

UFSC-CTC-INE-PPGCC

INE 410131 – Gerência de Dados para Big Data

Bancos de Dados em Memória: Visão Geral

BDs Tradicionais X BDs em Memória

- BDs convencionais (BDCs)
 - principal dispositivo de armazenamento de dados é o disco
- BD em memória (*In-memory DB - IMDB*)
 - totalidade (ou quase) do BD está na memória RAM
 - motivação
 - minimizar o gargalo de acesso dos BDCs
 - HW mais desenvolvido
 - aplicações OLTP e OLAP

*“Memory is the new disk.
Disk is the new tape.”*
Jean Gray, DB scientist

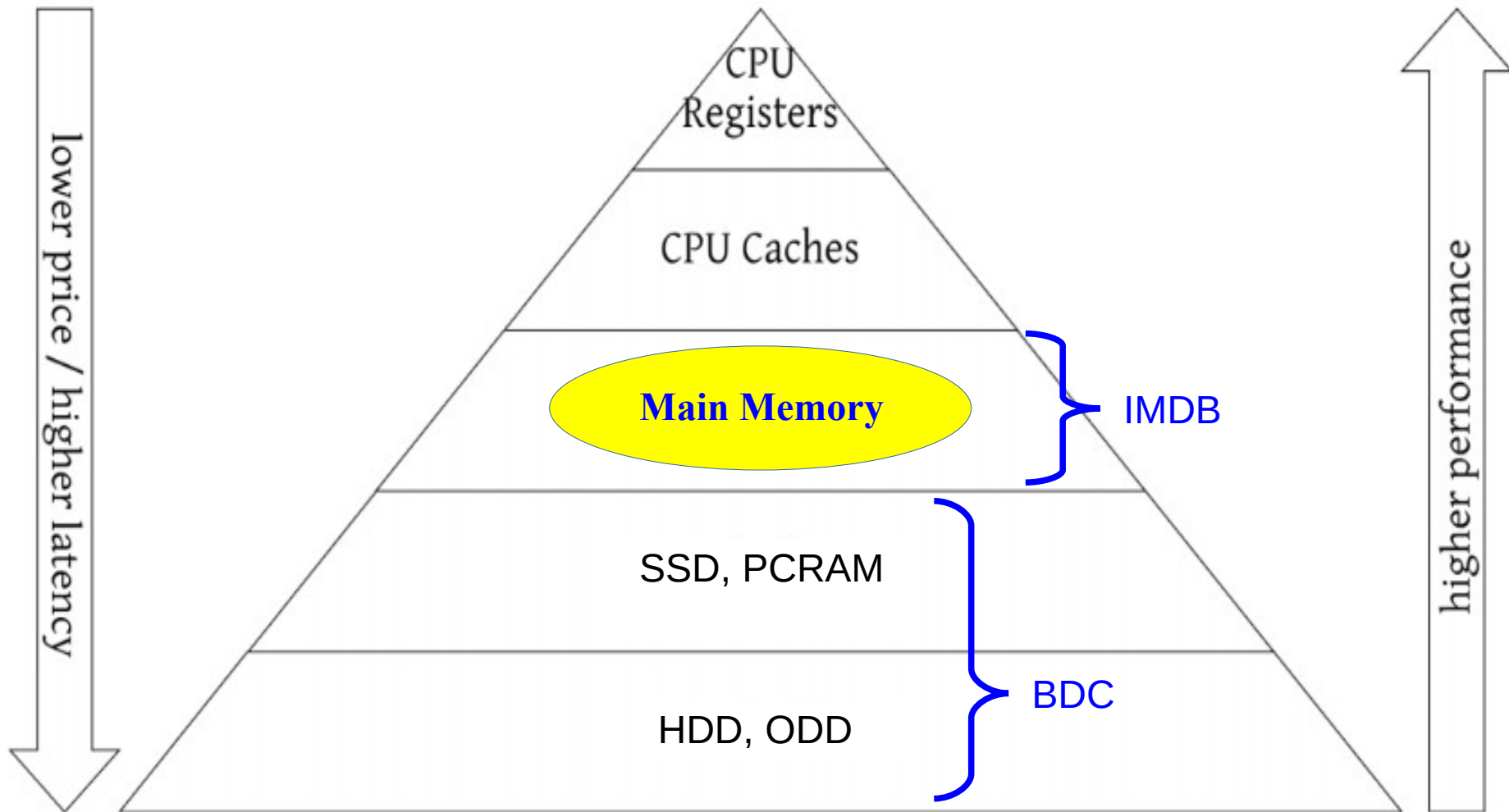
Alguns Conceitos

- Non-Volatile Memory (NVM):
 - HDD (Hard-Disk Drive): armazenamento em meio magnético
 - ODD (Optical-Disk Drive): armazenamento em dispositivo eletrônico semicondutor c/ o uso de laser
 - taxa de transferência de dados menor que HDD, mas tempo de busca maior que HDD
 - SSD (Solid-State Drive): dispositivo de baixa latência e custo mais alto, com armazenamento em *chips* de memória
 - PCRAM (Phase Change RAM): menor espaço físico e menor consumo energético que SSD, custo elevado e latência ligeiramente menor que SSD

Alguns Conceitos

- DRAM (Dynamic RAM):
 - memórias voláteis de baixo custo e alta capacidade de armazenamento (Tb)
 - sua capacidade vem aumentando aproximadamente 10x a cada 5 anos

