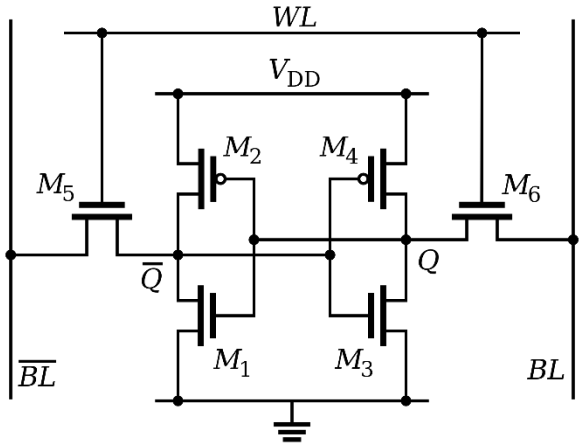
Paměťová hierarchie

Současné technologie polovodičových pamětí

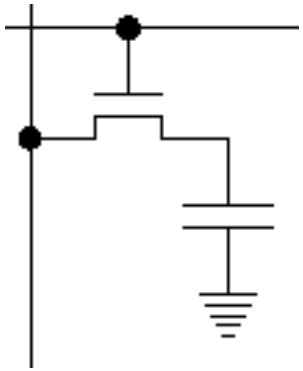
	Statická RAM	Dynamická RAM	Flash EPROM
Princip záznamu	Elektrický obvod se dvěma stabilními stavy	Nabitý kondenzátor	Nabitý kondenzátor
Výdrž záznamu	Po dobu napájení	Desítky milisekund	Desítky let
Princip čtení	Měření napětí na výstupu obvodu	Vybití kondenzátoru	Ovlivnění vodivosti elektrickým polem kondenzátoru
Princip zápisu	Změna napětí vstupů obvodu	Nabití kondenzátoru	Nabití/vybití kondenzátoru tunelováním
Tranzistorů na bit	6	1	1/3
Moorův zákon – dvojnásobná kapacita	2 roky	1,5 roku	1,4 roku
Latence	0.5-5 ns dle velikosti	10-20 ns	dle architektury
Moorův zákon – poloviční latence	?	> 7 let	?

