Non-blocking I/O and the C10K Problem

Efficient Networking and Threads

Effective Software — Lecture 9

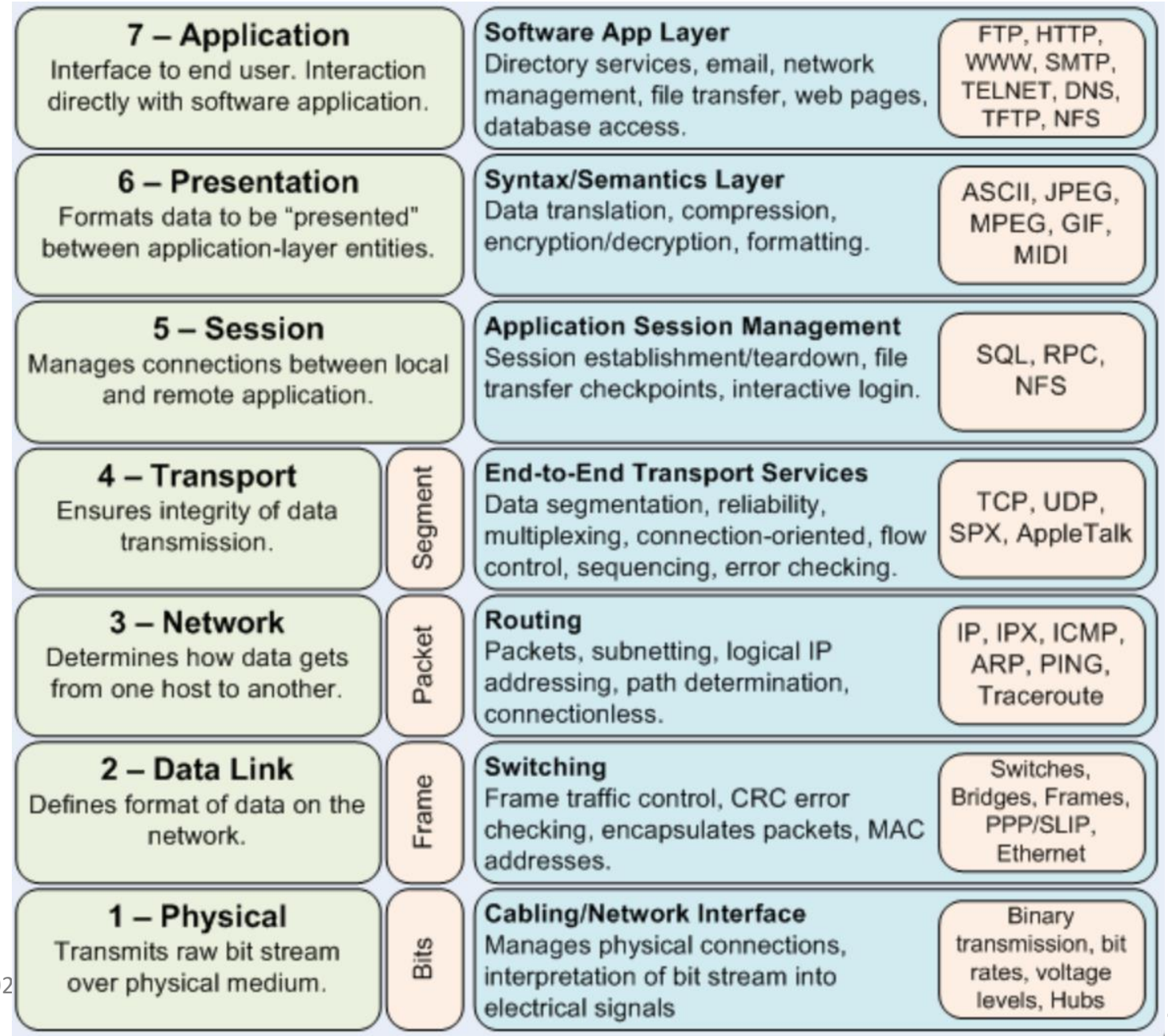
David Šišlák
david.sislak@fel.cvut.cz

[1] Tanenbaum, A. S., Wetherall, D. J.: Computer Networks. Pearson, 2011.
[2] Kegel, D.: The C10K problem. http://www.kegel.com/c10k.html
[3] Hitchens, R.: Java NIO. O'Reilly, 2002.
[4] Pressler, R., Bateman, A.: JEP 436 - Virtual Threads (second preview)

