

Multithreading Grundlagen
Multi-Thread Programmierung
JVM Thread Modell
Java ist ein Single Process System. JVM ist ein Prozess im Bsys. Main-Thread wird beim Aufstarten der JVM anhand <i>main()</i> Methode erzeugt. JVM läuft, solange Threads laufen (Ausnahme Daemon Threads).
var myThread = new Thread() -> (/* ... */); myThread.start();
Thread wird erst bei <i>start()</i> erzeugt. Führt <i>run()</i> -Methode des Runnable Interface aus. Thread endet beim Verlassen von <i>run()</i> . Nicht-Determinismus: Threads laufen ohne Vorkehrungen beliebig verzahnt oder parallel.
Explizite Runnable-Implementation:
<pre>class SimpleLogic implements Runnable { @Override public void run() { /* ... */ } } var myThread = new Thread(new SimpleLogic()).start();</pre>
Sub-Klasse von Thread
<pre>class SimpleThread extends Thread { @Override public void run() { /* ... */ } } var myThread = new SimpleThread().start();</pre>
Thread Join
Warten auf Beendigung eines Threads. <i>t2.join()</i> blockiert, solange t2 läuft.
Thread Passivierung
Thraed.sleep(ms): Laufender Thread geht in Wartezustand, dann ready. Thread.yield(): Gibt Prozessor frei, direkt ready.
Monitor Synchronisation
Java Synchronized Methoden
synchronized void f() (/* ... */) // <i>Object Lock</i> static synchronized void g() (/* ... */) // <i>Class Lock</i>
Monitor
Objekt mit internem gegenseitigem Ausschluss. Nur 1 Thread operiert im Monitor. Alle äusseren Methoden synchronized. Wait & Signal Mechanismus: Threads können im Monitor auf Bedingung warten und wartende aufwecken (signal).
<pre>// <i>CyclicBarrier</i> private final int parties; private int entered, exited = 0; public CyclicBarrier(int parties) { this.parties = parties; } public synchronized void await() throws InterruptedException { entered++; if(entered == parties) { exited = 0; notifyAll(); } while (entered < parties) { wait(); } if(exited == parties) { entered = 0; notifyAll(); } while (exited < parties) { wait(); } }</pre>
notify(): Bei Uniform Waiters & One-In-One-Out Bedingungen. notifyAll(): Bei mehreren Bedingungen / One-In-Multiple-Out. Pauschales wait & signal: Wartende müssen selber schauen, ob sie ein Signal interessiert. Signal and Continue: Signalisierender Thread behält Monitor nach notify. Aufgeweckter Thread muss um Monitor-Eintritt kämpfen.
Spezifische Synchronisationsprimitiven
Arten
Faire Semaphore: <i>new Semaphore(N, true)</i> , FIFO Warteschlange, langsamer.
<pre>private Semaphore upperL = new Semaphore(CAP, true); private Semaphore lowerL = new Semaphore(0, true); public void put(T item) throws InterruptedException { upperL.acquire(); synchronized (queue) { queue.add(item); } lowerL.release(); }</pre>
<pre>public T get() throws InterruptedException { T item; lowerL.aquire(); synchronized (queue) { item = queue.remove(); } upperL.release(); return item; }</pre>
Multi-Acquire/Release: <i>acquire(N)</i> .
Lock & Condition
Monitor mit mehreren Wartelisten für verschiedene Bedingungen. Lock-Objekt: Sperre für Eintritt in Monitor. Condition-Objekt: Wait & Signal für bestimmte Bedingung.
<pre>private Lock monitor = new ReentrantLock(true); // <i>fair</i> 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(); } }</pre>