Současné technologie polovodičových pamětí

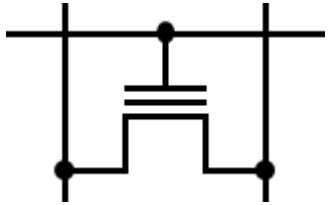
SRAM



DRAM

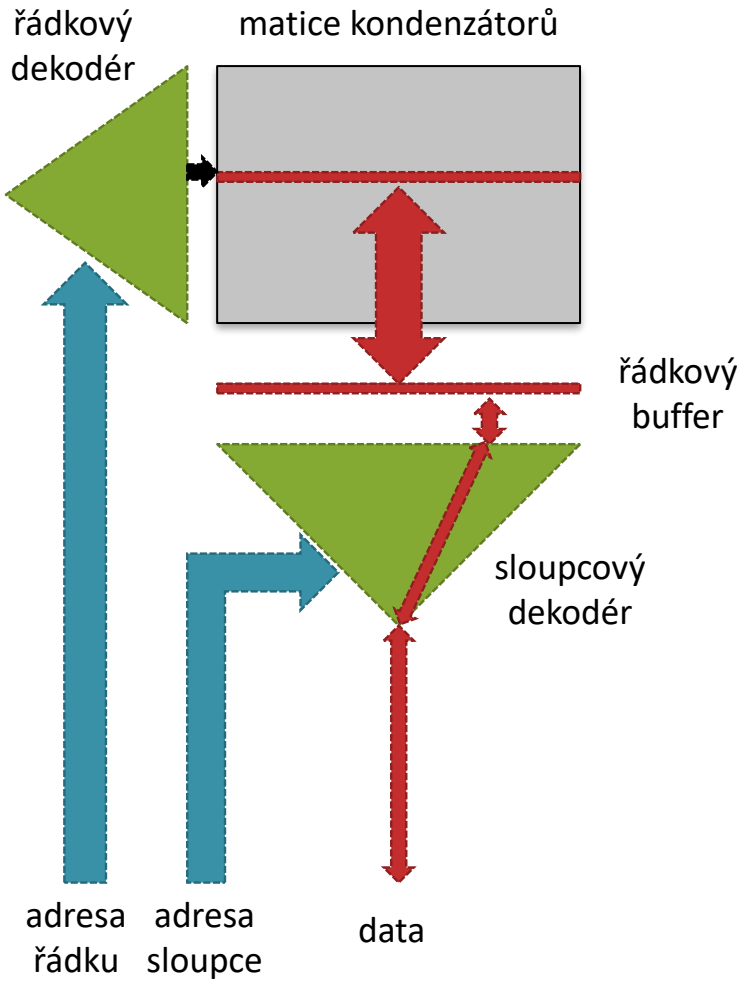


Flash



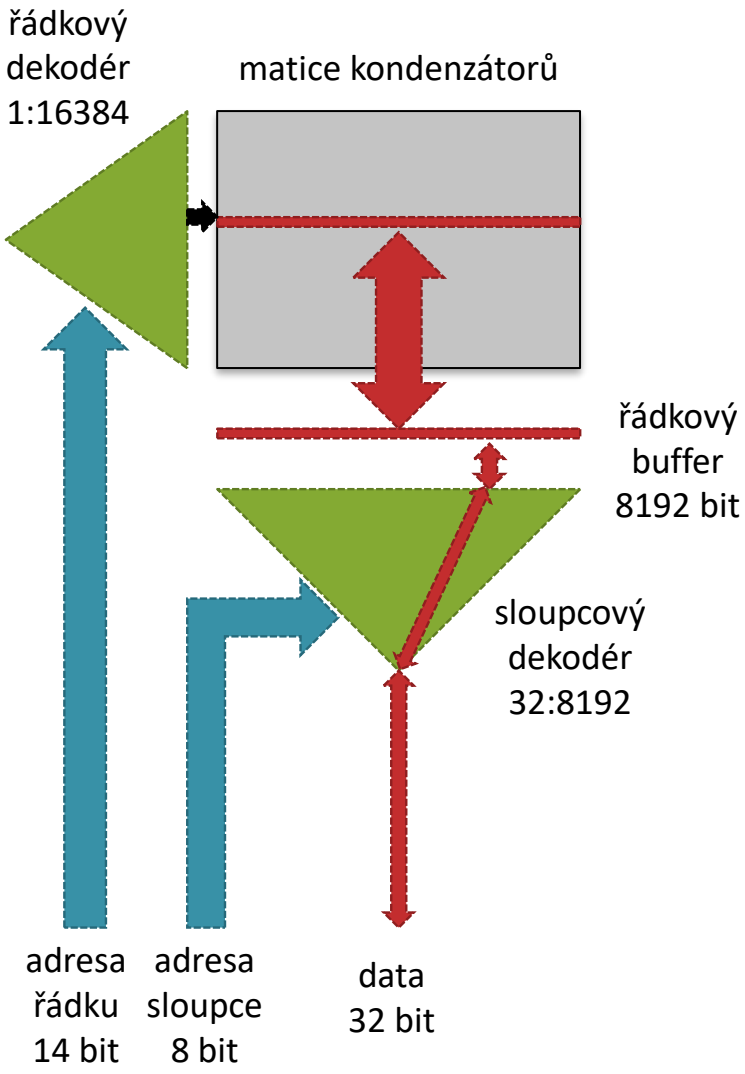
Dynamická RAM

Architektura paměti DRAM



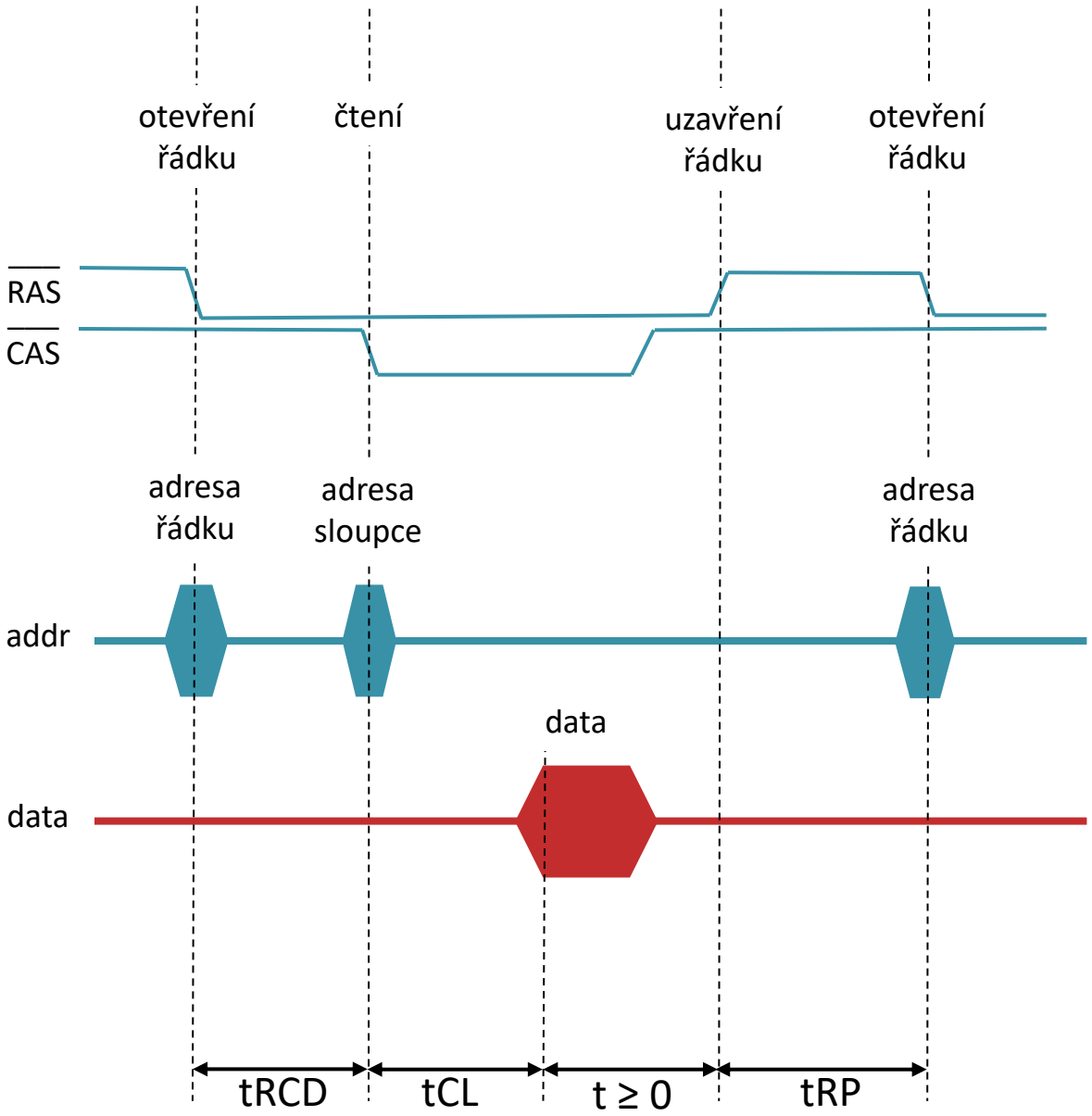
- Otevření řádku
  - Adresa řádku se dekóduje
  - Hodnota řádku se přenese do řádkového bufferu
    - Řádek se tím smaže!
- Čtení/zápis
  - Adresa sloupce se dekóduje
  - Bit v řádkovém bufferu se přečte/nastaví
- Zavření řádku
  - Hodnota z řádkového bufferu se zapíše do řádku
    - Slouží jako refresh celého řádku

Architektura paměti DRAM



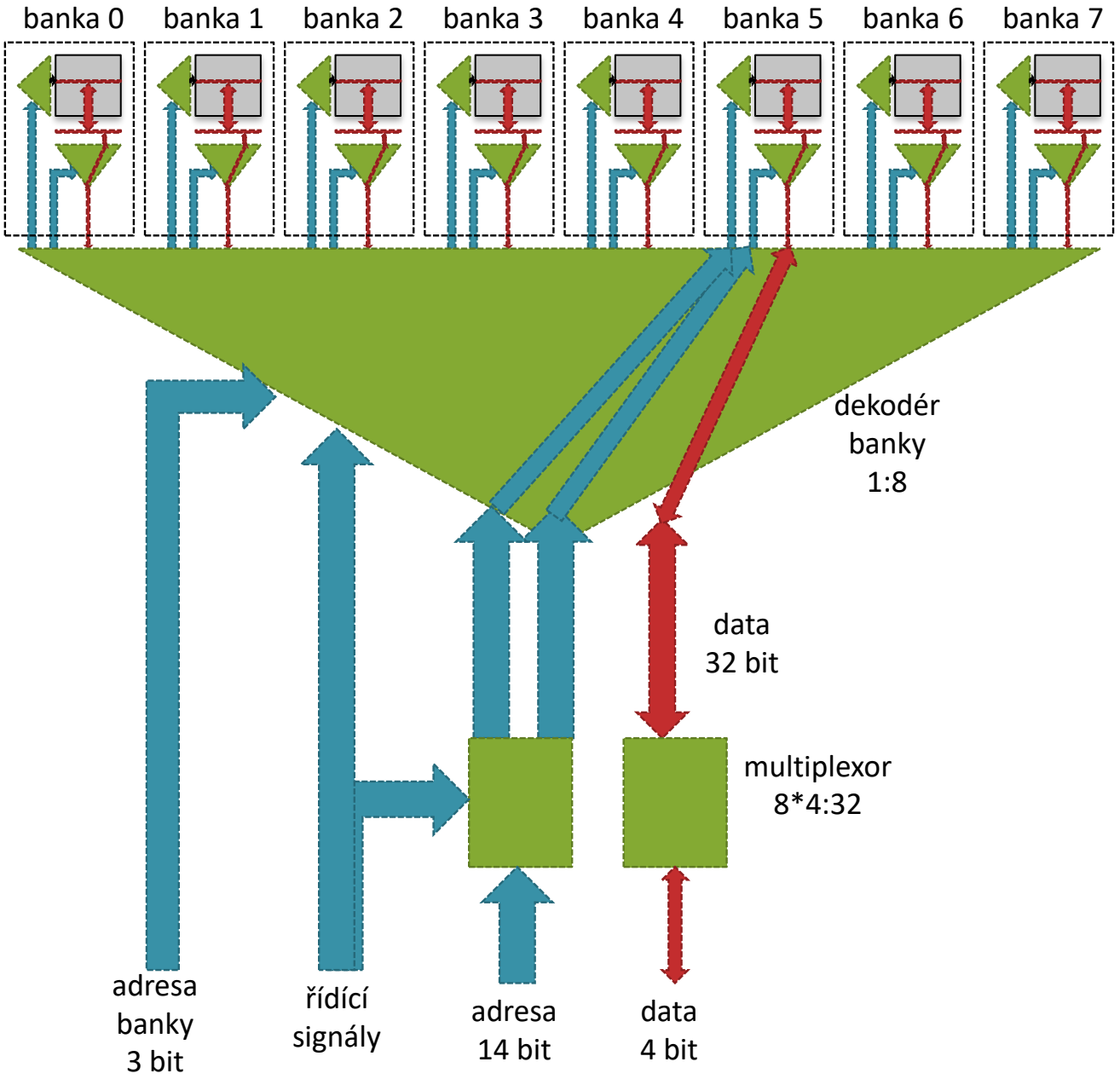
- Typické rozměry (2012)
  - 16384 řádků
  - 8192 sloupců
  - celkem 128 Mbit
- Typické časy
  - tRCD = 13 ns(2012) 10 ns(2020)
    - Otevření řádku
  - tCL = 13 ns(2012) 10 ns(2020)
    - Čtení/zápis
  - tRP = 13 ns(2012) 10 ns(2020)
    - Zavření řádku

