

Distribuovaná sdílená paměť



Přednášky z Distribuovaných systémů
Ing. Jiří Ledvina, CSc.

Úvod



- Sdílená paměť multiprocesorového systému
 - Jednoduchá implementace paralelního zpracování
 - Snaha o přenesení do multipočítačového prostředí
 - Distribuovaná sdílená paměť
- Systémy distribuované sdílené paměti
 - Založené na stránkách
 - Založené na sdílených proměnných
- Distribuovaná sdílená paměť
 - Soubor počítačů sdílí jeden virtuální adresní prostor

Úvod



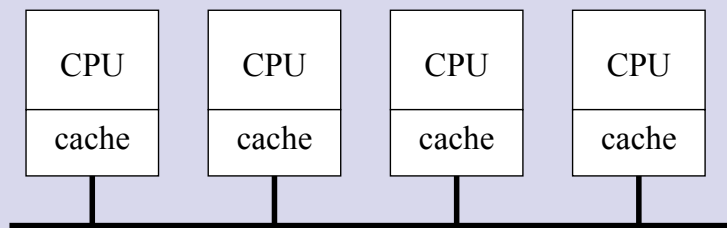
- **Implementace**
 - Odkazy na lokální stránky jsou realizovány hardwarem
 - Odkazy na vzdálené stránky způsobí výpadek stránky a stránka se natáhne ze vzdáleného systému
- **Optimalizace**
 - Sdílení pouze vybraných částí paměti
 - Replikace sdílených proměnných na více počítačů
- **Sdílená paměť v multiprocесorech**
 - Vícebranná paměť
 - Sběrníkové multiprocесory
 - Multiprocесory s kruhovou organizací
 - Přepínané multiprocесory
 - NUMA – Non-Uniform memory Access

Porovnání systémů sdílené paměti



- **Řízené MMU**
 - Sběrníkové multiprocесory (Sequent)
 - Přepínané multiprocесory
- **Řízené OS**
 - NUMA architektura
 - DSM založené na výměně stránek (Ivy)
- **Řízené aplikacemi (úroveň programovacích jazyků)**
 - DSM se sdílenými proměnnými (Munin)
 - Objektově orientované DSM (Orca)

Sběrníkové multiprocesory



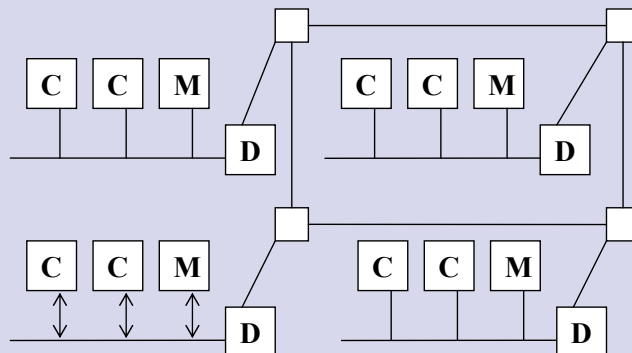
- Protokol pro udržení konzistentnosti cache
 - Write-through cache protocol
- Write through
 - Zneplatnění dat
 - „špinavá data“

Multiprocesory založené na kruhové topologii



- Bez centralizované globální paměti
- Paměťové bloky ve sdílené paměti mají home memory field
- Čtení:
 - čekání na token
 - odeslání požadavku
 - počítač který má požadovaný blok jej pošle v tokenu, mazání bitu exclusive
- Zápis:
 - lokální blok – pouze kopie, lokální zápis
 - blok je lokální – ne pouze kopie, poslání paketu zneplatnění, nastavení pole exclusive
 - blok není lokální: odeslání požadavku nebo zneplatnění

Přepínané multiprocesory



Non-uniform Memory Access (NUMA)