<pre>public T get() throws InterruptedException { monitor.lock(); try { while (queue.size() == 0) { nonEmpty.await(); } T item = queue.remove(); nonFull.signal(); return item; } finally { monitor.unlock(); } }</pre>
Read-Write Locks
Gegenseitiger Ausschluss ist unnötig streng für lesende Abschnitte. Erlaube parallele Lese-Zugriffe.
<pre>private Collection<String> names = new HashSet<>(); private ReadWriteLock rwLock = new ReentrantReadWriteLock(); public boolean exists(String pattern) { // <i>Read-only accesses</i> rwLock.readLock().lock(); try { return names.steram().anyMatch(n -> n.matches(pattern)); } finally { rwLock.readLock().unlock(); } } public void insert(String name) { // <i>Write / Read accesses</i> rwLock.writeLock().lock(); try { names.add(name); } finally { rwLock.writeLock().unlock(); } }</pre>
Count Down Latch
Synchronisationsprimitive mit Count Down Zähler. Threads können warten, bis Zähler <= 0 wird. await(): warten bis Count Down 0 ist. countDown(): Zähler - 1.
<pre>var ready = new CountDownLatch(N); // <i>Warte auf N cars</i> var start = new CountDownLatch(1); // <i>Einer gibt signal</i> // <i>N Cars:</i> ready.countDown(); start.await(); // <i>RaceControl:</i> ready.await(); start.countDown();</pre>
Cyclic Barrier
Treffpunkt für fixe Anzahl Threads. Anzahl treffender Threads muss vorgeben sein. Ist wiederwendbar (mehrere Runden).
<pre>var start = new CyclicBarrier(N); // <i>Treffende Autos</i> // <i>N Cars:</i> start.await(); // <i>braucht kein Race Control mehr</i> var gameRound = new CyclicBarrier(N); // <i>N Players:</i> while(true) { gameRound.await(); // <i>play concurrently with others</i> }</pre>
<pre>// <i>Mit Austausch</i> Exchanger.exchange(something); // <i>Blockiert bis anderer Thread auch exchange() aufruft.</i></pre>
Gefahren der Nebenläufigkeit
Neue Arten von Programmierfehler, die es bei single-Threading nicht gibt. Können sporadisch oder selten auftreten. Sehr schwierig durch Tests zu finden.
Race Condition
Mehrere Threads greifen auf gemeinsame Ressourcen ohne genügend synchronisation zu. Mögliche falsche Resultate oder falsches Verhalten (Lost Updates). Ursache oft ein Data Race, nicht immer. Keine Atomarität bei Time of Check / Time of Use
Data Race
Unsynchronisierter Zugriff auf gleichen Speicher. Selbe Variable oder Array Element (min. 1 schreibender Zugriff).
Race Condition ohne Data Race
Critical Sections nicht geschützt. Data Races mit Synchronisation eliminiert, aber nicht genügend grosse synchronisierte Blöcke.
<pre>synchronized int getBalance() { return balance; } synchronized void setBalance(int x) { balance = x; } // <i>Mehrere Threads, Kein Atomares Inc - Lost Update moeglich</i> account.setBalance(account.getBalance() + 100);</pre>
Kombinationen
Alles Synchronisieren? Hilft nichts. Race Condition trotzdem möglich. Weitere Nebenlaufigkeitsfehler. Synchronisationskosten sind relativ teuer.
Synchronisation: Verzichtbare Fälle
Immutability (Unveränderlichkeit): Objekte mit nur lesendem Zugriff. Confinement (Einsperrung): Objekt gehört nur einem Thread zu einer Zeit.
Immutable Objects
Instanzvariablen alle <i>final</i> . Primitive Datentypen. Referenzen wiederum auf Immutable Objekte. Methoden mit nur lesendem Zugriff. Konstruktor initialisiert Instanzvariablen. Nach Konstruktor kann Objekt ohne Synchronisation von Threads verwendet werden.
Confinement
Struktur garantiert, dass auf ein Objekt nur durch einen Thread zur gleichen Zeit zugegriffen wird. Thread Confinement: Objekt gehört nur einem Thread und wird nur von demjenigen verwendet. Object Confinement: Objekt in anderem bereits synchronisierten Objekt eingekapselt.
Kapselungsbrüche: 1. Inneres Objekt ist aussen zugreifbar. 2. Rückgabe einer Referenz auf inneres Objekt. 3. Holder installiert selber Referenz ausserhalb. 4. Inneres Objekt gibt selber <i>this</i> raus.
Thread Safety
Klassen / Methoden, die intern synchronisiert sind. Keine Race Conditions innerhalb dieses Codes. Kritischer Abschnitt nur pro Methode erfüllt. Aber: Kein kritischer Abschnitt über mehrere Methodenaufrufe. Andere Nebenlaufigkeitsfehler möglich.

