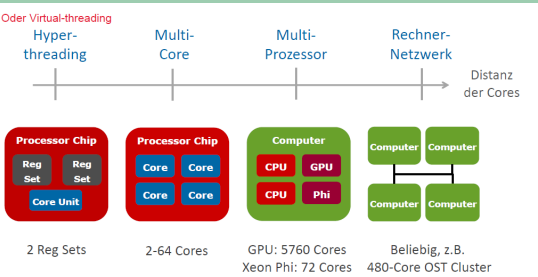


1 Multi-Threading Grundlagen



**Parallelität (Parallelism):** Mehrere Teilabläufe, welche gleichzeitig auf mehreren Prozessoren/Cores laufen. → Schnellere Programme

**Nebenläufigkeit (Concurrency):** Gleichzeitig oder verzahnt ausführbare Abläufe, welche auf gemeinsame Ressourcen zugreifen (Logisch unabhängig) → Einfachere Programme

**User-Level Threads:** Im Prozess implementiert (keine echte Parallelität)

**Kernel-Level Threads:** Im Kernel implementiert (Multi-Core Ausnutzung) → Kontextwechsel vom Prozess per SW-Interrupt

1.1 Thread Scheduling

**Processor Sharing:** Mehr Threads als Prozessoren und bei Wartebindung Proz. an anderen bereiten Thread abgeben.

**Verzahnte Ausführung:** Prozessor führt Instruktionen von mehreren Threads in Teilsequenzen aus. → Quasiparallelität

**Synchron:** Warten auf Bedingung → Waiting Threads

**Asynchron:** Zeitablauf → Nach gewisser Zeit Proz. abgeben

**Kooperativ:** Threads müssen explizit beim Scheduler in Abständen Kontextwechsel synchron initiieren

**Preemptiv:** Scheduler kann per Timer-Interrupt den laufenden Thread asynchron unterbrechen

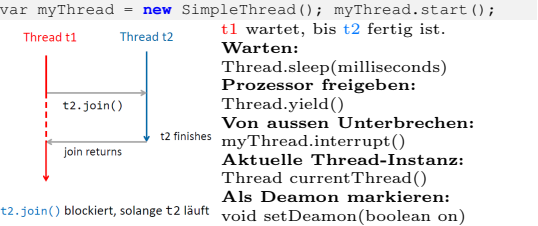
1.2 Multi Thread Programmierung

JVM ist ein Prozess im Betriebssystem. → Programmierere kann weitere Threads starten

Die JVM läuft, solange Threads laufen.

**Ausnahme:** Deamon Threads (z.B. Garbage Collector). Mit `System.exit()/Runtime.exit()` kann die JVM direkt terminiert werden (unsauber)

```
// A,B Ausgaben können durcheinander sein:
public class MultiThreadTest {
    public static void main (String[] args) {
        var a = new Thread(Thread(() -> multiPrint("A")));
        var b = new Thread(Thread(() -> multiPrint("B")));
        a.start();
        b.start();
        System.out.println("main finished");
    }
    static void multiPrint (String label) {
        for (int i = 0; i < 10; i++) {
            System.out.println(label + ": " + i);
        }
    }
}
// Explizite Runnable-Implementation:
class SimpleLogic implements Runnable {
    @Override
    public void run () {
        // thread behavior
    }
}
var myThread = new Thread(new SimpleLogic());
myThread.start();
// Sub-Klasse von Thread
class SimpleThread extends Thread {
    @Override
    public void run ()
        // thread behavior
    }
}
```



2 Thread Synchronisation

Threads teilen sich Adressraum und Heap.

2.1 Race Condition

Wenn mehrere Threads das Selbe Objekt lesen und anhand vom Resultat dieses überschreiben. (z.B. deposit: zuerst lesen dann += amount → Zwischen lesen und schreiben könnte ein anderer Thread die balance geändert haben)

**Lösung:** Gegenseitiger Ausschluss (Mutual Exclusion)

→ Geht nicht mit einfachem locked Boolean (Keine atomaren Instruktionen)

2.2 Synchronized (Monitor Lock)

Das Keyword `synchronized` belegt einen Lock für das Objekt.