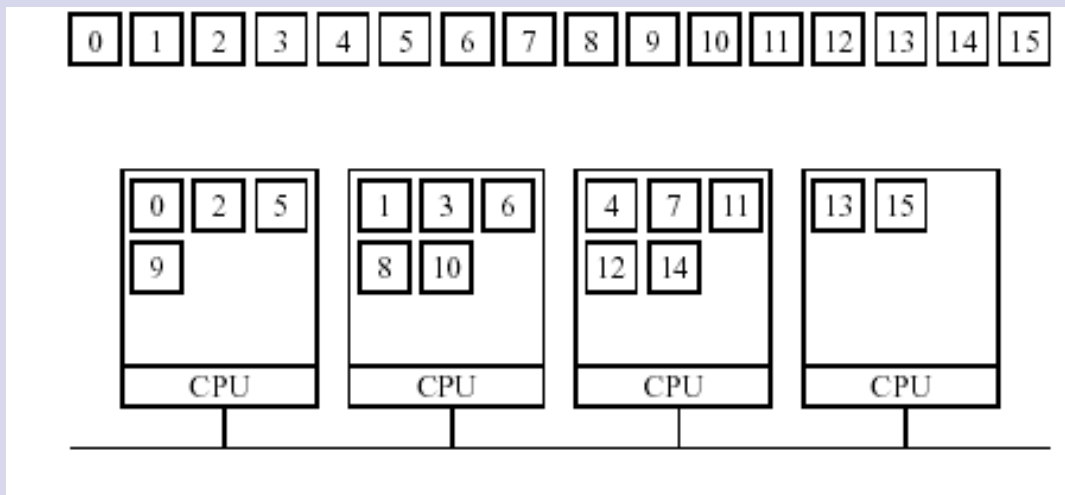


- **NUMA**
 - Procesor může přímo pracovat s lokálními i vzdálenými paměťovými místy
 - Bez podpory programového vybavení
- **Pracovní stanice na síti**
 - Mohou pracovat pouze s lokální pamětí
- **Cíl distribuované sdílené paměti**
 - Přidat software aby umožnil pracovat s multiprocesorovým kódem
 - Zjednodušit programování



Základní návrh

- Emulace cache multiprocesoru s použitím MMU a systémového software



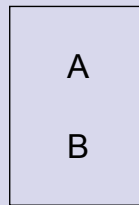
Podmínky návrhu

- Replikace
 - Replikovat read/only části
 - Replikovat Read/write části
- Granularita (zrnitost)
 - Omezení: části paměti jsou násobkem stránek
 - Klady velkých částí
 - Snižují režii protokolu
 - Locality of reference
 - Zápory velkých částí
 - Falešné sdílení