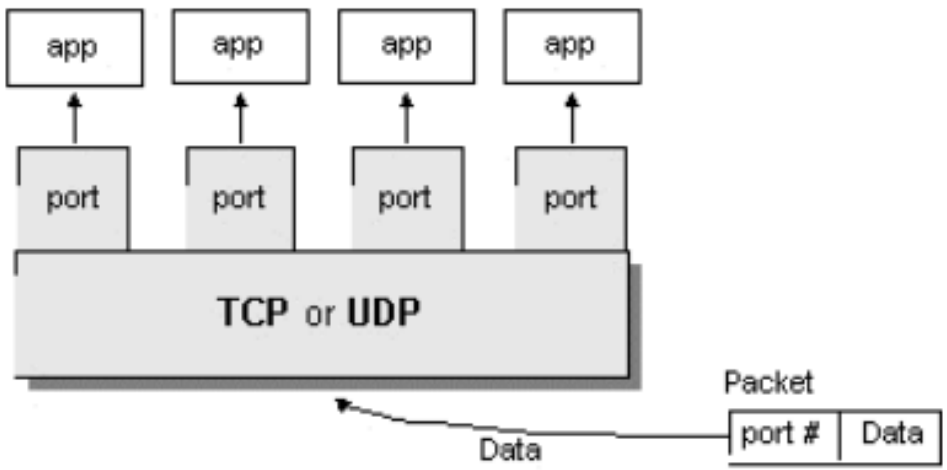
Outline

- Network communication
  - OSI model
- C10K problem
  - Thread-per-request vs. event-based approach
- Non-blocking I/O
  - select()
  - poll()
  - epoll
  - Java Non-blocking I/O
    - Native memory buffer
    - NIO
- Threads
  - Thread pools
  - Virtual threads

Network Communication – OSI Model Overview



Network Communication – Introduction



TCP		UDP	
FTP	20,21	DNS	53
SSH	22	BooTPS/DHCP	67
Telnet	23	TFTP	69
SMTP	25	SNMP	161
DNS	53		
HTTP	80		
POP3	110		
NTP	123		
IMAP4	143		
HTTPS	443		