BDC X IMDB



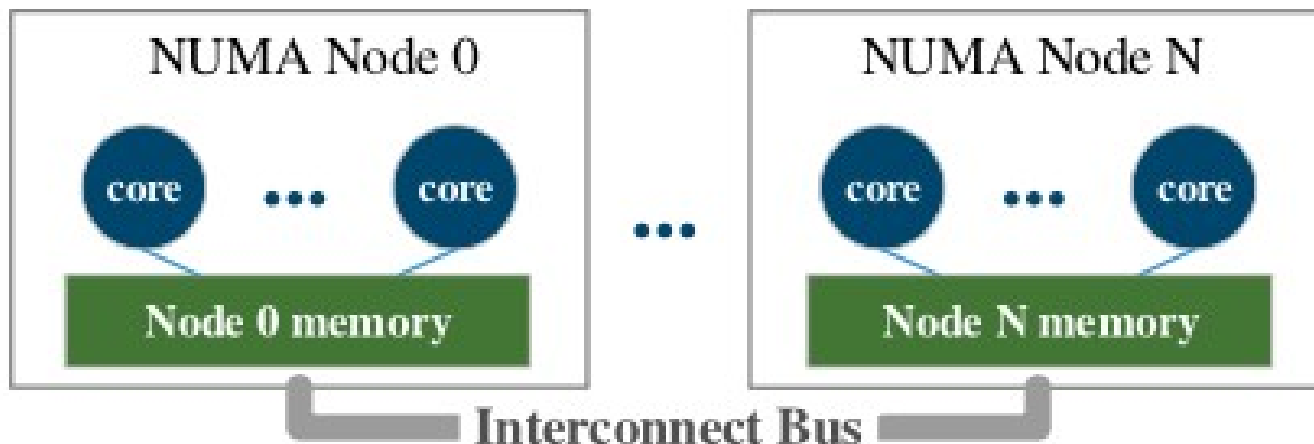
IMDB – Principais Tecnologias HW

1. Arquitetura NUMA

2. NVDIMM

1 – Arquitetura NUMA

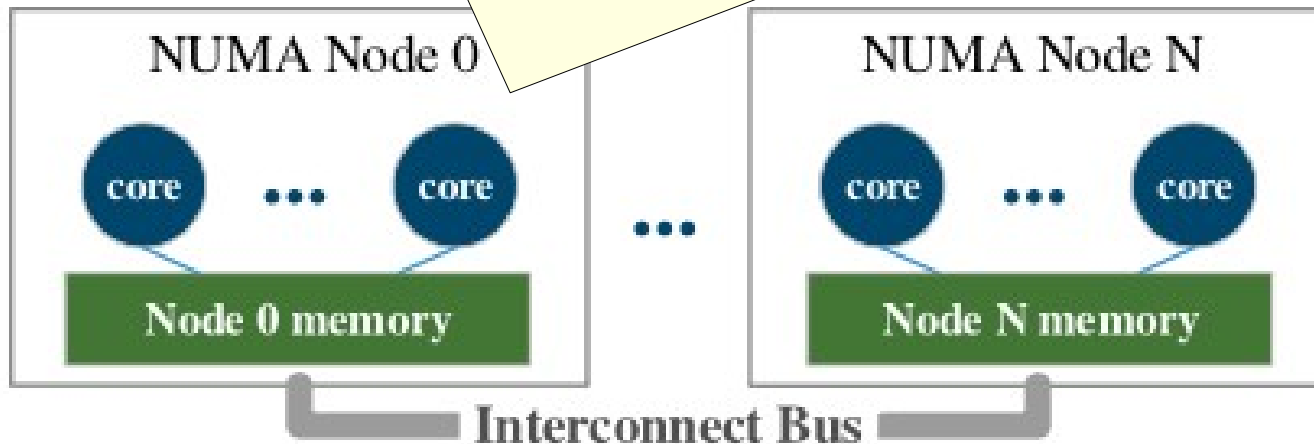
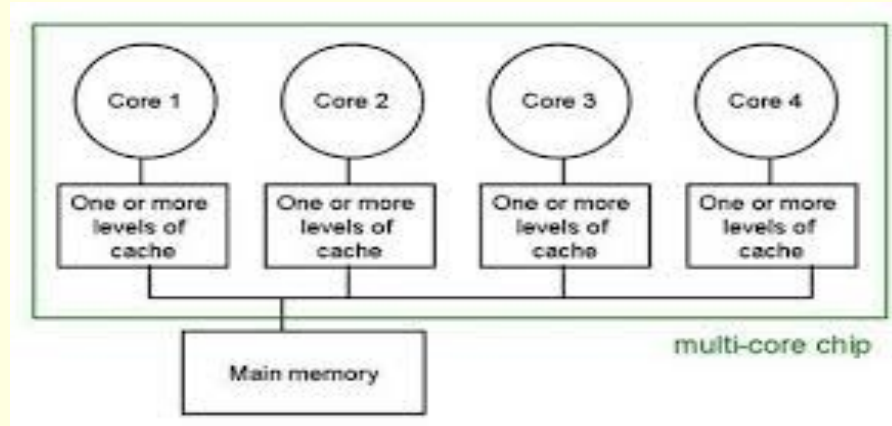
- Arquitetura composta por um conjunto de processadores *multicore* interconectados