Falešné sdílení

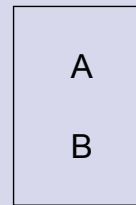


Procesor 1



Zapisuje A

Procesor 2



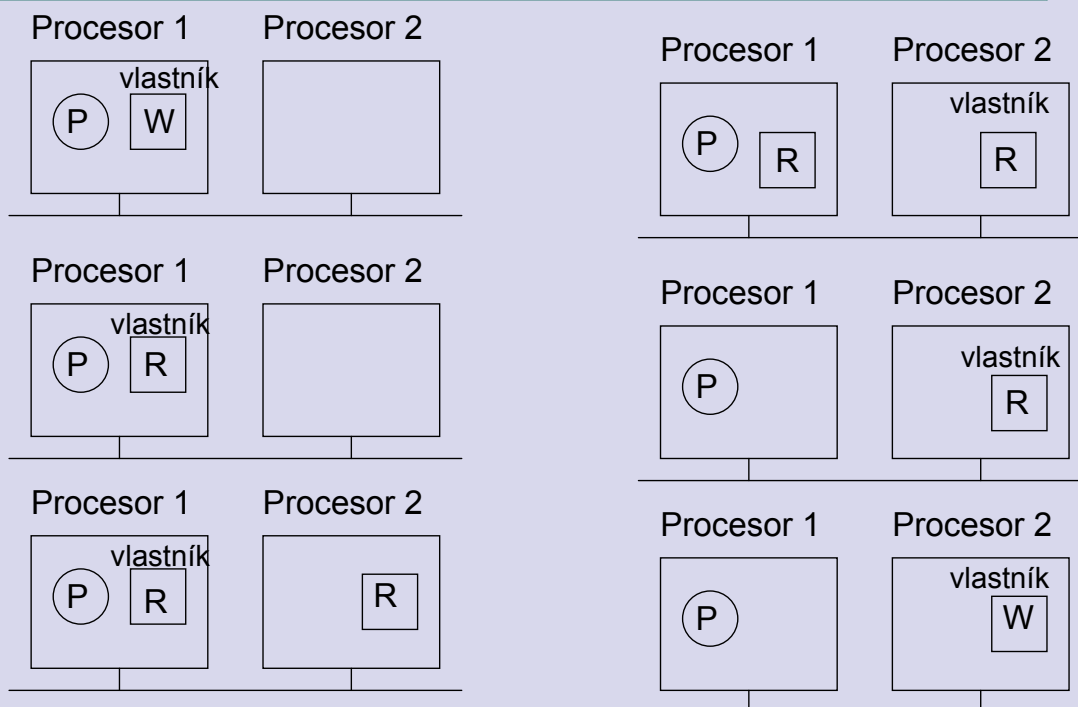
Zapisuje B

Sekvenční konzistentnost



- **Pouze jedna kopie každé stránky**
 - Konzistentnost je zaručena (triviální)
- **Replikované stránky**
 - Read/only – v pořádku
 - Read/write
 - Operace čtení – instalace lokální kopie, nastavena na R/O
 - Operace zápisu – oprava nebo zneplatnění ostatních kopií
- **Typický protokol**
 - R(readable), W(writable and readable) stránky
 - Každá stránka má vlastníka: proces, který zapisoval do stránky naposledy

Protokol pro sekvenční konzistentnost



Hledání vlastníka

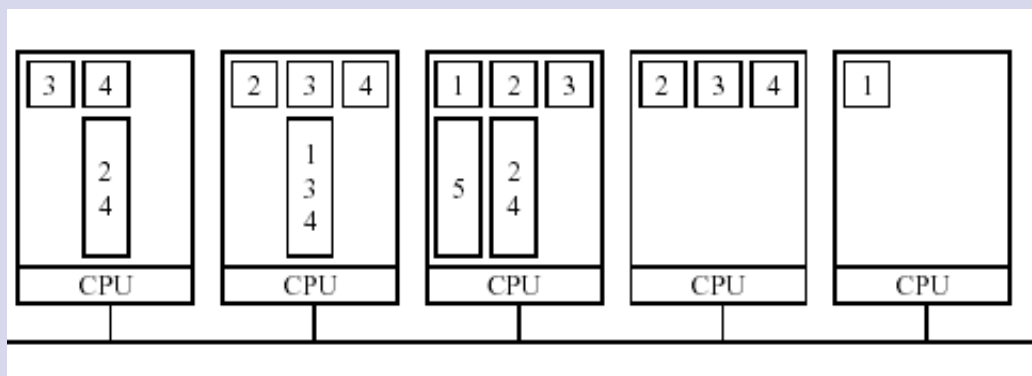


- **Poslání požadavku na vlastníka pomocí broadcastu**
 - Kombinace požadavku s požadovanou operací
 - Problém: broadcast osloví všechny účastníky (přeruší všechny procesy), využívá šířku pásma sítě
- **Manager stránek**
 - Možné úzké místo
 - Více managerů stránek, hashování adres stránek
- **Pravděpodobný vlastníka**
 - Každý proces si pamatuje pravděpodobného vlastníka
 - Periodicky obnovují informaci o stávajících vlastnících



Hledání kopií

- Jak najít kopie pokud musí být zneplatněny
- Požadavek ve formě broadcastu
 - Co když není broadcast spolehlivý
- Copyset (soubor kopií)
 - Je udržován managerem stránek nebo vlastníkem



Prostředky synchronizace

- Synchronizace
 - Zámky
 - Semaforey
 - Bariery
 - Tradiční synchronizační mechanismy pro multiprocesory nefungují
 - Managery synchronizace