Java Collections - Thread Safety
Alte Java 1.0 Collections (Vector, Stack, Hashtable): JA . Moderne Collections (HashSet, TreeSet, ArrayList, etc.): NEIN . Concurrent Collections (ConcurrentHashMap, etc.): JA .
Verstärktes Multi-Threading
Finalizers: Laufen über separaten Finalizer-Thread. Timers: Handler durch separaten Thread ausgeführt (ausser GUI). External Libraries & Frameworks: z.B. Abarbeitung von Web-Service Aufrufen.
Deadlock
Beide Threads sperren sich gegenseitig aus:
Spezialfall: LiveLocks
Threads haben sich gegenseitig permanent blockiert. Führen aber noch Wartestruktionen aus. Verbrauchen CPU während Deadlock.
<pre>// <i>Thread 1</i> b = false; while (!a) { } b = true; // <i>Thread 2</i> a = false; while (!b) { } a = true;</pre>
Deadlock Erkennung
Deadlock = Zyklus im Betriebsmittelgraph
Deadlock Voraussetzungen: Geschachtelte Locks, Zyklische Wartabhängigkeiten
Deadlock Vermeidung
Lineare Sperrordnung der Ressourcen einführen. Nur geschachtelt in aufsteigender Reihenfolge sperren. Eliminiert zyklische Wartabhängigkeiten.
Grobgranulare Locks wählen. Wenn lineare Sperrordnung nicht möglich/sinnvoll ist. Sperre gesamte Bank bei Kontenzugriff. Eliminiert Schachtelung von Locks.
Starvation
Ein Thread kriegt nie die Chance, auf eine Ressource zuzugreifen, obwohl sie immer wieder frei wird. Andere Threads überholen andauernd. Liveness/Fairness Problem.
<pre>do { // <i>Starvation moeglich</i> success = account.withdraw(100); } while (!success);</pre>
Vermeidung
Faire Synchronisationskonstrukte (bei Semaphore, Lock & Condition, ReadwriteLock möglich). Java Monitor hat ein Fairness Problem (Starvation anfällig).
Parallelität Korrektheitskriterien
Safety: Keine Race Conditions, Keine Deadlocks. Liveness: Keine Starvation.
Thread Pool
Work-Stealing: Vorteil: Effizienz durch weniger Contention. Nachteil: Fairnessprobleme bei unausgeglichenen Verteilungen.
Vorteile
Beschränkte Anzahl von Threads: Viele Threads verlangsamen das System oder überschreiten Speicherlimit. Recycling der Threads: Spare Thread-Erzeugung und Freigabe. Höhere Abstraktion: Vereinfacht Task-Beschreibung von Task-Ausführung. Anzahl Threads pro System konfigurierbar: <i>#Worker Threads = #Prozessoren + #I/O-Aufrufe</i>
Einschränkung
Dæmon: Laufen nicht garantiert zu Ende bei Fire & Forget. Tasks dürfen nicht aufeinander warten, sonst Deadlock. Run to Completion: Task muss zu Ende laufen, bevor Worker Thread anderen Task ausführen kann. Ausnahme: geschachtelte Tasks.
Java Fork-Join-Pool
<pre>var threadPool = new ForkJoinPool(); Future<Integer> future = threadPool.submit() -> { }; Int result = future.get(); // <i>Blockiert bis Task beendet</i></pre>
Rekursive Task Erstellung
Anzahl Tasks: Java: Arraylänge, THRESHOLD. .NET: <i># Freien Worker Threads falls < Arraylänge</i>
<pre>class PairSwapTask extends RecursiveTask<Integer> { private static final int THRESHOLD = 1000; private final int[] array; private final int from, to; public PairSwapTask(int[] array, int from, int to) { this.array = array; this.from = from, this.to = to; } @Override protected Integer compute() { if(to - from <= THRESHOLD) { for(int i = from; i < to; i++) { int offset = 2 * i; int temp = array[i]; array[offset] = array[offset + 1]; array[offset + 1] = temp; } } else { int middle = (from + to) / 2; // <i>l = new, r = new, fork(), fork(), return join() + join();</i> invokeAll(new PairSwapTask(array, from, middle); new PairSwapTask(array, middle, to);); } } }</pre>
<pre>// <i>Ausfuehrung</i> var threadPool = new ForkJoinPool(); threadPool.invoke(new PairSwapTask(array, 0, array.length / 2));</pre>
Fork Join Pool Internals
Automatischer Parallelitätsgrad: Default: <i>#Worker Threads = #Prozessoren</i> Dynamisches Hinzufügen / Wegnehmen von Threads
Common Pool: Verhindert Engpässe durch zu viele Thread Pools.

