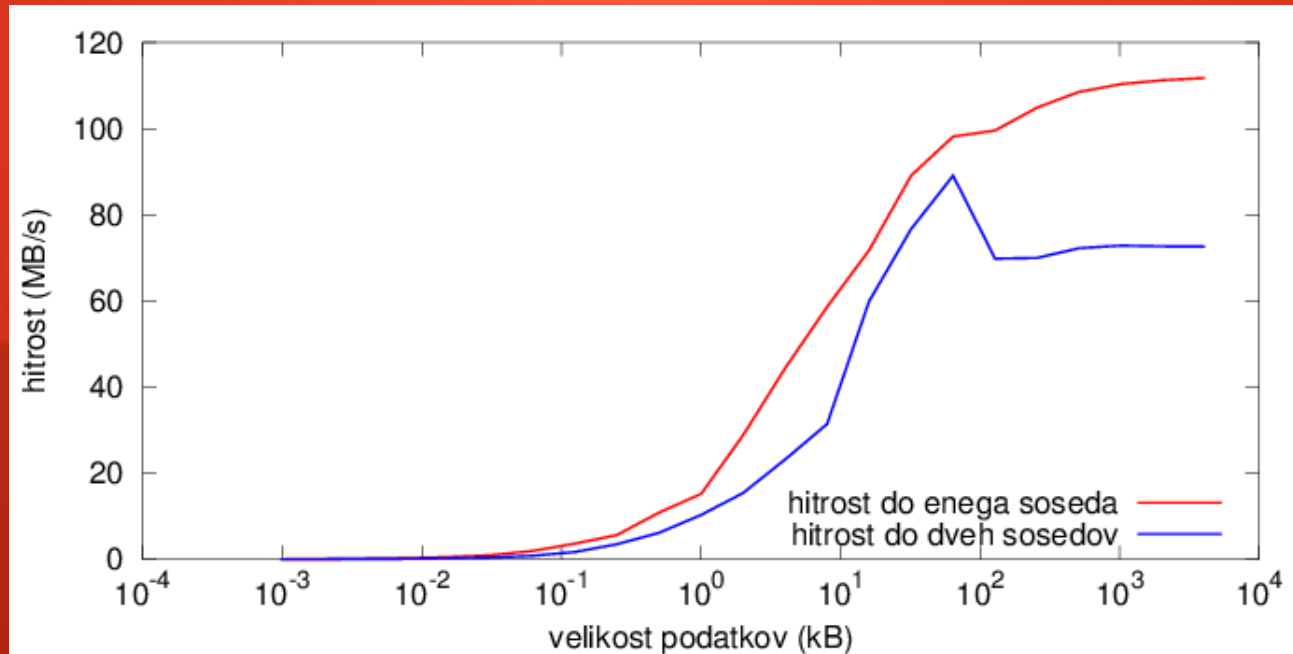
Model komunikacije

- t_s je čas potreben za začetek prenosa (startup)
- t_w je čas potreben za prenos ene besede (en bajt ali nekaj bajtov)
- L je dolžina sporočila
- $t_{\text{msg}} = t_s + L * t_w$
- Majhna sporočila so draga zaradi velike latence
- Hitrost prenosa dolgih sporočil je odvisna le še od pasovne širine



Primer hitrosti komunikacije

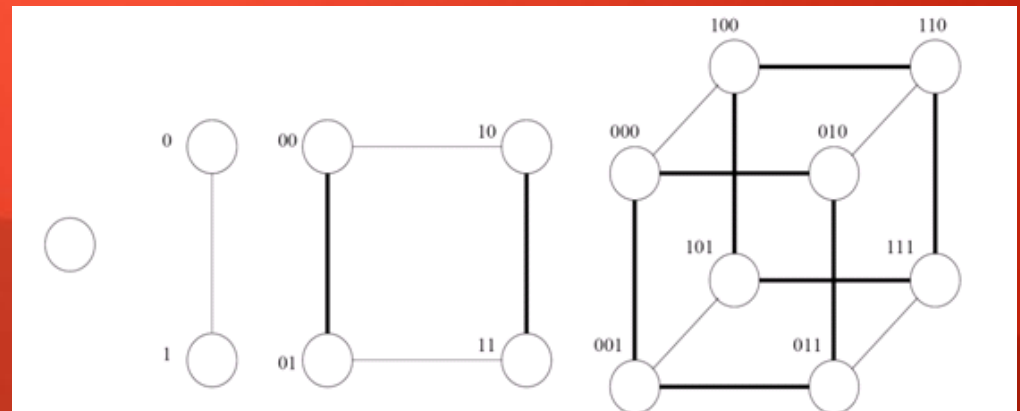
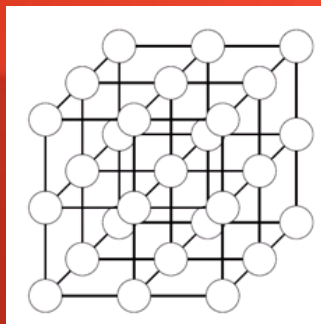
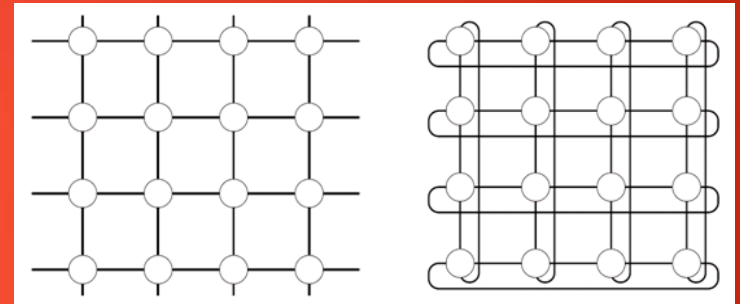
- 1 Gb Ethernet
- Pošiljanje sporočil z MPI
- 1 ali 2 soseda, povezava prek mrežnega stikala



Povezovalne mreže

- Tipi:

- Obroč (ring)
- Zvezda
- Regularne mreže (2D torus, 3D torus, hiperkocka)
- Neregularne mreže



Lastnosti povezovalnih mrež

- Število vozlišč, ki jih povezuje
- Število povezav na vozlišče
- Uporaba dodatne mrežne opreme (stikala)
- Pasovna širina prereza
- Največja in povprečna razdalja med vozlišči

Programska oprema

- Neizkoriščenost računalnika je nezaželjena
- Uporabnikov je veliko
- Hkratna uporaba delov računalnika
- Različna okolja za različne uporabnike
- Oddaljen dostop
- Oddaljeno upravljanje

Operacijski sistemi

- Linux
- Ostale Unix varjante
- Microsoft

„Obvezna“ programska oprema

- Programski jeziki in knjižnice za izkoriščanje paralelnosti
 - MPI
 - OpenMP
- Orodja za oddaljeno upravljanje (posodabljanje sistema, dodajanje programske opreme, ugašanje/prižiganje, ...)
- Skupen datotečni sistem
- Razporejevalnik (scheduler) opravil
- Vzporedno programje
- Dostopne točke za uporabnike

Uporaba HPC

- Simulacije kompleksnih sistemov (vreme, klima, mehanske naprave, elektronska vezja, ...)
- Simulacije proizvodnih procesov, fizikalnih procesov, kemičnih reakcij, eksplozij, ...
- Obdelava velike količine podatkov (video, računalniški posebni efekti v filmih, risanke, genetika, internet, ...)
- Napovedi stohastičnih in kaotičnih procesov (finance, ekonomija, vreme, ...)
- Modeliranje in inverzni problemi (geo-modeliranje, računalniška tomografija, ...)