» **Ports** – 16-bit numbers

» **IPv4** – 32-bit address

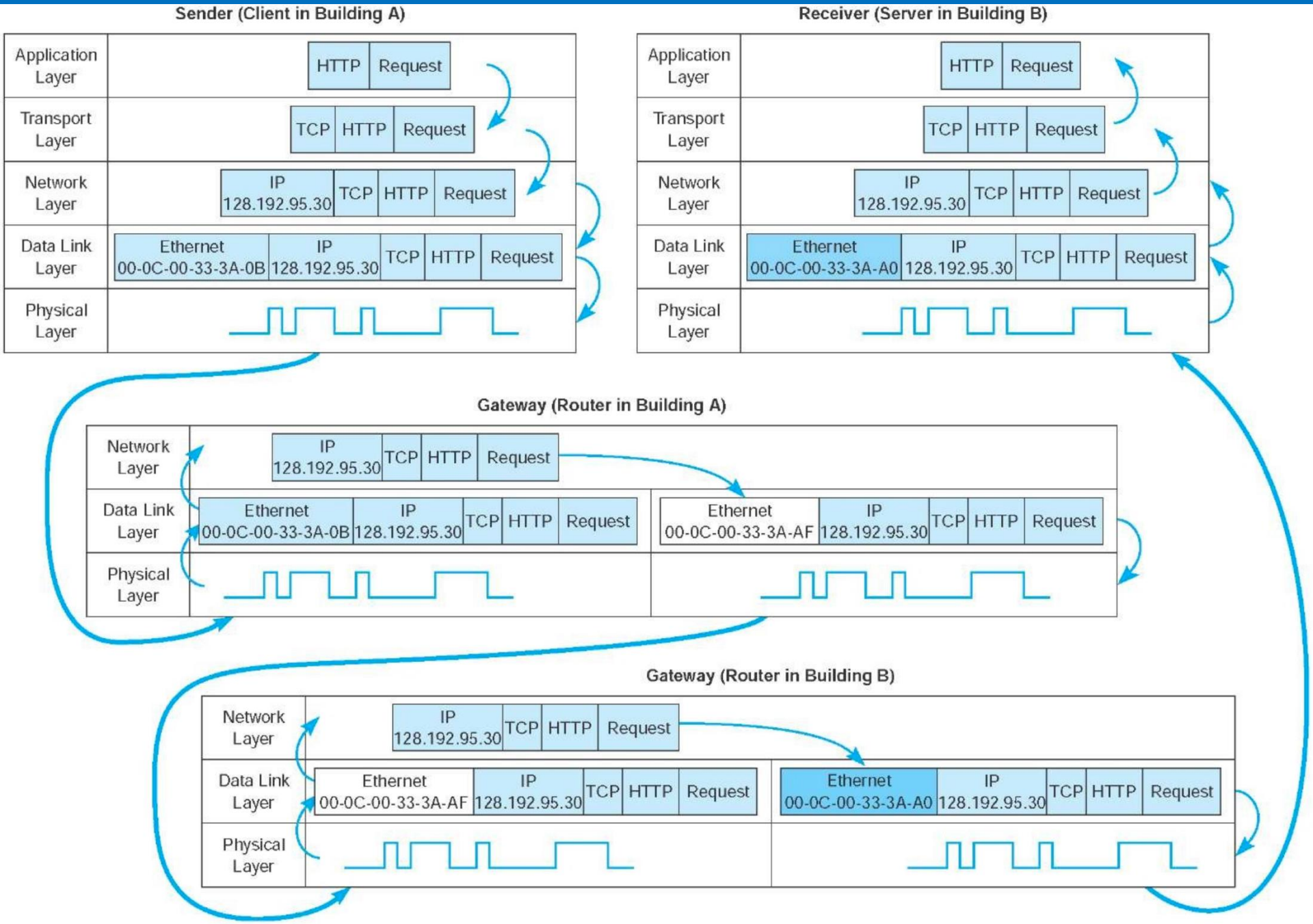
» **IPv6** – 128-bit address

- 48-bit (or longer) routing prefix, up to 16-bit subnet ID, 64-bit interface

http://[1fff:0:a88:85a3::ac1f]:8080/index.html

» TCP/UDP connection identification – **4-tuple** – src IP, src port, dst IP, dst port

Network Communication – HTTP Example – Encapsulation



C10K Problem

» Handling a large number of clients (10,000+) at the same time (late 90s)
  - **a large number of concurrent connections handled by a single server,** requiring efficient scheduling
  - not related to requests per second
» sometimes known as C1M or C10M problem (nowadays)
» approach
  - **thread-per-request servers** (*Apache*)
    - each connection is handled by its **own thread**/process (pooled but limited)
    - connection operations usually use **blocking** operations
    - thread scheduling does not scale (high cost of context switching)
    - thread scheduling used as packet scheduling
  - **event-driven I/O servers** (*Nginx*)
    - you handle packet scheduling yourself – **single/multi-threaded event loop**
    - using **non-blocking** (asynchronous) operations with **event interceptors**
    - **multi-core scalability** with controlled number of worker threads
    - reuse thread-based data structures, avoid locks (atomics, non-blocking)

Non-Blocking I/O Approach

- **interrupts**
  - hardware interrupts in kernel mode
- **polling**
  - looping to regularly check status (readiness for I/O)
  - wastes CPU cycles
- **signals**
  - OS generated signals on I/O readiness
  - may leave the process in an inconsistent state
- **callbacks**
  - pointer to handler function
  - stack growth issue (callbacks issuing nested I/O operations)
- **event-based**
  - select()
  - poll()
  - epoll

Event-Based I/O - select

» **select**

- defined in POSIX (Portable Operating System Interface)
- originally used for blocking I/O
- the passed **lists of *descriptors* cannot be reused** in subsequent calls as they are modified by the system call
- **not scalable** – limited *descriptors* + must be iterated over to find events

```
int
select(int nfds, fd_set *restrict readfds, fd_set *restrict writefds, fd_set *restrict errorfds,
struct timeval *restrict timeout);

void
FD_CLR(fd, fd_set *fdset);

void
FD_COPY(fd_set *fdset_orig, fd_set *fdset_copy);

int
FD_ISSET(fd, fd_set *fdset);

void
FD_SET(fd, fd_set *fdset);

void
FD_ZERO(fd_set *fdset);
```