Asynchrone Programmierung
Unnötige Synchronität: Langlaufende Rechnungen, I/O Aufrufe. Asynchroner Aufruf: Aufrufer soll während der Operation weitermachen.
<pre>// <i>runAsync()</i> falls kein Rueckgabewert CompletableFuture<long> future = CompletableFuture.supplyAsync(()->{}); // <i>other work</i> process(future.get());</pre>
Ende des async Aufrufs
Caller-zentrisch (Pull): Caller warted auf Task-Ende und holt sich Resultat, Future abfragen. Callee-zentrisch (Push): Async Operation informiert direkt über Resultat. Completion Callback.
Continuation
Folgeaufgabe an asynchrone Aufgabe anhängen.
<pre>// <i>thenApply()</i> fuer Continuation mit Rueckgabe future.thenAccept(res -> System.out.println(res));</pre>
Ausführung: durch beliebigen Thread, durch Initiator, wenn Future bereits Resultat hat. Asynchrone Continuations: <i>thenAcceptAsync()</i> bzw. <i>thenApplyAsync()</i> .
Multi-Continuation
<pre>// <i>allOf(futures).thenRun()</i> -> <i>syso()</i>; CompletableFuture.allOf(f1, f2).thenAcceptAsync(() -> { }); CompletableFuture.any(f1, f2).thenAcceptAsync(() -> { });</pre>
Fire and Forget
Task starten, ohne das Ende abzuwarten. Submitter ruft <i>get()</i> oder <i>join()</i> auf.
<pre>CompletableFuture.runAsync(() -> { });</pre>
Dæmon Workers: Workers Threads in Fork-Join-Pools sind Dæmon. Anwendung kann vor Task-Ende stoppen.
Inorgierte Exceptions: Exceptions in Fire & Forget Task werden ignoriert.
Task Parallel Library
.NET Threads
Keine Vererbung: Delegate bei Konstruktor. Exception in Thread: Abbruch des Programs.
<pre>var myThread = new Thread() => { /* ... */ }; myThread.start(); /* ... */ myThread.join();</pre>
C# Lambda kann umgebende Variablen zugreifen (auch scrheidend).
Monitor in .NET
FIFO Warteschlange, Pulse informiert längst Wartenden. <i>Wait()</i> in Schlaufe. <i>PulseAll()</i> bi mehreren Bedingungen oder Erfüllungen mehrerer Threads. Synchronisation mit HilfsObj als Best Practice.
<pre>private object syncObject = new(); // <i>Monitor auf HilfsObj</i> public void Withdraw(decimal amount) { lock (syncObject) { while (amount > balance) { Monitor.Wait(syncObject); } balance -= amount; } } public void Deposit(decimal amount) { lock (syncObject) { balance += amount; Monitor.PulseAll(syncObject); } }</pre>
.NET Synchronisationsprimitiven
Fehlen: kein Fairnessflag, kein Lock & Condition. Zusätzlich: ReadwriteLockSlim für Upgradeable Read/Write. Semaphore auch auf OS-Stufe nutzbar. Mutex (binärer Semaphore). Collections nicht Thread-safe: Ausser <i>System.Collections.Concurrent</i>
.NET Task Parallel Library (TPL)
Work Stealing Thread Pool. Verschiedene Abstraktionsstufen: Task Parallelization: Explizite Tasks starten und warten. Data Parallelization: Parallele Statements und Queries. Asynchrone Programmierung: mit Continuation Style.
Thread Injection
TPL fügt zur Laufzeit neue Worker Threads hinzu. Hill Climbing Algorithmus: Misst Durchsatz & variiert Anzahl Worker Threads. Kein Deadlock bei Task Abhängigkeiten, aber ineffizient, nicht dafür gemacht.
Task Parallelisierung
<pre>Task task = Task.Run(() => { }); // <i>/* ... */ task.Wait();</i> // <i>Task mit Rueckgabe</i> Task<int> task = Task.Run(() => { return 0; }); Console.WriteLine(task.Result); // <i>Blockiert</i> // <i>Geschachtelte Tasks</i> Task.Run(() => { var left = Task.Run(() => { }); var right = Task.Run(() => { }); int res = left.Result + right.Result; });</pre>
Parallele Statements
Menge an Statements potentiell parallel ausführen. Als Task starten. Barriere der Tasks am Ende.
<pre>Parallel.Invoke(() => MergeSort(l, m); () => MergeSort(m, l););</pre>
Parallele Loop
Schlaufen-Bodies potentiell parallel ausführen. Gruppierung der Bodies in Tasks. Barriere dieser Tasks am Ende.
<pre>Parallel.ForEach(list, file => Convert(file)); Parallel.For(0, array.Length, i => Calc(array[i]));</pre>