Sdílené proměnné v distribuované sdílené paměti



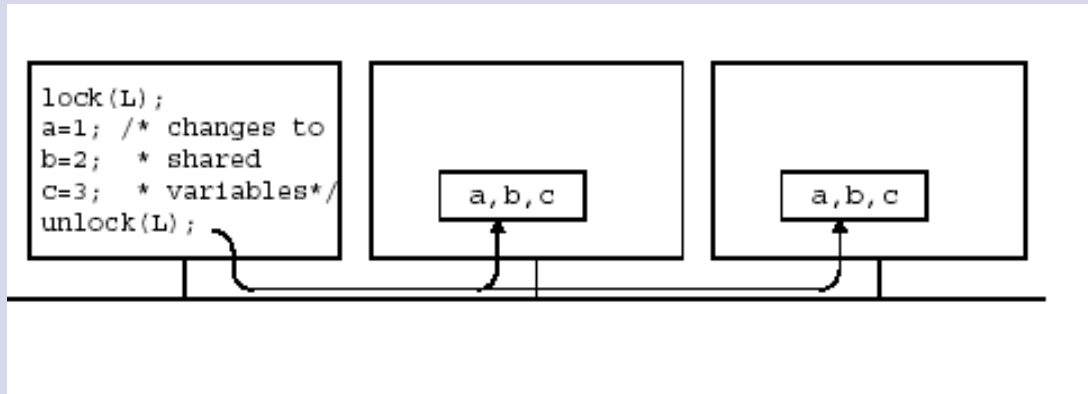
- **Není nutné sdílet celý adresní prostor**
- **Sdílení jednotlivých proměnných**
- **Větší vůle v algoritmech pro opravování replikovaných proměnných**
- **Příležitost eliminovat falešné sdílení**
- **Příklad: Munin**

Munin



- **Používá MMU: umístění každého sdíleného objektu ve zvláštní stránce**
- **Explicitní deklarace sdílených proměnných**
 - Klíčové slovo `shared`
 - Překladač ukládá proměnné do zvláštních stránek
- **Synchronizace:**
 - Uzamykání proměnných
 - Bariéry
 - Podmíněné proměnné
- **Uvolňovací konzistentnost**

Munin



Munin



- **Kritické sekce**
 - Zápis do sdílené proměnné se objeví uvnitř kritické sekce
 - Čtení se může objevit kdekoliv
 - Pokud je opuštěna kritická sekce, modifikované proměnné jsou opraveny ve všech počítačích
- **Rozeznává tři třídy proměnných**
 - Obyčejné proměnné – nejsou sdíleny, mohou být modifikovány pouze procesem, který je vytvořil
 - Sdílené proměnné – jsou viditelná ve více procesech, zůstávají sekvenčně konzistentní
 - Synchronizační proměnné
 - Jsou dostupné pouze systémovými procedurami
 - Pro zámkové jsou to lock/unlock, pro bariéry increment/wait



Munin – sdílené proměnné

- **Read-only**
 - Nejsou měněny po inicializaci, nejsou s nimi žádné problémy
 - Chráněny MMU
- **Migratory (potulné)**
 - Nejsou replikovány: migrují od počítače k počítači podle vstupů do kritických sekcí
 - Spojeny se zámkem
- **Write-shared (sdílené pro zápis)**
 - Chráněny před vícenásobným zápisem programů
 - Používají protocol pro řešení vícenásobného zápisu do jedné proměnné
- **Conventional (obecné)**
 - Chovají se jako v konvenčním page-based DSM: pouze jedna kopie přepisovatelné stránky, přenášeny mezi procesory



Munin – dvojice stránek

- Na počátku je stránka sdílená pro zápis označena jako read-only
- Objeví-li se zápis, je vytvořena kopie stránky a original je určen pro čtení a zápis
- uvolnění:
 - porovnání upravených stránek se svými dvojčaty slovo po slovu
 - poslání odlišností do všech procesů, které to potřebují
 - nastavení stránky na read-only
 - porovnání příchozích stránek pro modifikovaná slova
 - je-li modifikováno lokální i příchozí slovo, signalizuje se chyba za běhu