Event-Based I/O - poll

» **poll**

- the number of polled descriptors is not limited
- descriptors can be reused
- better but you still **need to iterate over descriptors** to find events

```
int
poll(struct pollfd fds[], nfds_t nfds, int timeout);
```

```
struct pollfd {
    int    fd;       /* file descriptor */
    short  events;   /* events to look for */
    short  revents;  /* events returned */
};
```

Event-Based I/O - epoll

- **epoll**
  - Linux only (e.g. Windows has IOCP – I/O Completion Ports)
  - **scalable**
  - monitored events can be modified while polling (via syscall)
  - returns only triggered events
- **API**
  - epoll_create & epoll_create1 – initialize an epoll instance (kernel structure)
  - epoll_ctl – add/modify/remove descriptors to epoll instance
  - epoll_wait – waits for events up to a timeout
- **modes**
  - **level triggered** – wait returns immediately if an event is available
  - **edge triggered** (EPOLLET) – readiness returned upon incoming event only
    (you have to process all pending events before next wait !)
- **events**
  - EPOLLIN, EPOLLOUT, EPOLLPRI
  - EPOLLRDHUP, EPOLLHUP
  - EPOLLERR

Epoll Usage

Epoll structure:

```
typedef union epoll_data
{
  void        *ptr;
  int          fd;
  __uint32_t   u32;
  __uint64_t   u64;
} epoll_data_t;

struct epoll_event
{
  __uint32_t   events; /* Epoll events */
  epoll_data_t data;   /* User data variable */
};
```

Initialization:

```
int epfd = epoll_create1(0);
...
struct epoll_event ev;
int client_sock;
...
ev.events = EPOLLIN | EPOLLPRI | EPOLLERR | EPOLLHUP;
ev.data.fd = client_sock;
int res = epoll_ctl(epfd, EPOLL_CTL_ADD, client_sock, &ev);
```

Epoll Event Loop

```
while (1) {
    // wait for something to do...
    int nfds = epoll_wait(epfd, events,
                                MAX_EPOLL_EVENTS_PER_RUN,
                                EPOLL_RUN_TIMEOUT);
    if (nfds < 0) die("Error in epoll_wait!");

    // for each ready socket
    for(int i = 0; i < nfds; i++) {
      int fd = events[i].data.fd;
      handle_io_on_socket(fd);
    }
}
```

Java Blocking Networking – TCP Client

» **Socket**

- client endpoint of a TCP/IP network connection
- is bound to a specific destination IP and port
- each TCP/IP connection is uniquely identified by its two endpoints
- provides input/output streams

```
try (
    Socket echoSocket = new Socket( host: "localhost",  port: 7);
    PrintWriter out = new PrintWriter(echoSocket.getOutputStream(),  autoFlush: true);
    BufferedReader in = new BufferedReader(new InputStreamReader(echoSocket.getInputStream()));
    BufferedReader stdIn = new BufferedReader(new InputStreamReader(System.in))
) {

    String userInput;

    while ((userInput = stdIn.readLine()) != null) {
        out.println(userInput);
        System.out.println("echo: " + in.readLine());
    }
}
```

Java Blocking Networking – TCP Server

» **ServerSocket**

- server socket representing a listening TCP/IP endpoint
- in the constructor, you specify the port, and optionally IP where it should be bound
- waits for a connection to be established using the method Socket **accept**()

Java Blocking Networking – TCP Server - Example

thread-per-request server example – each handler runs in its own thread with blocking I/O

```
ExecutorService clientRunner = Executors.newCachedThreadPool();
try (
        ServerSocket serverSocket = new ServerSocket( port: 7)
) {
    while (true) {
        final Socket s = serverSocket.accept();
        clientRunner.execute(() -> {
            try (
                    BufferedReader in = new BufferedReader(new InputStreamReader(s.getInputStream()));
                    PrintWriter out = new PrintWriter(s.getOutputStream(), autoFlush: true)
            ) {
                String line;
                while (s.isConnected()) {
                    if ((line = in.readLine()) != null) {
                        out.println(line);
                    }
                }
            } catch (IOException e) {
                e.printStackTrace();
            }
        });
    }
} catch (Exception e) {
    e.printStackTrace();
} finally {
    clientRunner.shutdownNow();
}
```