Parallele Loop Partitionierung
Schlaufe mit vielen sehr kurzen Bodies ist ineffizient. TPL gruppiert automatisch mehrere Bodies zu Task. Aufteilung gemäss verfügbaren Worker Threads. Explizite Partitionierung; Vorteil: Weniger Body-Delegates, Nachteil: Künstliche Unterschleife.
Parallel. ForEach(Partitioner.Create(0, array.Length), (range, _) => { for (int i = range.Item1; i < range.Item2; i++) { Calc(array[i]); } })
Asynchrone Programmierung mit TPL
Task Continuation
task1.ContinueWith(task2).ContinueWith(task3); // Multi-Continuation Task.WhenAll(task1, task2).ContinueWith(continuation); Task. WhenAny(task1, task2).ContinueWith(continuation);

GUI and Threading
GUI Frameworks erlauben nur Single-Threading. UI Thread: Loop zur Ausführung der Ereignisse aus einer Queue.
UI Thread Confinement

Wiso basieren GUI-Frameworks auf Single-Thread Modell?
Synchronisationskosten: Locking in allen Komponenten und Methoden relativ teuer. **Deadlock-Risiko:** Bei zyklischen geschachtelten Aufrufen (z.B. MVC).

Swing: Dispatching an UI Thread
Komponentenzugriffe an UI Thread delegieren.
// Benutzung der Klasse SwingUtilities static void invokeLater(Runnable doRun); // <i>Async</i> static void invokeAndWait(Runnable doRun); // <i>Synchron</i> // Example unordered, parallel link check button.addActionListener(e -> { var url = input.getText(); CompletableFuture.runAsync(() -> { var html = webClient.download(url); // <i>blocking</i> for (var link : webClient.extractLinks(html)) { CompletableFuture.runAsync(() -> { if (!webClient.isAccessible(link)) { SwingUtilities.invokeLater(() -> output.addElement(link + " is broken")); } }) }); }); }); // Example (same order of downloads) - Recursive void download(List<URL> links, OutputStream output) { if (links.isEmpty()) { statusLabel.setText("All done!"); } var url = links.get(0); statusLabel.setText("Downloading" + url); if (cancelBox.isSelected()) { statusLabel.setText("cancelled!"); return ; } CompletableFuture.runAsync(() -> { url.openStream().transferTo(output); var remaining = links.subList(1, links.size()); SwingUtilities.invokeLater(() -> download(remaining, output)); }); }
.NET UI Threading Modell
Gleiches Prinzip wie Java. UI Thread: Aufrufer von <i>Application.Run()</i> . UI Event Dispatching: WPF: <i>control.Dispatcher.InvokeAsync(action)</i> WinForm: <i>control.BeginInvoke(delegate)</i>
Async / Await

Async Methode läuft teilweise synchron, teilweise asynchron. Aufrufer führt Methode solange synchron aus bis ein <i>await</i> anliegt. Compiler zerlegt Methode in Abschnitte. Abschnitt nach <i>await</i> läuft später nach Task-Ende (Continuation). Methode läuft synchron bis <i>await</i> , springt dann zurück zum Aufrufer.
Verschiedene Ausführungen: Fall 1: Aufrufer ist normaler Thread, Abschnitt wird durch TPL Worker-Thread ausgeführt. Fall 2: Aufrufer ist UI-Thread, Abschnitt wird als Event vom UI-Thread ausgeführt.
public async Task<int> LongOperationAsync() { } Task<int> task = LongOperationAsync(); /* ... */ int res = await task; // Warte auf Beendigung
Async Rückgabetypen: <i>void</i> : fire-and-forget. <i>Task</i> : Kein Rückgabetyp erlaubt warten. <i>Task<T></i> : Rückgabetyp T.
async Task<string> ConcatAsync(string url1, string url2) { HttpClient client = new HttpClient(); Task<string> d1 = client.GetStringAsync(url1); Task<string> d2 = client.GetStringAsync(url2); string site1 = await d1; string site2 = await d2; return site1 + site2; }
Memory Models

Lock-Freie Programmierung: Korrekte nebenläufige Interaktionen ohne Locks. Garantien des Speichermodells nutzen.

Ursachen für Probleme

Weak Consistency: Speicherzugriffe werden in verschiedenen Reihenfolgen aus verschiedenen Threads gesehen. Ausnahme: Synchronisationen/Speicherbarrieren **Optimierungen:** Compiler,

Laufzeitsystem und CPUs, Instruktionen werden umgeordnet/we-
goptimiert.