Architektura paměti DRAM



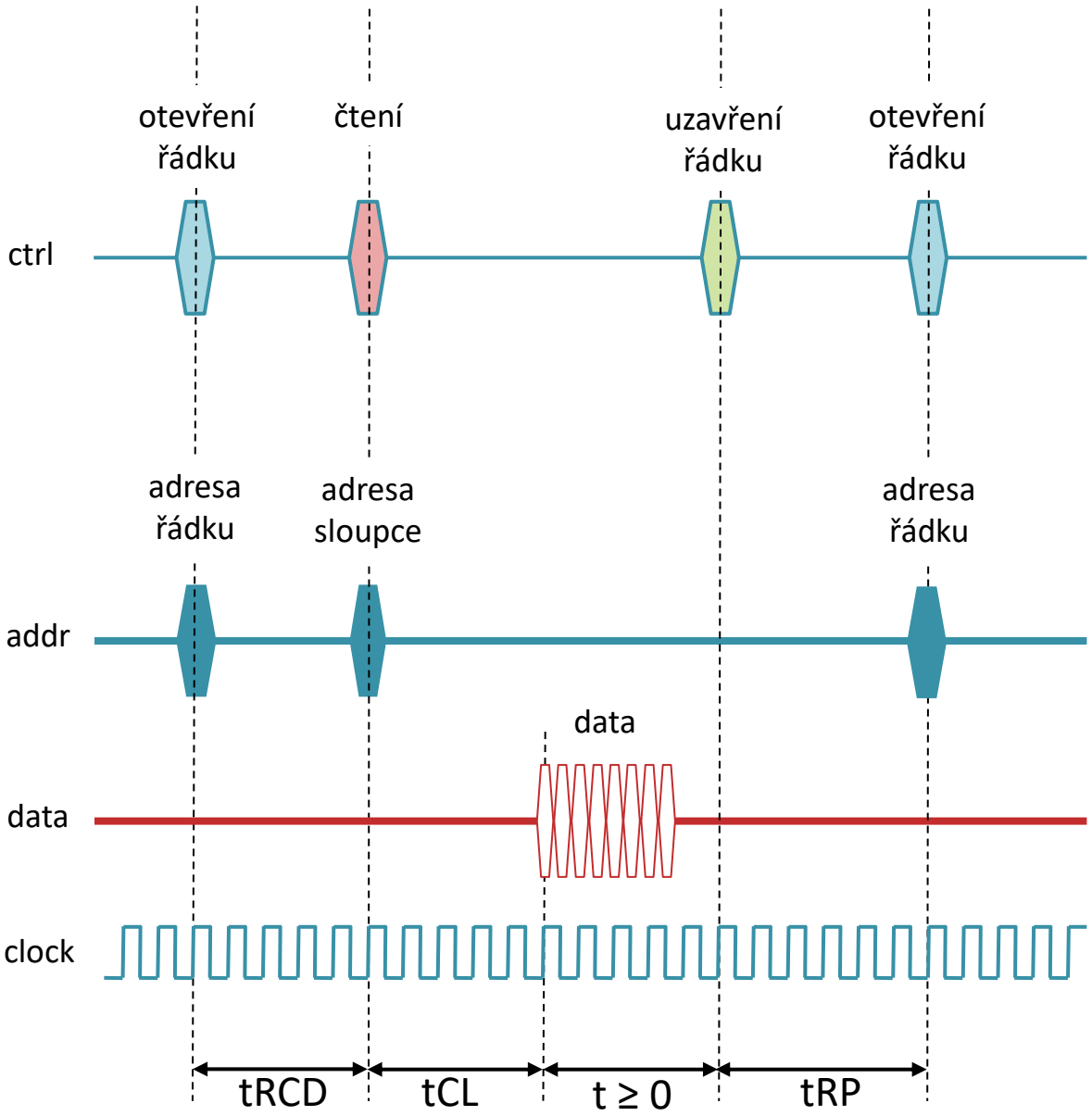
- Typické časy
  - tRCD = 13 ns(2012) 10 ns(2020)
    - Otevření řádku
  - tCL = 13 ns(2012) 10 ns(2020)
    - Čtení/zápis
  - tRP = 13 ns(2012) 10 ns(2020)
    - Zavření řádku
  - V rámci jednoho otevření řádku je možno provést více read/write operací se sloupci
    - V běžných CPU je takových příležitostí velmi málo