→ Nur ein Thread kann eine `synchronized` Methode in derselben Instanz zur gleichen Zeit ausführen.

```
// deposit und withdraw sind im gegenseitigem Ausschluss
class BankAccount {
    private int balance = 0;
    public synchronized void deposit (int amount) {
        this.balance += amount;
    }
    public synchronized boolean withdraw (int amount) {
        if (amount <= this.balance) {
            this.balance -= amount;
            return true;
        } else {return false; } }
}
// synchronized(object){ statements } ist eine explizite
// Angabe, auf welcher Instanz gelockt wird:
public void deposit (int amount) {
    synchronized(this) {this.balance += amount; } }
// Problem: Wenn man auf Bedingung warten will. → sleep() und
// yield() geben den Monitor-Lock nicht frei.
// Lösung: Wait & Signal Mechanismus. → Threads können im
// Monitor auf Bedingung warten oder können Wartende Threads
// aufwecken (Änderung signalisieren)
// wait() gibt lock frei und wartet auf Signal. notify() signalisiert
// nur einen beliebigen (zufälligen), wartenden Thread. → Wartet
// vlt auf andere Bedingung und kann zu ewigem Warten führen
// notifyAll() signalisiert alle wartenden Threads.
```

**Wichtig:**

- Signalisierender Thread behält Monitor
- Thread ist im inneren Warteraum, bis Signal kommt
- Bei `notify()` kommt Thread wieder in äusseren Warteraum

```
class BoundedBuffer <T> {
    private Queue<T> queue = new LinkedList<>();
    private int limit = 1; // or initialize in constructor
    public synchronized void put (T item) throws
        InterruptedException {
        while (queue.size () == limit) {
            wait(); // await non full
        } queue.add(item); notifyAll(); } // signal non empty
    public synchronized T get() throws
        InterruptedException {
        while (queue.size () == 0) {
            wait();// await non empty
        } var item = queue.remove(); notifyAll(); // non full
        return item; } }
}
```

3 Spezifische Synchronisationsprimitiven

3.1 Semaphore

**Ziel:** Vergabe einer beschränkten Anzahl freier Ressourcen. Semaphore ist also ein Objekt mit Zähler.

**acquire():** Bezieht freie Ressource. Wartet, falls keine verfügbar (Zähler <= 0). Dekrementiert Zähler wenn erfolgreich.

**release():** Ressource freigeben. Zähler inkrementieren.

**Verwendung:** `new Semaphore(N)`; oder `new Semaphore(N, true)`; → true für Fairness (FIFO-Prinzip), aber langsamer

**Mutex Semaphore als alternative zu synchronized**

```
class BoundedBuffer <T> {
    private Queue<T> queue = new LinkedList<>();
    private Semaphore upperLimit=new Semaphore(cap, true);
    private Semaphore lowerLimit = new Semaphore(0, true);
    // private Semaphore mutex = new Semaphore(1, true);
    public void put (T item) throws InterruptedException {
        upperLimit.acquire();
        synchronized (queue) {queue.add (item);}
        // mutex.acquire(); queue.add(item); mutex.release();
        lowerLimit.release(); }
    public T get() throws InterruptedException {
        T item; lowerLimit.acquire();
        synchronized (queue) { item = queue.remove(); }
        // mutex.acquire(); T item = queue.remove(); mutex.
        // release();
        upperLimit.release(); return item; } }
}
```