Java Memory Model
Atomicity
var myInt = new AtomicInteger(1); var myObj = new AtomicReference(null); myInt.updateAndGet(x -> x * x); var temp = myInt.getAndSet(2); // <i>.setAndGet(2)</i> var temp1 = myInt.incrementAndGet(); // <i>.getAndDecrement()</i> ;
Einzelnes Lesen / Schreiben ist atomar für: Primitive Datentypen, Obj-Referenzen, long und double nur mit <i>volatile</i> Keyword.
Unteilbarkeit ≠ Sichtbarkeit: Nach Write sieht anderer Thread vlt. noch alten Wert.
Visibility
Garantierte Sichtbarkeit zwischen Threads: Locks Release & Acquire, Volatile Variable, Thread/Task-Start und Join, Initialisierung von final Variablen.
Ordering
Program Order: 'as-if-serial', Sequentielles Verhalten jedes Threads bleibt erhalten. (Andere Threads dürfen es anders sehen) Synchronization Order (Total Order): Synchronisationsbefehle werden zueinander nie umgeordnet. Happens-Before Relation (Partial Order): Alles andere kann umgeordnet werden, ausser es gibt garantierte Sichtbarkeit unter den Threads.
Java Synchronization Order

// Keine Umordnung in Java, weil alles volatile volatile boolean a = false , b = false ; a = true ; while (b) { } // <i>Thread 1</i> b = true ; while (a) { } // <i>Thread 2</i> // Nicht korrekt, da nicht atomar private volatile boolean locked = false ; public void acquire() { while (locked) { } locked = true ; } // Spin-Lock mit atomarer Operation private AtomicBoolean locked = new AtomicBoolean(false) ; public void acquire() { while (locked.getAndSet(true)) { } } public void release() { locked.set(false); } public void acquire() { boolean success = false ; do { boolean old = locked.get(); if (old != true) { success = locked.compareAndSet(old, true); } } while (!success); } public void release() { do { boolean old = locked.get(); while (locked.compareAndSet(old, false)); } // <i>Atomares Compare and Set, setzt Update falls Wert gleich expect</i> boolean compareAndSet(boolean expect, boolean update); // <i>Optimistische Synchronizaton</i> do { oldValue = var.get(); newValue = calcChange(oldValue); } while (!var.compareAndSet(oldValue, newValue)); }
.NET Memory Model

Unterschied zu Java: Atomicity: long/double auch ohne volatile atomar. Visibility: Nicht definiert, implizit durch Ordering. Ordering: nur Half und Full Fences.
Atomare Instruktionen: *Interlocked* Klasse
Volatile Half Fences
Volatile Write: Vorangehende Zugriffe bleiben davor. (Release Semantik) **Volatile Read:** Nachfolgende Zugriffe bleiben danach. (Acquire Semantik)

Full Fence: Memory Barrier
Thread.MemoryBarrier(); // <i>Verbietet Umordnung in beide Richtungen</i>
Actor Model
Async: nicht Blockierend. Substantiell anderes Programmierkonzept. Aktive Objekte: haben nebenläufiges Innenleben Kommunikation: Objekte senden und empfangen Nachrichten. Kein Shared Memory: Nur Austausch von Nachrichten. Vorteile: Actor, CSP
Inhäerente Nebenläufigkeit: Alle Obj (Actors) laufen nebenläufig. Maschine kann Grad an Nebenläufigkeit ausnutzen. Keine Race Conditions: Kein Shared Memory. Nachrichtenaustausch synchronisiert implizit. Gute Verteilbarkeit: Kein Shared Memory. Nachrichtenaustausch für Netz prädestiniert.
Akka
class ParFrogActor extends UntypedActor { private final int number; private ActorRef next; public ParFrogActor(int number) { this .number = number; } public void onReceive(Object message) { if (message instanceof ActorRef) { next = (ActorRef)message; if (number == 0) { next.tell(number, getSelf()); } } else if (message instanceof Integer) { int value = (int)message; if (number == 0) { syso("res: " + value); } else { next.tell(value + number, getSelf()); } } } }

// Hauptprogramm var system = ActorSystem.create("System"); var actors = new ActorRef[SIZE]; for (int i = 0; i < SIZE; i++) { actors[i] = system.actorOf(Props.create(ParFrogActor. class , i)); } for (int i = 0; i < SIZE; i++) { var next = actors[(i + 1) % SIZE]; actors[i].tell(next, ActorRef.noSender()); } GPU Parallelisierung 512, 1024, 3584, 5760 Cores. Sehr spezifische langsamere Prozessoren. GPU Aufbau SP (Streaming Processor): 8-192 SPs pro SM SM (Streaming Multiprocessor): z.B. 1-30 SM SIMD: SM ist prinzipiell SIMD (Single Instruction Multiple Data), Vektorparallelität Alle Cores führen gleiche Instruktion aus, Einzelne können sie auch nicht ausführen. GPU vs. CPU GPU: hohe Datenparallelität, wenig Verzweigungen, Kein beliebiges Warten, Kleine Caches pro Core. Ziel: Hoher Gesamtdurchsatz GPU: Gegenteil, Ziel: Niedrige Latenz pro Thread NUMA Modell Non-Uniform-Memory-Access: Kein gemeinsamer Hauptspeicher zwischen GPU und CPU. Explizites Übertragen.
CUDA - Computer Unified Device Architecture