Architektura paměti SDRAM – příklad čipu



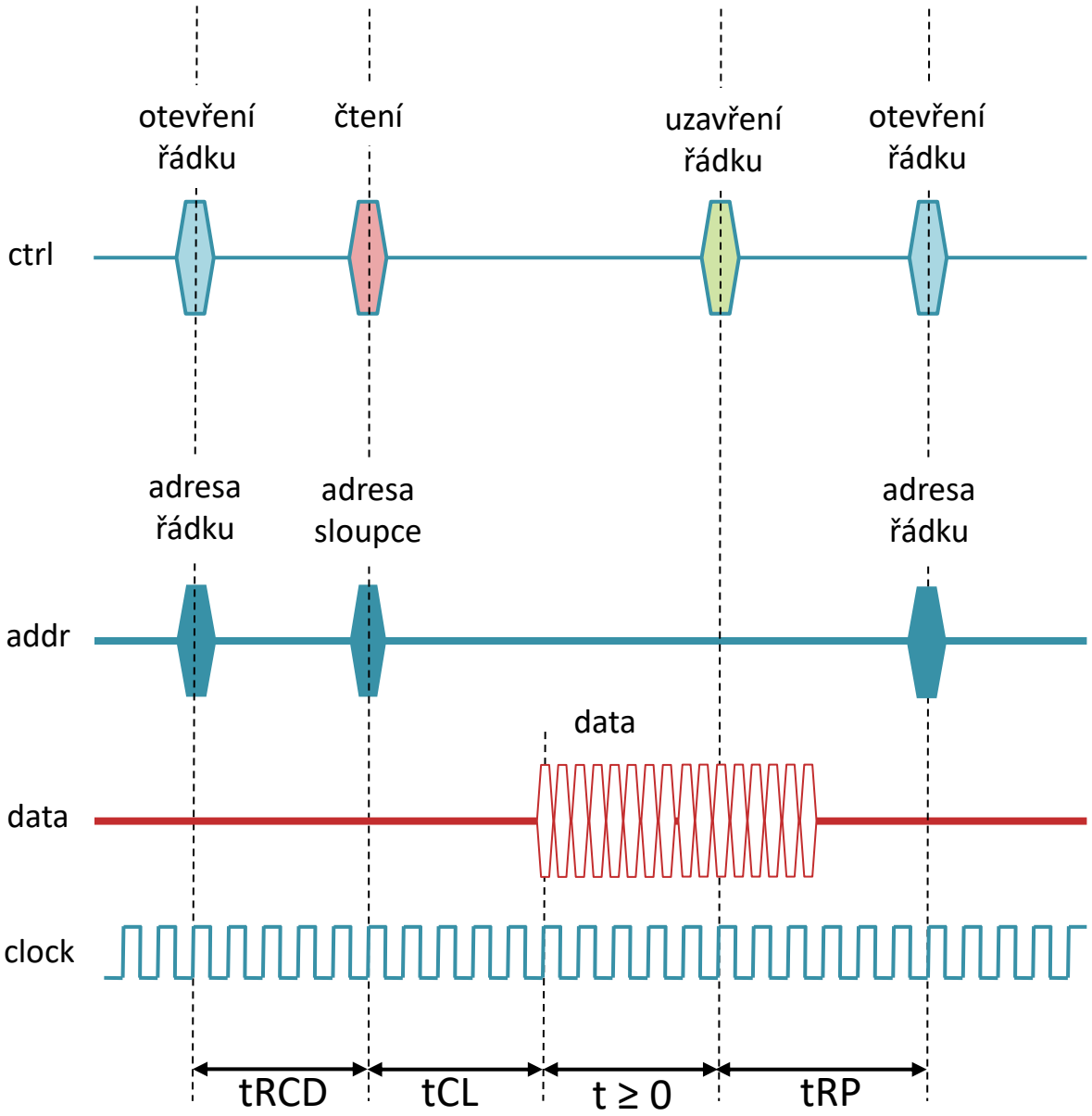
- Typický čip (2012)
  - matice 128 Mbit
  - 8 bank
  - celkem 1 Gbit
  - 256 M * 4 bit
- Paralelismus
  - Banky mohou pracovat nezávisle
  - V okamžiku předávání příkazu, adresy nebo dat je připojena jen jedna
  - Synchronizováno hodinovým signálem – programované zpoždění mezi příkazy a daty
  - Časový multiplex dat 8:1

Architektura paměti SDRAM – Časování



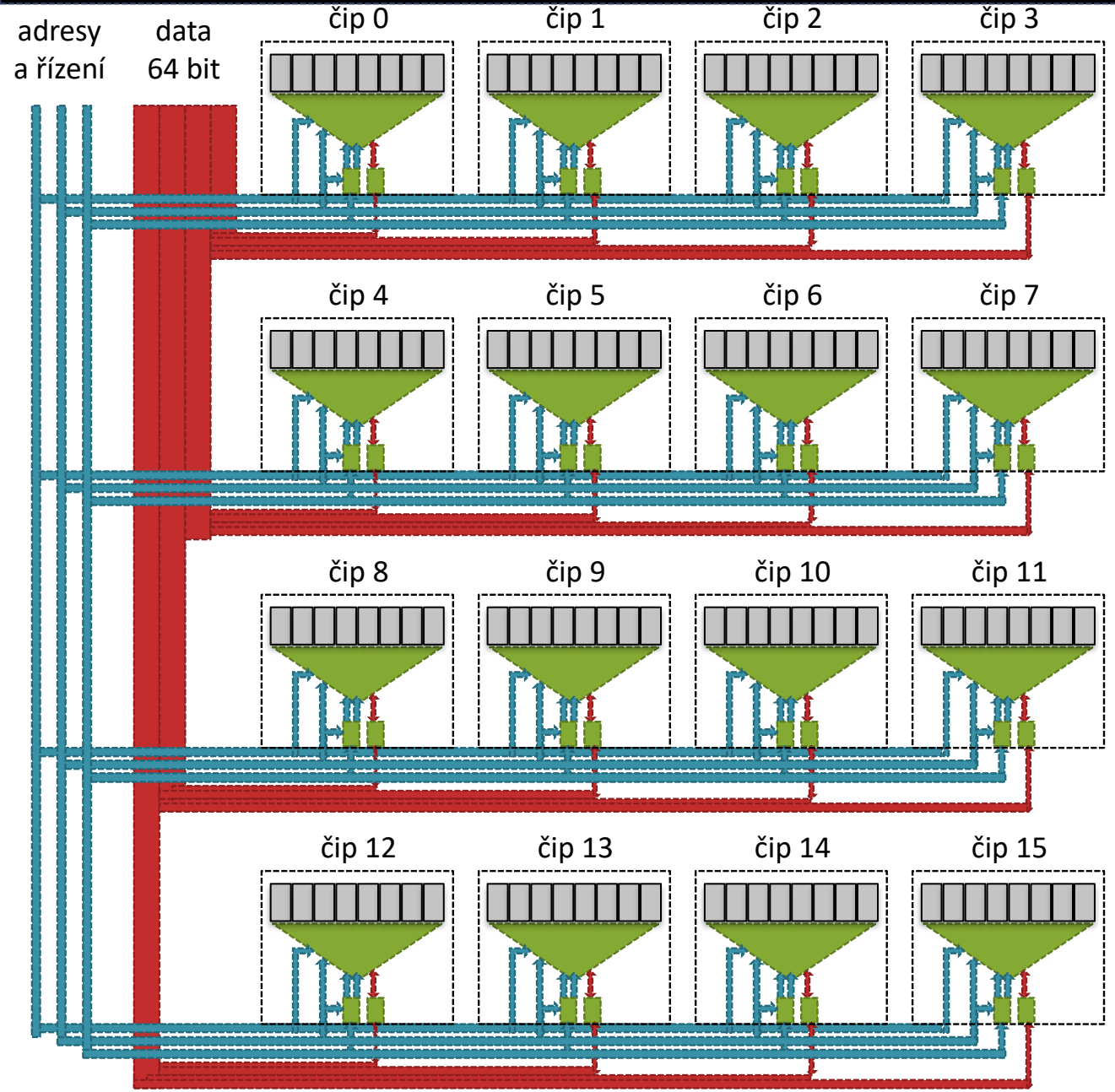
- Synchronous DRAM
  - Všechny signály synchronizovány hodinami
  - Časování udáváno v cyklech hodin
- Double-data-rate
  - 8 balíků dat za 4 cykly
  - Označení pamětí udává frekvenci dat
    - Dvojnásobek hodin
- Typické časy (2016)
  - DDR4-2400
    - hodiny = 1.2 GHz
  - tRCD = 15 cyklů
  - tCL = 15 cyklů
    - = 12.5 ns
  - tRP = 15 cyklů
  - Čas přenosu dat je vždy 4 cykly
    - = 3.33 ns