3.2 Lock & Condition

**Ziel:** Monitor mit mehreren Wartelisten für verschiedene Bedingungen. Braucht kein `synchronized`, `wait`, `notify`, `notifyAll`

```
class BoundedBuffer<T> {
    private Queue<T> queue = new LinkedList<>();
    private Lock monitor = new ReentrantLock(true); //fair
    private Condition nonFull = monitor.newCondition();
    private Condition nonEmpty = monitor.newCondition();
    public void put(T item) throws InterruptedException {
        monitor.lock();
        try {
            while (queue.size() == Capacity) {nonFull.await();}
            queue.add(item); nonEmpty.signal();
        } finally { monitor.unlock(); } }
}
```

3.3 Read-Write Lock

**Ziel:** Gegenseitiger Ausschluss ist unnötig streng für rein lesende Abschnitte. → Erlaube parallele Lese-Zugriffe, Gegenseitiger Ausschluss bei Schreiben.

```
var rwLock = new ReentrantReadWriteLock(true);
rwLock.readLock().lock(); // read-only accesses
rwLock.readLock().unlock();
rwLock.writeLock().lock(); // read and write accesses
rwLock.writeLock().unlock();
```

3.4 Count Down Latch

**Ziel:** Synchronisationsprimitive mit Count Down Zähler. Threads können warten, bis Zähler <= 0 ist. `countDown()`, um Zähler zu dekrementieren aber **kein** `countUp()`.

3.5 Cyclic Barrier

**Ziel:** Treffpunkt für fixe Anzahl Threads. Ist wiederverwendbar.

```
// CountDown Latch
var ready = new CountDownLatch(N); //Warte auf N Threads
var start = new CountDownLatch(1); // Einer gibt Signal
// Cyclic Barrier
var start = new CyclicBarrier(N); // Kein start nötig
```

4 Gefahren der Nebenläufigkeit

4.1 Race Condition

**Problem:** Ungenügend synchronisierte Zugriffe auf gemeinsame Ressourcen. → Falsche Resultate oder falsches Verhalten möglich. (meistens wegen Data-Race)

**Ohne Data Race weil:** Critical Sections nicht geschützt.

**Kombinationen:**

**Data Race + Race Condition:** Fehlerhaftes Programmverhalten

**Data Race ohne Race Condition:** Korrektes Programmverhalten, aber formal falsch

**Race Condition ohne Data Race:** Fehlerhaftes Prog.verhalten

**Weder noch:** Richtig

**Wann kann man auf Synchronisation verzichten?**

**Immutability** (Unveränderlichkeit): Instanzvariablen sind alle final, Methoden mit nur Lesezugriff → Konstruktor initialisiert die Instanzvariablen.

**Confinement** (Einsperrung): Objekt gehört nur einem Thread zu einer Zeit. Oder Objekt ist in anderem bereits synchronisiertem Objekt eingekapselt

4.2 Deadlocks

**Problem:** Gegenseitiges Aussperren von Threads.

```
class BankAccount {
    private int balance;
    public synchronized void transfer(Acc to, int amount)
    {
        balance -= amount;
        to.deposit(amount); } // implizit geschachtelter Lock
    public synchronized void deposit (int amount) {
        balance += amount; } }
// Thread 1 Thread 2
a.transfer(b, 20); b.transfer(a, 50);
```

Deadlock Vermeidung:

- Lineare Sperrordnung der Ressourcen einführen → Nur geschachtelt in aufsteigender reihenfolge sperren
- Grobgranulare Cores wählen → Wenn lineare nicht möglich

4.3 Starvation

**Problem:** Kontinuierliche Fortschrittsbehinderung von Threads wegen Fairness-Problem. → Ist ein Liveness/Fairness Problem

**Starvation Vermeidung:**

- Länger wartende Threads haben Vortritt
- Fairness einschalten mit **Semaphore, Lock & Condition, read-Write Lock**
- Java Monitor hat ein Fairness-Problem