- *NUMA (Non-Uniform Memory Access)*
 - maior capacidade de processamento de dados
 - uso paralelo de memórias DRAM



1 – Arquitetura NUMA

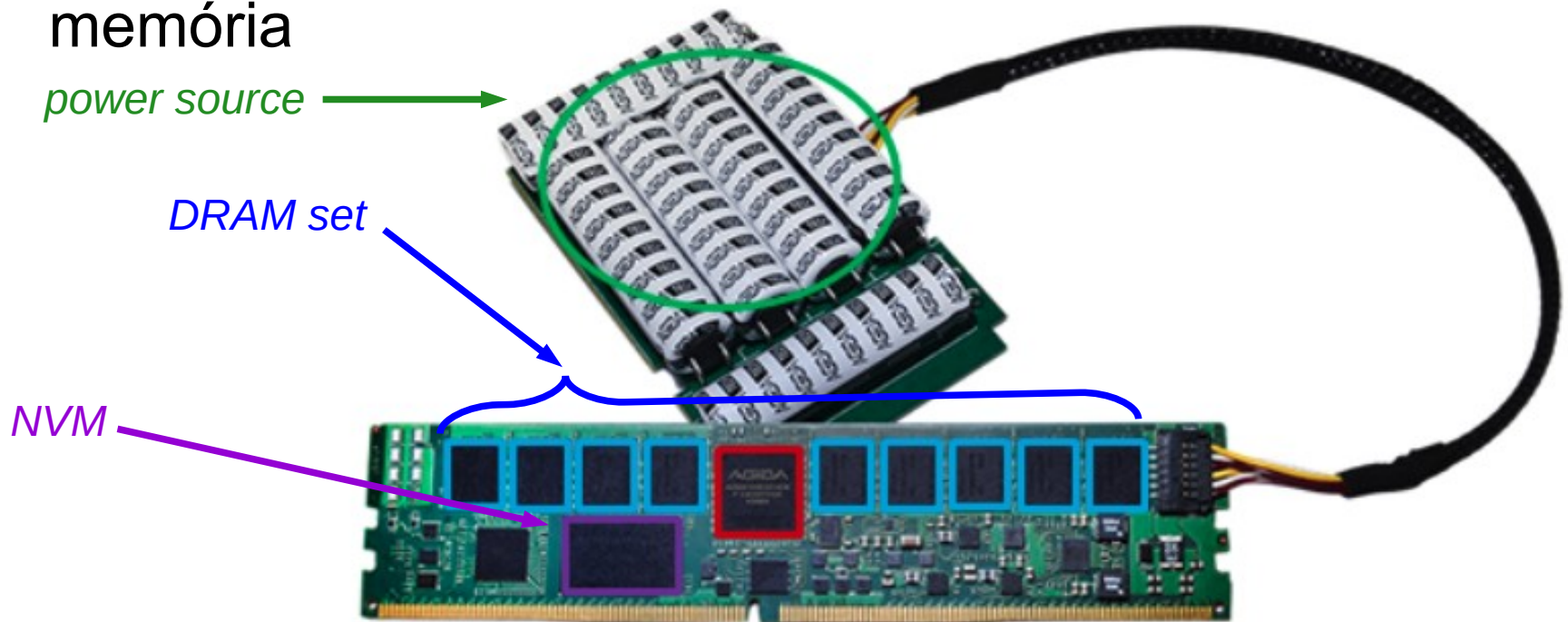
- Arquitetura processadora
- **NUMA** (Non-Uniform Memory Access)
 - maior capacidade
 - uso paralelo