Architektura paměti SDRAM – Časování



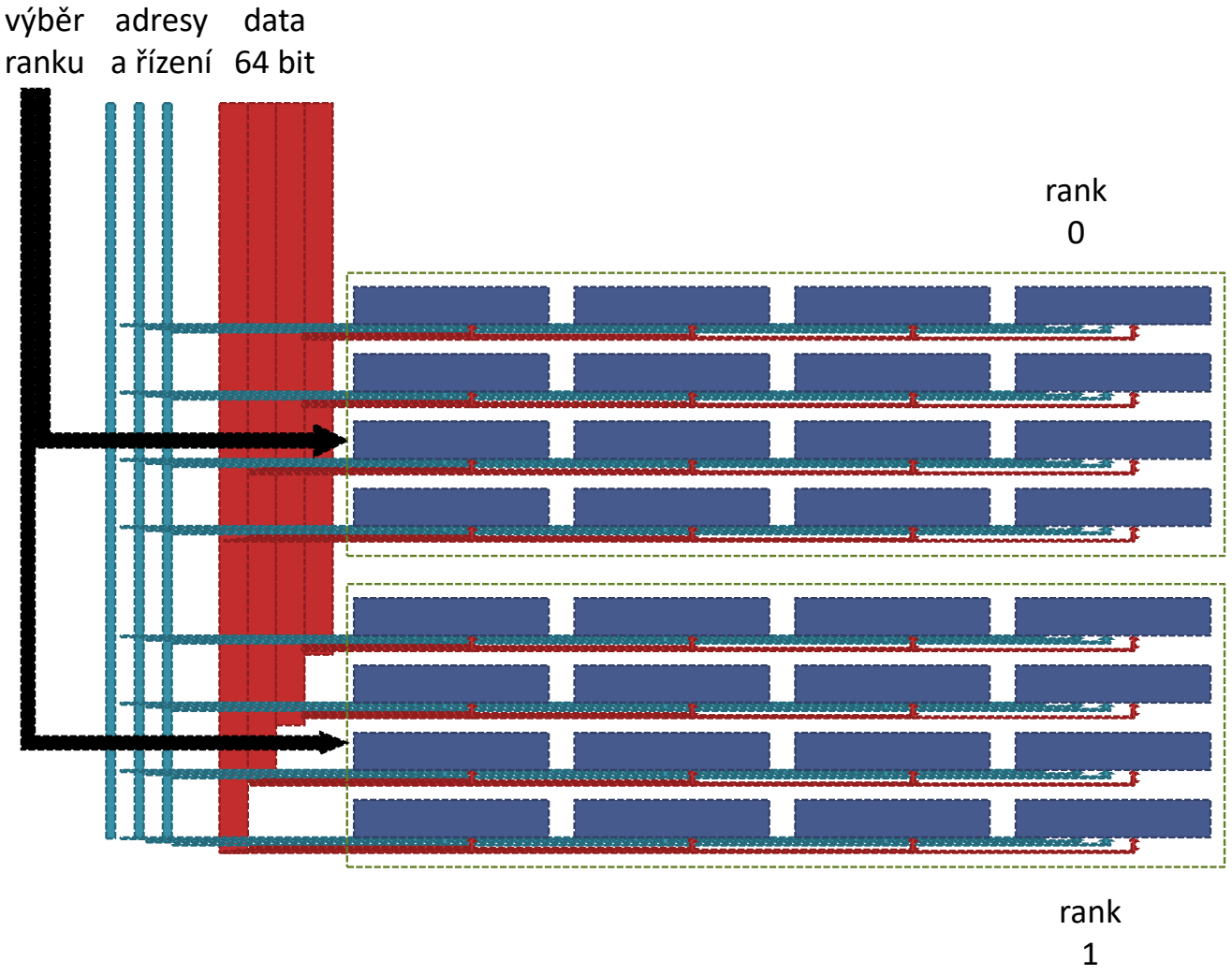
- Synchronous DRAM
  - Všechny signály synchronizovány hodinami
  - Časování udáváno v cyklech hodin
- DDR5
  - 16 balíků dat za 8 cyklů
  - Označení pamětí udává frekvenci dat
    - Dvojnásobek hodin
- Typické časy (2023)
  - DDR5-6000
    - hodiny = 3 GHz
  - tRCD = 30 cyklů
  - tCL = 30 cyklů
    - = 10 ns
  - tRP = 30 cyklů
  - Čas přenosu dat je vždy 8 cyklů
    - = 2,66 ns