Java Blocking Networking - UDP

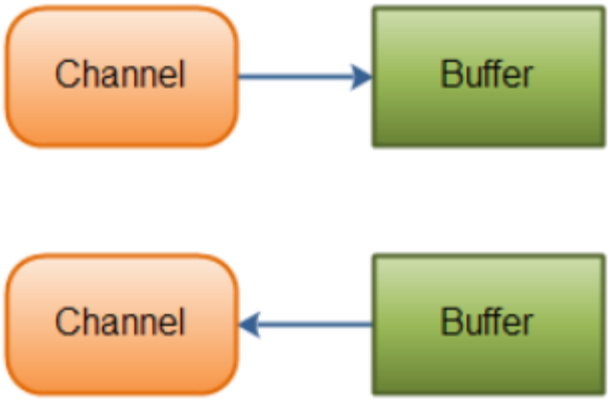
- **DatagramPacket**
  - independent, self-contained message sent over the network
  - similar to a network **packet**
    - InetAddress address, int port – destination
    - byte data[], int length, int offset
    - SocketAddress sa – sender
- **DatagramSocket**
  - endpoint for **sending or receiving packets**
  - can be bound to any available port (using default constructor)
  - connect(InetAddress,int) – can send or receive packets only from/to a specified host, if not set in the DatagramPacket, it is filled automatically
  - send(DatagramPacket p), receive(DatagramPacket p) – blocking I/O
- **MulticastSocket**
  - additional capabilities for joining/leaving multicast groups, loopback
  - multicast IP (IGMP – Internet Group Management Protocol)

    224.0.0.0 – 239.255.255.255

Java Non-blocking Networking - NIO

- **scalable I/O** – asynchronous I/O requests and polling
- high-speed **block-oriented** binary and character I/O – including mapping files to the memory, using channels and selectors