cada processador *multicore* mantém 2 ou mais unidades CPU ou GPU que compartilham uma DRAM



2 – NVDIMM

- *Non-Volatile Dual In-Line Memory Module*
- Dados são mantidos na ausência de energia pois possui fonte de energia dedicada e uma 2ª memória não-volátil: DRAM+NVM
 - permite a descarga de dados em tempo para a 2ª memória



Operação de Eviction

- Decisão quanto aos dados que devem ser transferidos da DRAM para NVM (*swap*)
- Como funciona
 - dados não frequentemente acessados (*“cold tuples”*) são transferidos eventualmente para NVM
 - um índice para tuplas frias mantém as suas localizações na NVM (*“tombstone index”*)
 - quando uma transação T_x deseja acessar *cold tuples*:
 - 1) T_x é abortada
 - 2) Uma *thread* é ativada para trazer esses dados para a DRAM (*“tuple resurrection”*)
 - 3) T_x é restartada

IMDB – Tópicos de Pesquisa

- Paralelismo

- processar múltiplos dados presentes em diferentes *cores* em uma única instrução
- lidar com múltiplos dados em um único registrador (*otimizar processamento em nível de CPU*)

IMDB – Tópicos de Pesquisa

- Controle de Concorrência: *Very Lightweight Locking (VLL)*, *MVCC-adapted* e *HTM-based timestamp-based schedulers*
 - técnicas de *scheduling* que **mantêm informações**, como o tipo de bloqueio, fila de transações em espera ou o *timestamp*, **junto com o dado na memória** e não em estruturas de dados separadas
 - **separar** transações OLTP e OLAP
 - OLAP não faz atualização de dados e pode ser executada com maior paralelismo




IMDB – Tópicos de Pesquisa

- Recuperação de Dados (*Recovery*)
 - *logs* persistidos em NVM rápidos (SSD ou PCM) dedicados e replicados
 - *command logging*: guardam apenas a identificação do dado e a operação realizada sobre ele no *log* (ao invés dos valores antigo e novo do dado)
 - *group commit*: agrupam operações de *commit* para então persistir no *log*

IMDB – Tópicos de Pesquisa

- Indexação
 - estruturas de acesso rápido pois estão na DRAM
 - 1) estruturas de índice *baseadas em árvore* com foco em *range queries* para *data analytics*
 - *T-Tree, CSS-Trees, CSB+-Trees, delta-Tree, BD-Tree, ...*
 - para dados mantidos ordenados e contíguos em memória por um atributo X, basta indexar apenas o 1º registro de X em um *range*. Os demais são rapidamente acessados (*índice por range de dados*)
 - 2) estruturas de índice *baseados em hash*
 - utilizados por alguns SGBDs NoSQL chave-valor (*Redis, Memcached, RAMCloud*)

Soluções IMDB (além dos BD NewSQL)

- Apache Ignite 
 - *memory-centric distributed DB, caching, and processing platform for transactional, analytical, and streaming workloads, delivering in-memory speeds at petabyte scale. Multimodel: relational; key-value*
- OrigoDB 
 - *fully ACID in-memory DB. Multimodel: relational; document; key-value; graph*
- GemFire 
 - *real-time, consistent access to data-intensive applications throughout widely distributed cloud architectures. An in-memory DBMS that provides reliable asynchronous event notifications and guaranteed message delivery. Multimodel: document; key-value*

IMDB – Atividade

https://en.wikipedia.org/wiki/In-memory_database

1) O D do ACID (Durabilidade) é um requisito desafiador para um IMDB, pois os dados estão prioritariamente na memória RAM. Para garantir esta propriedade, algumas técnicas de apoio sugeridas são *checkpoint*, *logging* e *replication*. Como elas funcionam?

2) Qual a diferença entre um IMDB e uma *cache* de dados?