Architektura paměti DRAM - příklad



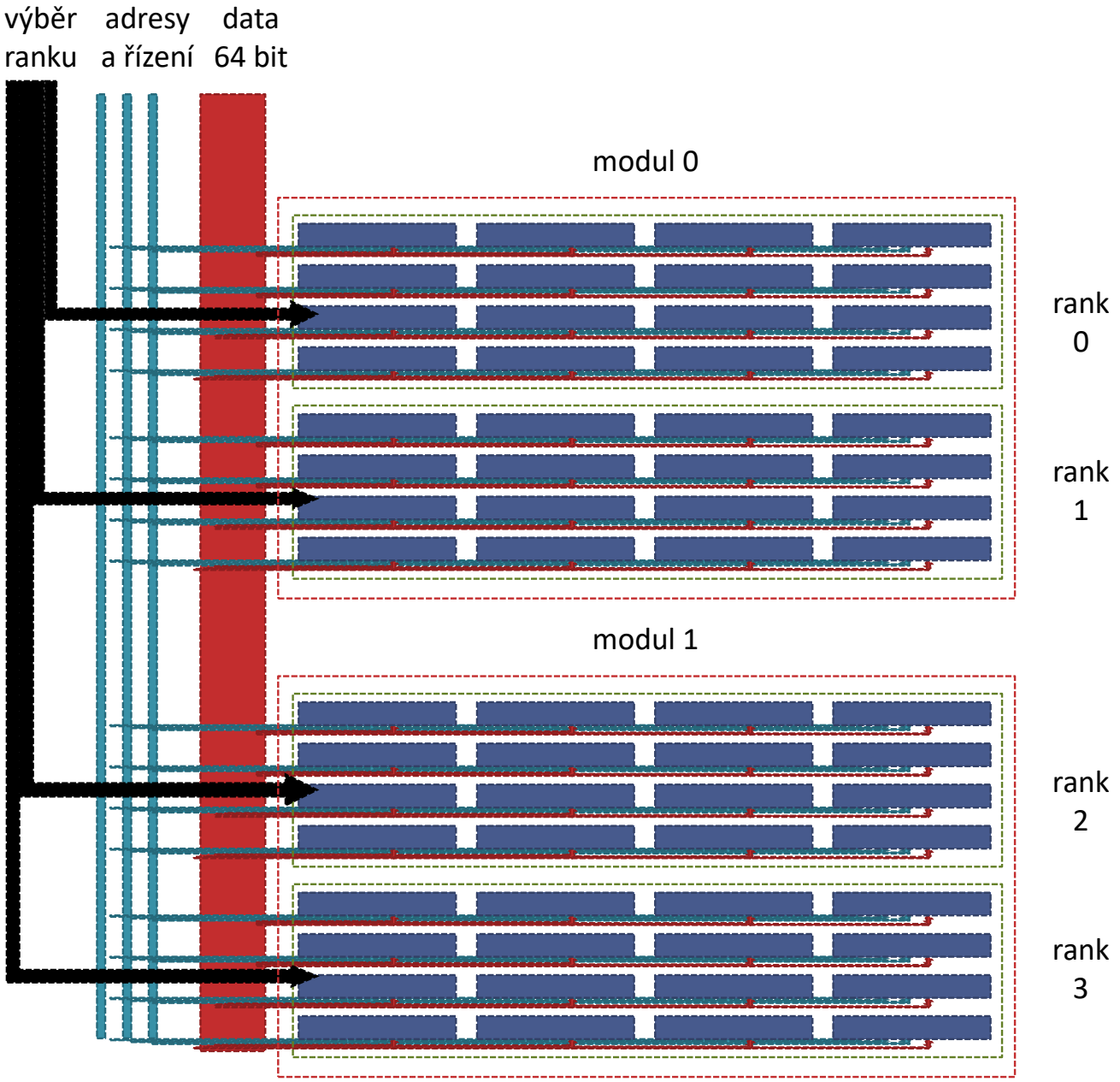
- Rank
  - čip 256 M * 4 bit
  - 16 čipů
  - celkem 2 GB
  - 256 M * 64 bit
- Paralelismus
  - Všechny čipy dělají totéž
  - Každý řeší 4 bity
- Doplňky
  - ECC: +2 čipy – paritní kontrola

Architektura paměti DRAM – typický 4GB DDR3 modul



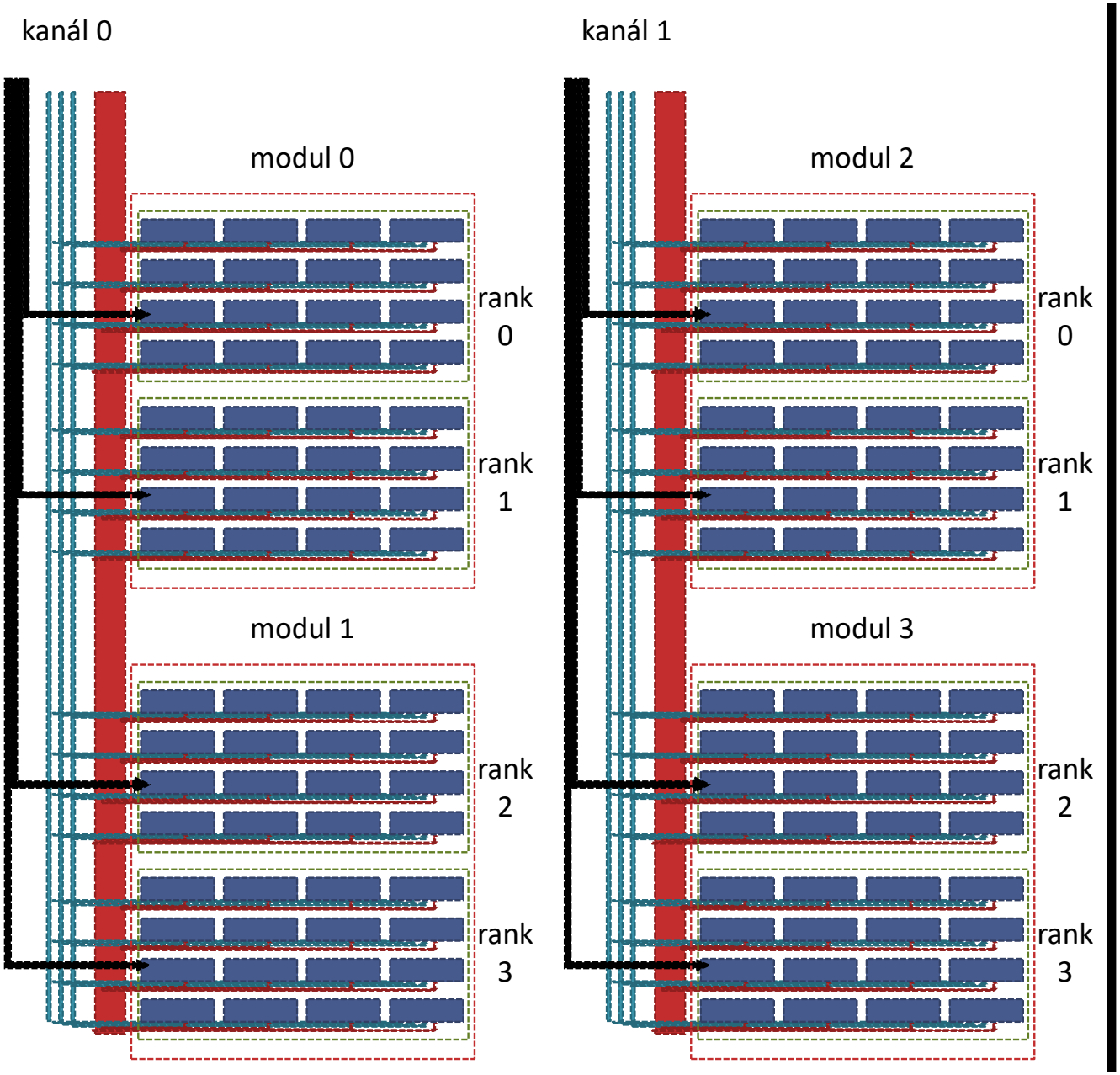
- DDR3 modul
  - dual rank
  - každý rank 256 M * 64 bit
  - celkem 4 GB
  - 32 čipů
  - 512 M * 64 bit
- Paralelismus
  - Ranky pracují nezávisle
  - Ke sběrnici může přistupovat jen jeden
- Doplňky
  - „registered“: opakovač signálů