5 Thread Pools

5.1 Konzept und Funktionsweise

**Tasks:** Implementieren potentiell parallele Arbeitspakete. Auszuführende Tasks werden in Warteschlange eingereiht.

**Thread Pool:** Beschränke Anzahl von Worker-Threads. Holen Tasks aus der Queue und führen sie aus.

5.2 Vorteile und Einschränkungen

**Vorteile:**

- Beschränkte Threadanzahl (Zu viele verlangsamten System)
- Recycling der Threads (Thread-Erzeugung und Freigabe)
- Höhere Abstraktion (Trenne TaskBeschreibung/Ausführung)
- Anzahl Threads pro System konfigurierbar

**Einschränkungen:**

- Tasks dürfen nicht aufeinander warten (Deadlock)
- Task muss zu Ende laufen, bevor Worker Thread anderen Task ausführen kann (Ausnahme: Geschachtelte Tasks)

5.3 Fork & Join Pool

**Future Konzept:** Repräsentiert ein zukünftiges Resultat. Proxy wartet auf Resultat, muss das Ende der Berechnung abwarten.

```
// Future Verwendung
var threadPool = new ForkJoinPool();
Future<Integer> future = threadPool.submit(() -> {
    return value; }); // Tasks können Rückgabe haben
int result = future.get() // Blockiert, bis Task beendet
// Einfaches Beispiel:
var left = threadPool.submit(() -> count(leftPart));
var right = threadPool.submit(() -> count(rightPart));
result = left.get() + right.get();
```

5.3.1 Rekursive Tasks

Tasks können Untertasks starten und abwarten. Erben von `RecursiveTask<T>`

- `T compute():` Task Implementierung
- `fork():` Starte als Sub-Task in einem anderen Task
- `T join():` Warte auf Task-Ende und frage Resultat ab
- `T invoke():` Ein Sub-Task starten und abwarten
- `invokeAll():` Mehrere Sub-Tasks starten und abwarten

```
// Eventuell THRESHOLD einbauen und zwischen parallel
// und sequentiell unterscheiden
class CountTask extends RecursiveTask<Integer> {
    private final int lower, upper;
    public CountTask(int lower, int upper) {
        this.lower = lower; this.upper = upper; }
    protected Integer compute() {
        if (lower == upper) { return 0; }
        if (lower+1 == upper) {return isPrime(lower)? 1 : 0;}
        int middle = (lower + upper) / 2;
        var left = new CountTask (lower, middle);
        var right = new CountTask (middle, upper);
        left.for(); right.fork();
        return right.join () + left.join(); } }
}
```

5.4 Asynchrone Programmierung

**Ziel:** Aufrufer soll während der Operation weiterarbeiten. → Operation in Thread oder Thread Pool auslagern

```
// Klassisch:
Future<int> future = threadPool.submit(() -> doSmth());
// other work
process(future.get()); // Resultat über future
// Modern: Starte asynchrone Aufgabe in Standard Pool
- ForkJoinPool.commonPool();
CompletableFuture<int> future = CompletableFuture
    .supplyAsync (()->doSmth());//runAsync falls kein return
//other work
process(future.get());
```

5.4.1 Continuation

**Ziel:** Folgeaufgabe an asynchrone Aufgabe anhängen. Ausführung der Continuation durch beliebigen Thread.

```
// Continuation:
future.thenAccept(result -> syso(result));
// Tasks verketten:
future.thenApplyAsync(second).thenAcceptAsync(third);
// Multi-Continuation:
CompletableFuture.allOf(future1, future2)
    .thenAcceptAsync(continuation); //Warten aufeinander
CompletableFuture.any(future1, future2)
    .thenAcceptAsync(continuation)//Sobald einer fertig ist
// runAsync() ist Fire and Forget. Workers sind Deamons
// und ignoriert Exceptions
```