CUDA Blocks: Threads sind in Blöcke gruppiert. Blöcke sind im gleichen SM. Threads können innerhalb Block interagieren. Ausführungsmodell: Thread: virt. Skalarprozessor. Block: virt. Multiprozessor. Blöcke müssen unabhängig sein. Run To Completion. Beliebige Ausführungsreihenfolge. Grad der Parallelität durch GPU bestimmt. Automatische Skalierung. Ablauf: 1. Auf GPU allokieren <i>cudaMalloc</i> , 2. Daten auf GPU transferieren <i>cudaMemcpy</i> , 3. Kernel ausführen, 4. Rücktransfer, 5. Auf GPU deallozieren
Datenaufteilung
threadIdx.x: ThreadId im Block. blockIdx.x: Nummer des Blocks. blockDim.x: Blockgrösse. Programmierende modellieren Datenaufteilung selber.
Boundary Check
Falls Mehr Threads als zu bearbeitende Daten. Threads mit $i \geq N$ dürfen nicht auf Daten zugreifen.
Unified Memory
Automatischer Transfer CPU - GPU.
Speicherstufen

Shared Memory: Per SM. Schnell ca 4 Zyklen. Nur zwischen Threads innerhalb Block sichtbar. Paar KB. *shared float x*; **Global Memory:** Main Memory in GPU. Langsam, ca. 400-600 Zyklen. Allen Threads sichtbar. Mehrere GB. *cudaMalloc()*

Block Barriere
<i>_syncThreads()</i> ;
In if/else nur falls für alle Threads eines Blocks.
Warp
Block wird intern in Warps zerlegt (je 32 Threads). Ausführung: SIMD. SM kann alle Warps eines Blocks beherbergen. Nur wenige laufen echt parallel (1 bis 24).
Divergenz
Unterschiedliche Verzweigungen im selben Warp. <i>if/switch/while-do/for</i> . SM führt Instruktionen der einen Verzweigung durch, dann Instruktionen einer anderen Verzweigung. Performance Problem Schlechter Fall: Divergenz innerhalb derselben Warp. Guter Fall: Gleiche Verzweigung im Warp.
Memory Coalescing

Zugriffsmuster für Performance optimieren. **Burst:** Falls Threads auf 32-Byte-Bereiche zugreifen. Sonst teure Einzel-Zugriffe. (Je 400 Zyklen pro Global Memory)
Tiles einlesen, Multi-Dim Spalte auf x-Achse
Shared Memory Optimierung
Falls Threads auf mehrere Elemente zugreifen, im SM Cachen.

Example
1-Dimensional
global void pairwise_sum(int* array, int length) { int i = blockIdx.x * blockDim.x + threadIdx.x; int offset = 2 * i; if (offset + 1 < length) { array[offset] = array[offset + 1]; array[offset + 1] = 0; } } // Aufruf int blockSize = 1024; int gridSize = (len / 2 + blockSize - 1) / blockSize; pairwise_sum<<gridSize, blockSize>>>(d_a, len); 2-Dimensional
global void transpose(int* matrix, int rows, int cols, int* result) { int row = blockIdx.x * blockDim.x + threadIdx.x; int col = blockIdx.y * blockDim.y + threadIdx.y; if (row < rows && col < cols) { result[col * rows + row] = matrix[row * cols + col]; } } // Aufruf dim3 gridSize(32,32); dim3 gridSize((N + 31) / 32, (M + 31) / 32); transpose<<gridSize, blockSize>>>(matrix, N, M, result);