- A channel is a block-oriented abstraction working with buffers

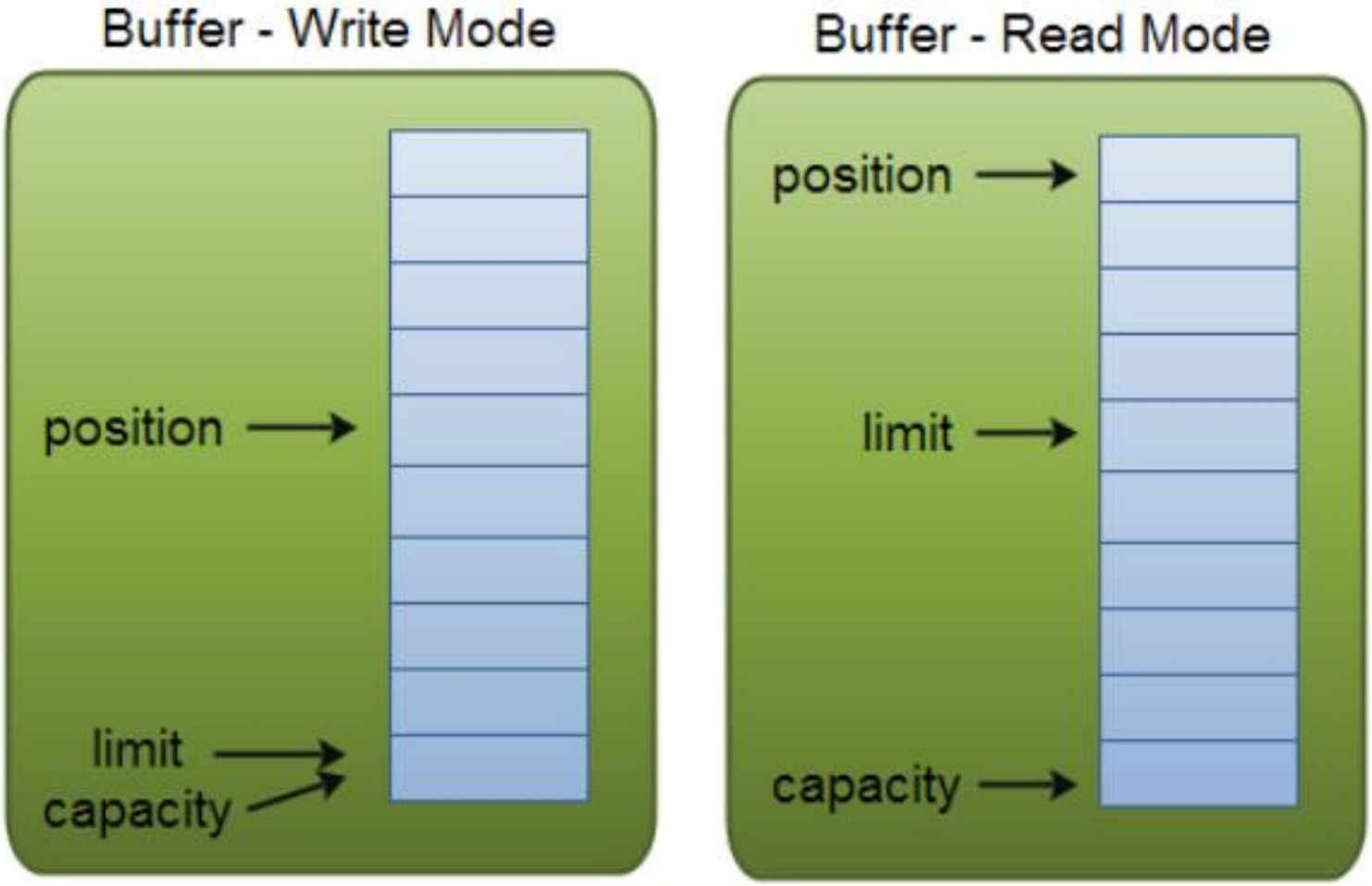


Java NIO - Buffer

» **java.nio.Buffer**

- **linear, finite sequence of elements** of a specific primitive type
  - ByteBuffer, CharBuffer, DoubleBuffer, FloatBuffer, IntBuffer, LongBuffer, ShortBuffer, MappedByteBuffer {FileChannel.map(…)}
- not thread safe, **multiple modes** for the same buffer (both read & write)
- **key properties** – 0 <= mark <= position <= limit <= capacity
  - capacity – number of elements, never changing !
  - limit – index of the first element that should not be read or written
  - position – index of the next element to be read or written
  - mark – index to which its position is set after reset()
- clear() – position=0, limit=capacity => **ready for channel read** (put)
- flip() – limit=position, position=0 => **ready for channel write** (get)
- rewind() – limit unchanged, position=0 => ready for re-reading
- mark() – mark = position
- reset() – position=mark

Java NIO - Buffer



- write mode – channel.read(buf); buf.put(…);
- read mode – channel.write(buf); … buf.get();

Java NIO - Buffer

» java.nio.Buffer

- isReadOnly() – can be read-only
- hasArray() – is backed by an accessible array (array())
- equals(), compareTo() – compare the remaining sequence