6 Task Library
<b>6.1 Threading in .NET</b>
Keine Vererbung: Delegate bei Konstruktor. Exception in Thread führt zu Abbruch des Programms. <b>.NET Monitor:</b> FIFO Warteschlange, <b>wait()</b> in schlaufe, <b>PulseAll()</b> für aufwecken, Synchronisation mit Hilfsobjekt ist Best Practice, Kein Fairness-Flag, Kein Lock & Condition. <b>Zusätzlich:</b> ReadWriteLockSlim für Upgradeable Read/Write, Semaphoren auch auf OS-Stufe nutzbar, Mutex (Binärer Semaphore auf OS-Stufe) var myThread = <b>new</b> Thread(() => { <b>for</b> ( <b>int</b> i = 0; i < 100; i++) { Console.WriteLine("MyThread step {0}", i); } }); myThread.Start(); myThread.Join(); // <b>Monitor in .NET</b> <b>class</b> BankAccount { <b>private</b> decimal balance; <b>private</b> object syncObject = <b>new</b> () //Monitor Hilfsobjekt <b>public void</b> Withdraw(decimal amount) { lock (syncObject) { // Analog zu synchronized <b>while</b> (amount > balance) { Monitor.Wait(syncObject); } balance -= amount; } } <b>public void</b> Deposit(decimal amount) { lock (syncObject) { balance += amount; Monitor.PulseAll(syncObject)}}} // wie notifyAll
<b>6.2 Task Parallelität</b>
<b>Work Stealing Thread Pool</b> mit Abstraktionsstufen: Task Parallelität: Explizite Tasks starten und warten Data Parallelität: Parallele Statements und Queries Asynchrone Programmierung: Mit Continuation Style task task = Task.Run(() => { implementation }); task.Wait() // <b>Warten auf task</b> Task< <b>int</b> > task = Task.Run(() => { <b>return</b> 2;}); Console.Write(task.Result) // <b>Wartet auf Task Ende</b>
<b>6.3 Datenparallelität</b>
// Parallele Statements wartet bis alle fertig sind Parallel.Invoke( () => MergeSort(l,m), () => MergeSort(m,r)); // <b>Parallele Loop</b> Parallel.ForEach(list, file => Convert(file)); // <b>Falls Iteration unabhängig:</b> Parallel.For(0, array.Length, i => doSmth(array[i]));
<b>6.4 Asynchrone Programmierung</b>
var task = Task.Run(LongOperation); // <b>ist asynchron</b> <b>int</b> result = task.Result; // <b>Task Continuations (Wie CompletableFutures</b> task1.ContinueWith(task2).ContinueWith(task3); // <b>Multi-Continuation</b> Task.WhenAll(task1, task2).ContinueWith(continuation); Task.WhenAny(task1, task2).ContinueWith(continuation);
7 GUI und Threading
<b>7.1 GUI &amp; Threading</b>
GUI Frameworks erlauben nur <b>Single-Threading</b> : Nur spezifischer UI-Thread darf auf UI-Komponente zugreifen. <b>GUI Implikationen:</b> Keine langen Operat. in UI Events, Kein Zugriff durch fremde Threads (blockieren & Race Condition) <b>UI Interaktion:</b> UI Operationen müssen als Events in die UI Event Queue eingereiht werden. <b>Dispatching:</b> Benutzung der Klasse <b>SwingUtilities</b> button.addActionListener(event -> { <b>new</b> Thread() -> { var text = readHugeFile(); SwingUtilities.invokeLater(() -> { // asynchron textArea.setText(text); });//synchron:invokeAndWait }).start(); } }); // <b>Sauberer Setup:</b> frame.pack() und setVisible(true) (Auf JFrame bezogen) in SwingUtilities ausführen // <b>Background Worker:</b> <b>class</b> BackgrCalc <b>extends</b> SwingWorker<Integer, Void> { @Override // <b>Int:</b> Resultat, <b>Void:</b> Zwischenresultat <b>public</b> Integer doInBackground () { <b>return</b> longComputation (); } @Override // <b>UI Thread</b> <b>protected void</b> done() {