Architektura paměti DRAM - kanály



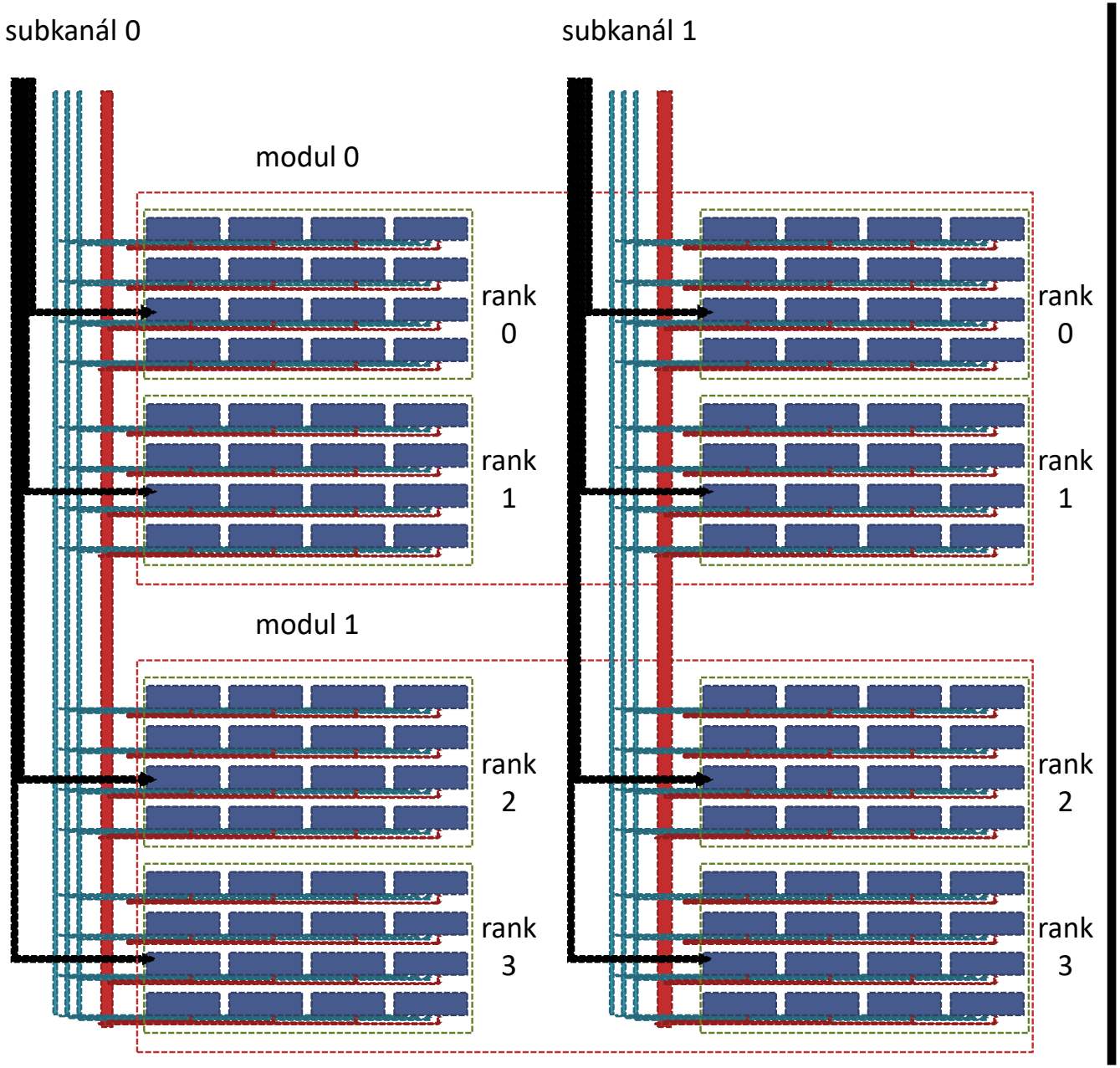
- DDR3 kanál
  - dva paralelně propojené moduly
  - každý modul 512 M * 64 bit
  - celkem 8 GB
  - 1 G * 64 bit
- Paralelismus
  - Jednotlivé ranky obou modulů pracují nezávisle
  - Data přenáší pouze jeden
- Varianty
  - „registered“: až 4 moduly
  - větší a pomalejší

Architektura paměti DRAM – dual channel



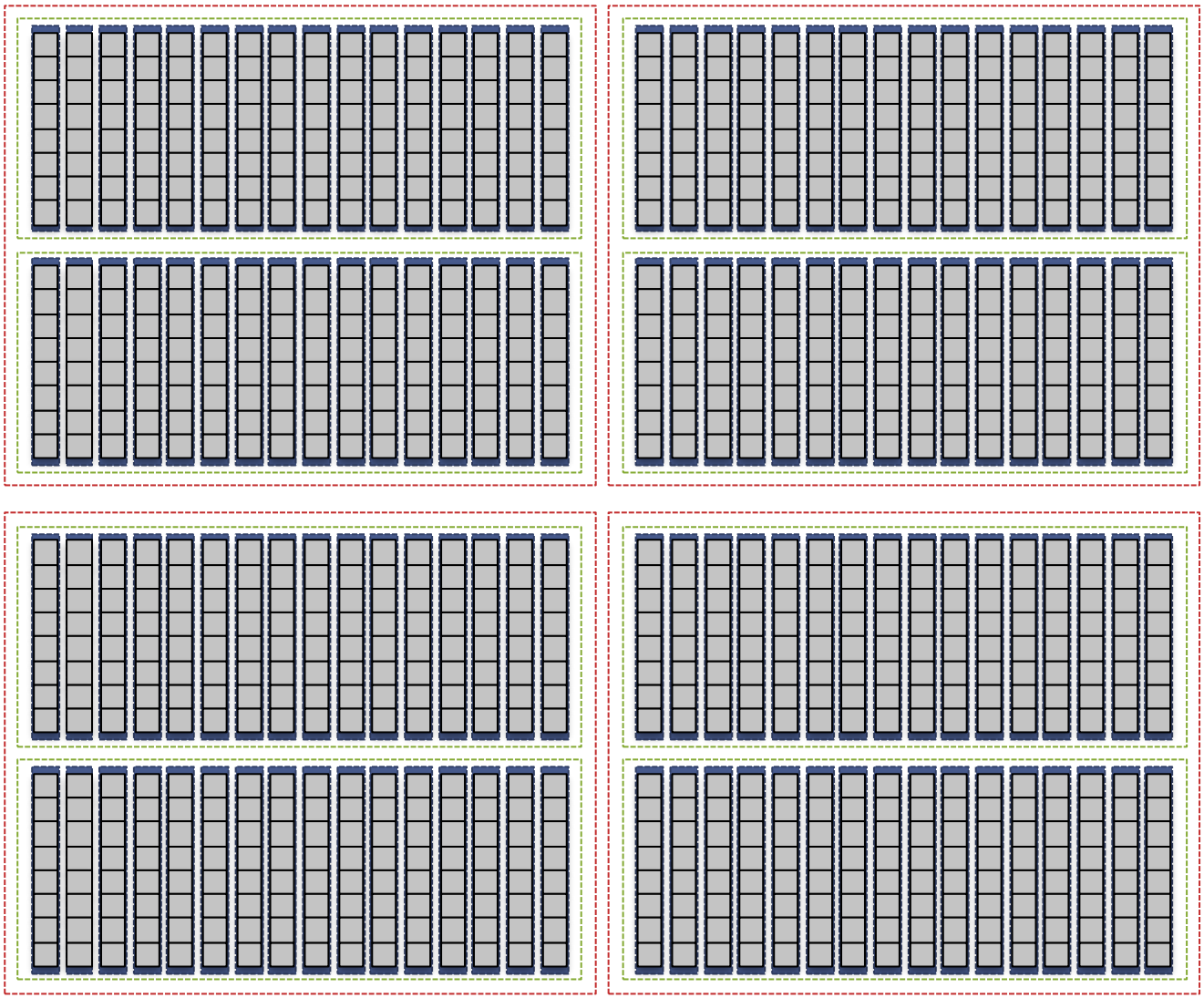
- Dual channel
  - 2 kanály
  - každý kanál 1 G * 64 bit
  - celkem 16 GB
  - 1 G * 128 bit nebo 2 G * 64 bit
- Paralelismus
  - Kanály dělají totéž nebo
  - Kanály pracují a přenášejí data nezávisle