Cluster Parallelisierung
Synchron: Send & Recv sind blockierend. <i>mpicxx -n 100 programm.exe</i> Motivation: möglichst hohe Beschleunigung. Viele CPU Cores statt nur viele GPU Cores. Verteiltes Programmiermodell: Programm auf mehreren Nodes ausführen. Kein Shared Memory (NUMA) zwischen nodes, nur für Cores im Node (SMP). Message Passing Interface (MPI): Basiert auf Actor/CSP Prinzip.
MPI_Init (&argc, &argv); // <i>MPI Initialisierung</i> MPI_Comm_rank (MPI_COMM_WORLD, &rank); // <i>Prozess Identifikation</i> MPI_Finalize (); // <i>MPI Finalisierung</i> MPI_Send (&val, 1, MPI_INT, recRank, tag, MPI_COMM_WORLD); MPI_Recv (&val, 1, MPI_INT, senderRank, tag, MPI_COMM_WORLD, MPI_STATUS_IGNORE); MPI_Barrier (MPI_COMM_WORLD); // <i>Wartet auf alle Prozesse</i> MPI_Allreduce (&val, &total, 1, MPI_INT, MPI_SUM, MPI_COMM_WORLD); // <i>Aggregation von Teilergebnaten</i> MPI_Reduce (&val, &total, 1, MPI_INT, MPI_SUM, recRank, MPI_COMM_WORLD); // <i>Effizienter, kein Broadcast</i>
SPMD: Single Program Multiple Data. MPI Programm wird in mehrere Prozesse gestart. Prozesse können untereinander kommunizieren. Communicator: Gruppe von MPI-Prozessen. Communicator World: Alle Prozesse einer MPI-Programmausführung.
// maximum of array int main(int argc, char* argv[]) { MPI_Init(&argc, &argv); int rank, size; MPI_Comm_rank(comm, &rank); MPI_Comm_size(comm, &size); int* array = NULL ; int offset, length = 0; if (rank == 0) { array = read_array_file(&length); int partition = length / size; for (int to = 1; to < size; to++) { offset = rank * partition; if (rank == size - 1) { partition = length - offset; } MPI_Send(&partition, 1, MPI_INT, to, 0, MPI_COMM_WORLD); MPI_Send(array+offset, partition MPI_INT, to, 0, MPI_COMM_WORLD); } if (size > 1) { length = partition; } } else { MPI_Recv(&length, 1, MPI_INT, 0, 0, MPI_COMM_WORLD); MPI_Recv(array, length, MPI_INT, 0, 0, MPI_COMM_WORLD); } // <i>calculate maximum</i> int result; MPI_Reduce(&maximum, &result, 1, MPI_INT, MPI_MAX, 0, MPI_COMM_WORLD); print ("maximum: %i", result); MPI_Finalize(); }
Concurrency in Python

GIL: Global Interpreter Lock. Nur ein Thread kann Python Byte-Code ausführen. Kein Speedup für CPU-Bount Operationen möglich. Data Races durch Reordering dennoch möglich, Visibility nicht garantiert. Kein definiertes Memory Model.

from threading import Thread from multiprocessing import Process if __name__ == '__main__': t = Thread(# p = Process target=fibonacci, args=(10,)) t.start(); t.join(); # Shared Memory res = Value('i', -1, lock=False); // Typ, InitialWert, lock
Pools
Threads: <i>concurrent.futures.thread.ThreadPoolExecutor</i> Prozesse: <i>concurrent.futures.process.ProcessPoolExecutor</i> / <i>multiprocessing.Pool</i> if __name__ == '__main__': with ProcessPoolExecutor() as pool: future = pool.submit(fib, 10) # or: map(fib, [1,2,3]) print (future.res()); # Blockiert bis Task-Ende
Asynchrone Programmierung (asyncio)

Coroutine-Functions: werden erst beim <i>await</i> ausgeführt. Keine parallele Ausführung. Ausnahme: Coroutine wird als Task verpackt.
async def sub_routine(n): await asyncio.sleep(n); if __name__ == '__main__': asyncio.run(sub_routine(1))
JavaScript Concurrency
Grundsätzlich Single-Threaded mit einem Event-Loop. Kein Schutz vor Race-Condition.
function delay(ms) { let promise = new Promise((resolve, reject) => { setTimeout(() => resolve(), ms); }) } async function countTo(n) { for (let i = 1; i <= n; i++) { await delay(1000); } } }
Web-Worker
Entsprechen einem Thread im Browser. Datenaustausch primär über Messaging. Werden mit Quelldatei gestartet. Langiebiger Prozess mit eigenem Event-Loop.
onmessage = event => { // <i>Worker definieren</i> const n = event.data; const res = fib(n); postMessage(res); } const worker = new Worker('w.js'); // <i>Worker verwenden</i> worker.onmessage = event => { console.log(event.data); } worker.postMessage(42);