<pre>try {     int result = get();     label.setText("Result: " + result); } <b>catch</b> (InterruptedException ExecutionException e) { } } } // <b>get()</b> gibt <b>Resultat von doInBackground</b> <b>.NET UI Thread Modell:</b> Gleiches Prinzip wie in Java. // <b>UI Thread</b> ist der Aufrufen von Application.Run() // <b>UI Event Dispatching:</b> control.Dispatcher.InvokeAsync(action); // WPF control.BeginInvoke(delegate) // WinForm // <b>Sind beide asynchron. Synchron mit Invoke</b></pre>
<b>7.2 C# async/await</b>
<b>async</b> für Methode (Aufrufer wird blockiert). <b>await</b> für Tasks (Warten auf Ende eines TPL Task). <b>Rückgabetypen:</b> <b>void:</b> Fire & Forget, <b>Task:</b> Keine Rückgabe, erlaubt Warten, <b>Task&lt;T&gt;</b> Rückgabetyp T. <b>Ausführungsmodell:</b> Aufrufer führt <b>async</b> Methode <b>synchron</b> aus, bis ein blockierendes <b>await</b> anliegt. Erst danach <b>asynchron</b> . <b>async Task&lt;string&gt;</b> ConcatWebSitesAsync(string url1, string url2) { HttpClient client = <b>new</b> HttpClient(); Task<string> download1 = client.GetStringAsync(url1); Task<string> download2 = client.GetStringAsync(url2); string site1 = await download1; string site2 = await download2; <b>return</b> site1 + site2; }
8 Memory Models
<b>8.1 Java Memory Model</b>
<b>Problem bei Weak Consistency:</b> Speicherzugriffe werden in verschiedenen Reihenfolgen von verschiedenen Threads gemacht. (Ausnahme: Synchronisationen/Speicherbarrieren) <b>Problem bei Opmptimierungen:</b> Betrifft Compiler, Laufzeitsystem und CPUs. Sie ordnen Instruktionen um → Keine sequentielle Konsistenz bei Nebenläufigkeit <b>Atomicity:</b> Einzelnes lesen und schreiben ist atomar bei primitiven Datentypen bis 32 Bit → Sonst <b>volatile</b> Keyword. <b>Visibility:</b> Sieht Änderungen eines anderen Thread eventuell nicht oder viel später (z.B. wegen Compiler-/JIT-Optimierung) <b>Visibility Garantien:</b> <ul style="list-style-type: none"><li>• Locks Release &amp; Acquire: Änderungen vor Release werden bei Acquire sichtbar</li><li>• Volatile Variable: Änderungen bis zum write werden beim read sichtbar</li><li>• Thread/Task-Start und Join: Bei Start Eingabe und bei Join Ausgabe</li><li>• Final Variablen: Nach Ende des Konstruktors sichtbar</li></ul> <b>Ordering Garantien:</b> <ul style="list-style-type: none"><li>• Program Order: Sequentielles Verhalten jedes einzelnen Thread bleibt erhalten</li><li>• Synchronization Order: Syncbefehle werden zueinander nie ungeordnet (Locking, volatile, Thread start/join)</li><li>• Happens-Before Relation: Alles andere kann ungeordnet werden (ausser garantierte Sichtbarkeit unter Threads)</li></ul> <b>Atomic Klassen:</b> Gibt es für <b>Boolean, Integer, Long und Referenzen</b> (auch für Array Elemente). Diverse atomare Operationen wie <b>addAndGet(), getAndAdd()</b> etc. <b>public class</b> SpinLock { <b>private</b> AtomicBoolean locked= <b>new</b> AtomicBoolean( <b>false</b> ); <b>public void</b> acquire() { <b>while</b> (locked.getAndSet( <b>true</b> )) { } <b>public void</b> release() { locked.set( <b>false</b> ); } } <b>Lock-freie Datenstrukturen:</b> ConcurrentLinkedQueue<V>, ConcurrentLinkedDeque<V>, ConcurrentSkipListSet<V>, ConcurrentHashMap<K, V>, ConcurrentSkipListMap<K, V>
<b>8.2 .NET Memory Model</b>
<b>Unterschied zu Java Memory Model:</b> <b>Atomicity:</b> long/double nicht mit volatile atomar. <b>Visibility:</b> Nicht definiert, implizit durch Ordering. <b>Ordering:</b> Nur Half und Full Fences → Atomare Instruktionen mit <b>Interlocked</b> Klasse <b>.NET Volatile Half Fences:</b> <b>Volatile Write:</b> Vorangehende Zugriffe bleiben davor <b>Volatile Read:</b> Nachfolgende Zugriffe bleiben danach <b>.NET Volatile Full Fences:</b> Thread.MemoryBarrier(); → Verbietet Umordnung in beide Richtungen.
9 Actor Model
<b>Aktive Objekte:</b> Objekte haben nebenläufiges Innenleben <b>Kommunikation:</b> Objekte senden und empfangen Nachrichten <b>Kein Shared Memory:</b> Nur Austausch von Nachrichten über