Architektura paměti DRAM – DDR5



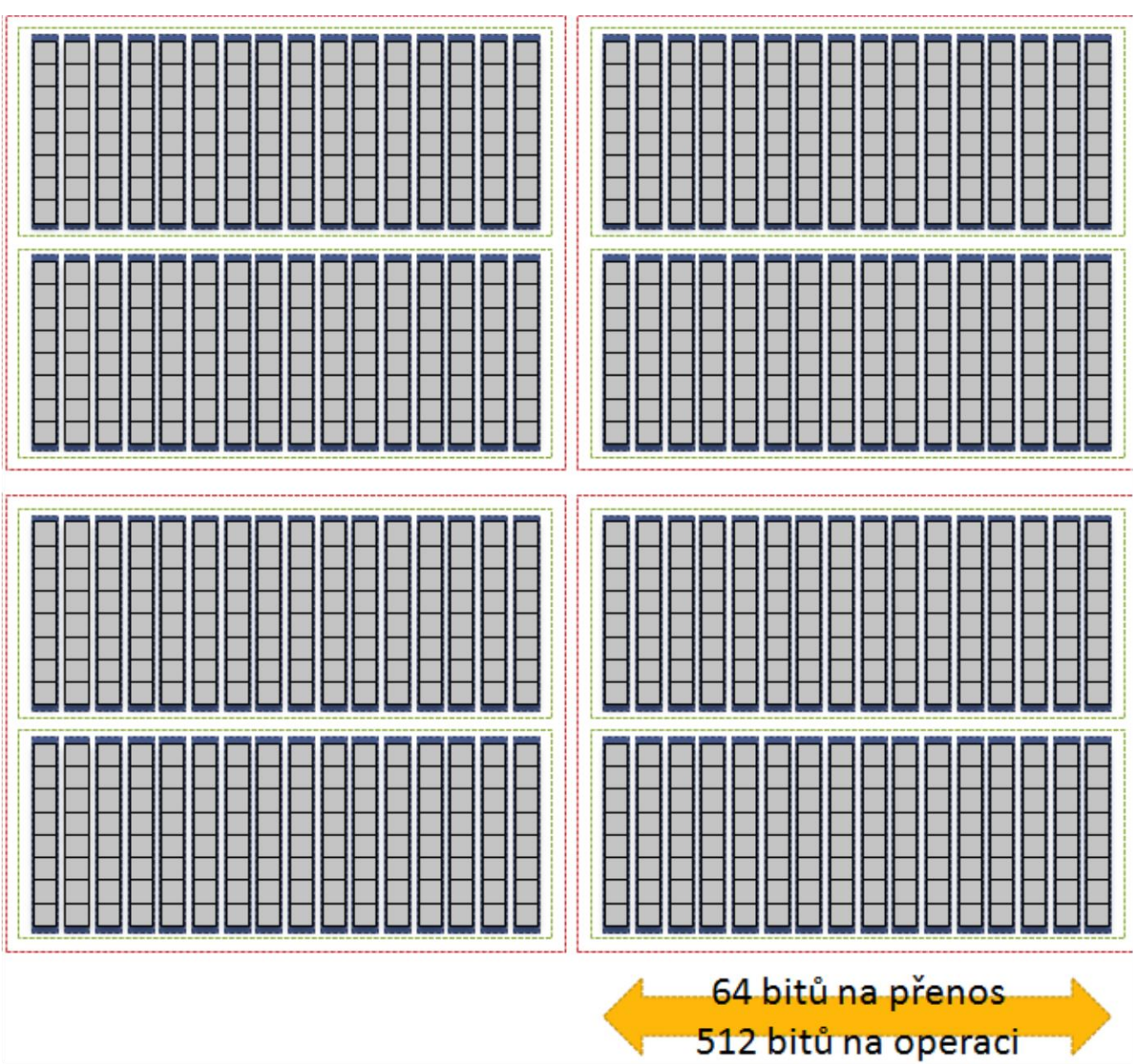
- DDR5
  - 2 subkanály po 32 bitech
  - 16 balíků dat v každém přenosu
  - 16*32=512 bitů
- Paralelismus
  - Subkanály pracují a přenášejí data nezávisle
  - Jsou fyzicky umístěny na témže modulu
  - Procesor může obsluhovat více kanálů (dvojic subkanálů)

Architektura paměti DRAM – Příklad: 4 * DDR3 4 GB DR x4



- Shrnutí
  - 2 kanály
  - 2*2 = 4 ranky
    nezávislá činnost
  - 16 čipů
    shodná činnost
  - 8 bank
    nezávislá činnost
  - 16384 řádků
    jeden aktivní
  - 8192 sloupců
    32 vybraných

Architektura paměti DRAM



- Atomická operace
  - časový multiplex: 8 přenosů po 64 bitech
  - celkem 512 bitů
  - při 1600 MT/s zaměstná 1 kanál na 5 ns
  - latence 26 ns, s úklidem 52 ns
- Paralelismus
  - Zbývajících 31 bank kanálu může dělat něco jiného
  - Propustnost kanálu stačí na 200 M operací/s
  - Kanály jsou dva (až 4)

Rozhraní procesor - DRAM

- Přímé připojení na DDR5
  - každý kanál až 6400 MT/s = 51.2 GB/s
- Přímé připojení na DDR4
  - každý kanál až 3200 MT/s = 25.6 GB/s
- Přímé připojení na DDR3
  - 1-4 kanály po 64 bitech, každý až 1600 MT/s = 12.8 GB/s
- Intel DMI, QPI, AMD HyperTransport
  - Serio-paralelní rozhraní, různé šířky, až 51.2 GB/s
- Intel FSB
  - 1 kanál, 64 bitů, až 1600 MT/s = 12.8 GB/s

- Komplikovanost rozhraní procesor – DRAM prodlužuje latenci
  - celkem 50-100 ns, samotná DRAM kolem 30 ns