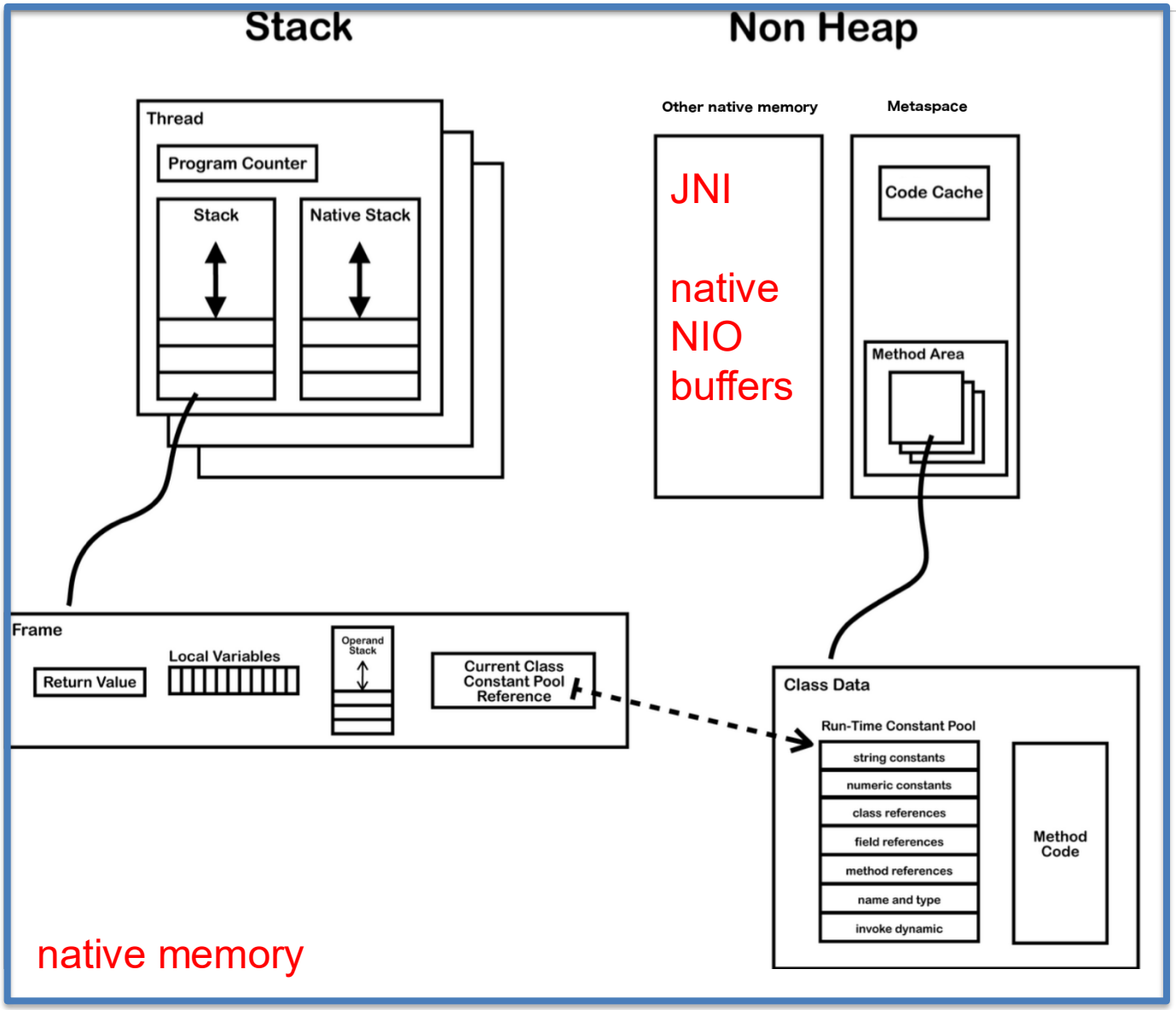
- can be **allocated to native memory** (see next slide)

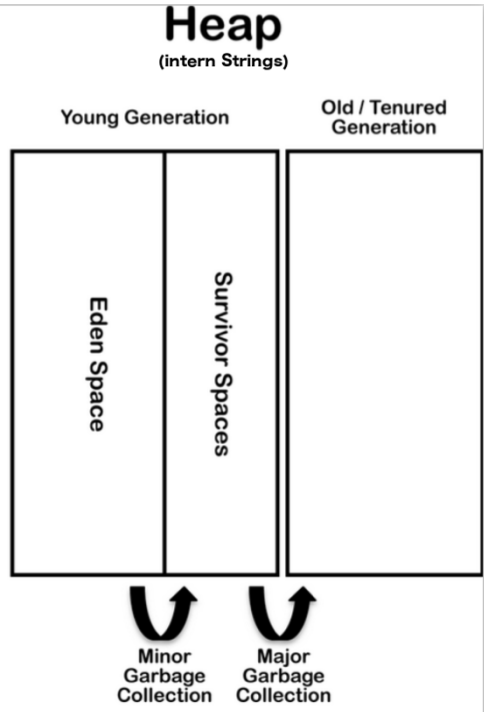
- **typical usage**

1. Write data into the Buffer
2. Call `buffer.flip()`
3. Read data out of the Buffer
4. Call `buffer.clear()` or `buffer.compact()`

Note: compact() – bytes between position and limit are copied to the beginning of the buffer and prepares the buffer for writing again