Kanäle/Mailbox <b>Akka Empfangsverhalten:</b> <b>Reaktion auf ankommende Nachricht:</b> Spezielle Behandlungsmethode wird ausgeführt. <b>Effekte per Behandlung:</b> Ändere privaten Zustand, Sende nachrichten, Erzeuge neue Actors. <b>Intern sequentiell:</b> Nur eine Nachricht auf einmal bedienbar (Buffer) // <b>Einfacher Akka Actor</b> <b>public class</b> NumberPrinter <b>extends</b> UntypedActor { <b>public void</b> onReceive( <b>final</b> Object message) { <b>if</b> (message <b>instanceof</b> Integer) { System.out.println(message); } } // <b>Erzeugen und Senden</b> ActorSystem system = ActorSystem.create("System"); ActorRef printer = system.actorOf(Props.create( NumberPrinter. <b>class</b> )); <b>for</b> ( <b>int</b> i = 0; i < 100; i++) { printer.tell(i, ActorRef.noSender()); } system.shutdown();
10 GPU Parallelisierung
Cores innerhalb Streaming Multiprocessor sind nur mit Vektor Parallelisierung effizient nutzbar. → Cores führen dieselbe Instruktion auf unterschiedlichen Daten/Speicherstellen aus <b>NUMA Modell:</b> Non-Uniform Memory Access <b>Kein gemeinsamer Hauptspeicher zwischen GPU und CPU</b> → Explizites Übertragen. <b>Unterschiedlicher Instruktionssatz/Architektur</b> → Code für GPU kompilieren/designen
<b>10.1 CUDA (Computer Unified Device Architecture)</b>
<b>CUDA Threads:</b> Gleicher Kernel wird von mehreren Threads ausgeführt. <b>SIMT:</b> Single Instruction Multiple Threads. <b>CUDA Blocks:</b> Threads sind in Blöcke gruppiert. → Ein Block: Gleicher SM, Threads können innerhalb von Block interagieren. <b>CUDA Ausführungsmodell:</b> <b>Thread:</b> Virtueller Skalarprozessor. <b>Block:</b> Virtueller Multiprozessor. <b>Blöcke müssen unabhängig sein:</b> Run To Completion, Beliebige Ausführungsreihenfolge (sequentiell, parallel), Blocks in Threadpool <b>Datenaufteilung:</b> Jede Kernel-Ausführung bestimmt seinen Datenteil. <b>threadIdx.x:</b> Nummer des Threads innerhalb Block, <b>blockIdx.x:</b> Nummer des Blocks, <b>blockDim.x:</b> Blockgrösse // <b>CUDA Kernel</b> __global__ // <b>Läuft auf GPU (Device)</b> <b>void</b> VectorAddKernel ( <b>float</b> *A, <b>float</b> *B, <b>float</b> *C, <b>int</b> N){ <b>int</b> i = threadIdx.x * blockDim.x + threadIdx.x; // eindeutiger Index basierend auf Block & Thread ID <b>if</b> (i < N) { C[i] = A[i] + B[i]; } } <b>int</b> main() { // <b>Läuft auf CPU (Host)</b> //kernel invocation <b>int</b> blockSize = 1024; <b>int</b> gridSize = (N + blockSize -1) / blockSize; VectorAddKernel<<<gridSize, blockSize>>>(A, B, C, N);} <b>CUDA Ausführung:</b> <ul style="list-style-type: none"><li>• Auf GPU allozieren: cudaMalloc</li><li>• Daten zu GPU transferieren: cudaMemcpy</li><li>• Kernel ausführen: &lt;&lt;&lt;gridD, blockD&gt;&gt;&gt;</li><li>• Daten von GPU zu PU transferieren: cudaMemcpy</li><li>• Auf GPU deallozieren: cudaFree</li></ul> <b>void</b> CudaVectorAdd( <b>float</b> * A, <b>float</b> * B, <b>float</b> * C, <b>int</b> N){ size_t size = N * sizeof( <b>float</b> ); <b>float</b> *d_A, *d_B, *d_C; cudaMalloc(&d_A, size); cudaMalloc(&d_B, size); cudaMalloc(&d_C, size); cudaMemcpy(d_A, A, size, cudaMemcpyHostToDevice); cudaMemcpy(d_B, B, size, cudaMemcpyHostToDevice); <b>int</b> blockSize = 1024; <b>int</b> gridSize = (N + blockSize - 1) / blockSize; VectorAddKerne<<<gridSize,blockSize>>>(&d_A,&d_B,&d_C,N); cudaMemcpy(C, &d_C, size, cudaMemcpyDeviceToHost); cudaFree(&d_A); cudaFree(&d_B); cudaFree(&d_C); //handleCudaError(...) um Malloc, Mempcy und Free
11 GPU Parallelisierung 2
<b>Speichermodell:</b> Global Memory ist relativ teuer (ca. 600 Zyklen). Die Threads lesen bei Matrix Multiplikation wiederholt die selben Elemente von A & B. → Effizienzsteigerung mit Cache Speicher <b>Synchronisation:</b> <b>syncthreads()</b> synchronisiert alle Threads innerhalb eines Blocks. (Nicht zwischen Blöcken) <b>Warps:</b> Ein Block wird intern in Wraps zerlegt (zu je 32 Threads). Alle Threads führen die gleiche Instruktion aus. Verzweigungen (if, switch, while, do, for) werden abwechselnd ausgeführt (Divergenz) → Einzelne Threads müssen warten

// multiple of 32 <b>const int</b> C_ROWS = 1024; <b>const int</b> C_COLS = 2048; <b>const int</b> A_COLS = 3072; <b>const int</b> A_ROWS = C_ROWS; <b>const int</b> B_ROWS = A_COLS; <b>const int</b> B_COLS = C_COLS; __shared__ <b>float</b> Asub[TILE_SIZE][TILE_SIZE]; __shared__ <b>float</b> Bsub[TILE_SIZE][TILE_SIZE]; <b>int</b> tx = threadIdx.x, ty = threadIdx.y; <b>int</b> col = blockIdx.x * TILE_SIZE + tx; <b>int</b> row = blockIdx.y * TILE_SIZE + ty; <b>int</b> nofTiles = (A_COLS + TILE_SIZE - 1) / TILE_SIZE; <b>float</b> sum = 0.0; <b>for</b> ( <b>int</b> tile = 0; tile < nofTiles; tile++) { Asub[ty][tx] = A[row*A_COLS + tile*TILE_SIZE + tx]; Bsub[ty][tx] = B[(tile*TILE_SIZE + ty)*B_COLS + col]; __syncthreads(); <b>for</b> ( <b>int</b> ksub = 0; ksub < TILE_SIZE; ksub++) { sum += Asub[ty][ksub] * Bsub[ksub][tx]; } __syncthreads(); } C[row * C_COLS + col ] = sum;
12 Cluster Parallelisierung
13 Concurrency in Python & JavaScript