JVM – Memory Layout – Native Memory



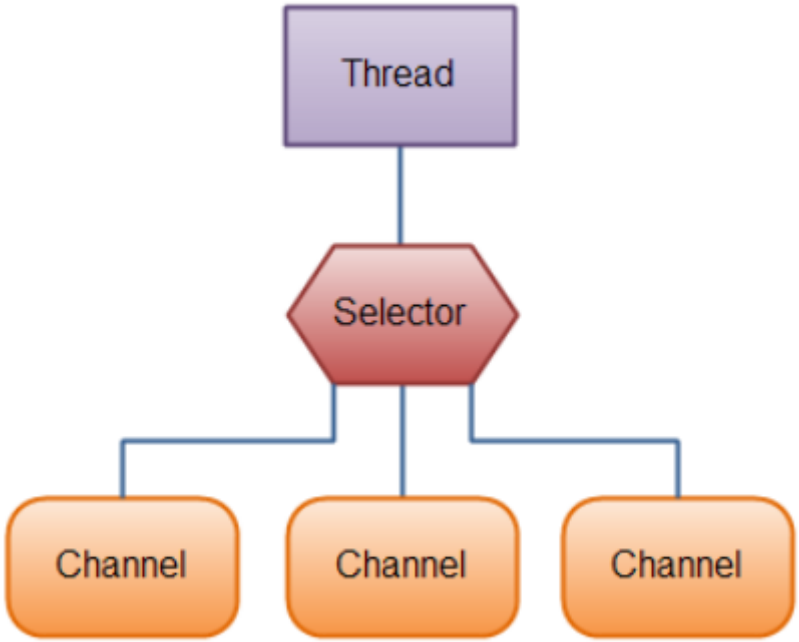


JVM NIO - Direct Buffers

- **ByteBuffer.allocateDirect(…)**
- stored out of Java heap in **native memory**
- allows native and Java code to **share data without copying**
  - useful for files and sockets
    - the same memory is passed to kernel during calls
- multiple buffers can share native memory
  - slice()/duplicate() – independent position, limit, mark, shared content
  - asReadOnlyBuffer() – read only view of shared content
- tuning/tracking
  - -XX:MaxDirectMemorySize=N (default: unlimited)
  - -XX:NativeMemoryTracking=off|summary|detail
  - -XX:+PrintNMTStatistics

Note: usage of heap buffers implies content copy out/in Java heap space due to possible relocations by GC

Java NIO - Channel, Selector



- **one thread** works with **multiple channels at the same time**
  - **epoll-based** if the OS supports epoll
- **Channels** cover UDP/TCP network I/O and file I/O
  - FileChannel – from Input/OutputStream or RandomAccessFile
  - DatagramChannel
  - MulticastChannel
  - SocketChannel
  - ServerSocketChannel

Java - NIO Channel

» **Channel**

- supports simultaneous read/write operations (streams are only one-way)
- always read/write from/to a **buffer**

» **FileChannel**

- supports only **blocking mode**
- support – direct buffers, mapped files, locking
- bulk transfers between channels
  - no copy at all, direct transfer e.g. to socket
  - **transferFrom**(sourceChannel, int pos, int count)
  - **transferTo**(int pos, int count, dstChannel)

Java - NIO Channel

- **SocketChannel** – client endpoint of TCP/IP
  - can be configured as **non-blocking** before connecting
  - SocketChannel sc = socket.getChannel();
  - SocketChannel sc2 = SocketChannel.open();
  - sch.connect(…)
  - write(…) and read(…) may return without having written/read anything for non-blocking channel
- **ServerSocketChannel** – server endpoint of TCP/IP
  - can be configured as **non-blocking**
  - can be created directly using open() or from ServerSocket
  - accept() – returns SocketChannel in the same mode

Java - NIO Selector

» **Selector**

- Selector selector = Selector.open();
- only channels in **non-blocking** mode can be registered

  channel.configureBlocking(false);

  SelectionKey channel.register(selector, SelectionKey.OP_READ);
- FileChannel doesn't support non-blocking mode

» **SelectionKey** – events you can listen for (multiple can be combined)

- OP_CONNECT
- OP_ACCEPT
- OP_READ
- OP_WRITE

» events are set by channels that are ready for a given operation

Java - NIO Selector

» **SelectionKey** – returned from register method
  - interest set – your configured ops
  - ready set – which ops are ready, sk.isReadable(), sk.isWritable(), …
  - channel
  - selector
  - optional attached object – sk.attach(Object obj);
    Object sk.attachment()

SelectionKey channel.register(selector, ops, attachmentObj);

Java - NIO Selector

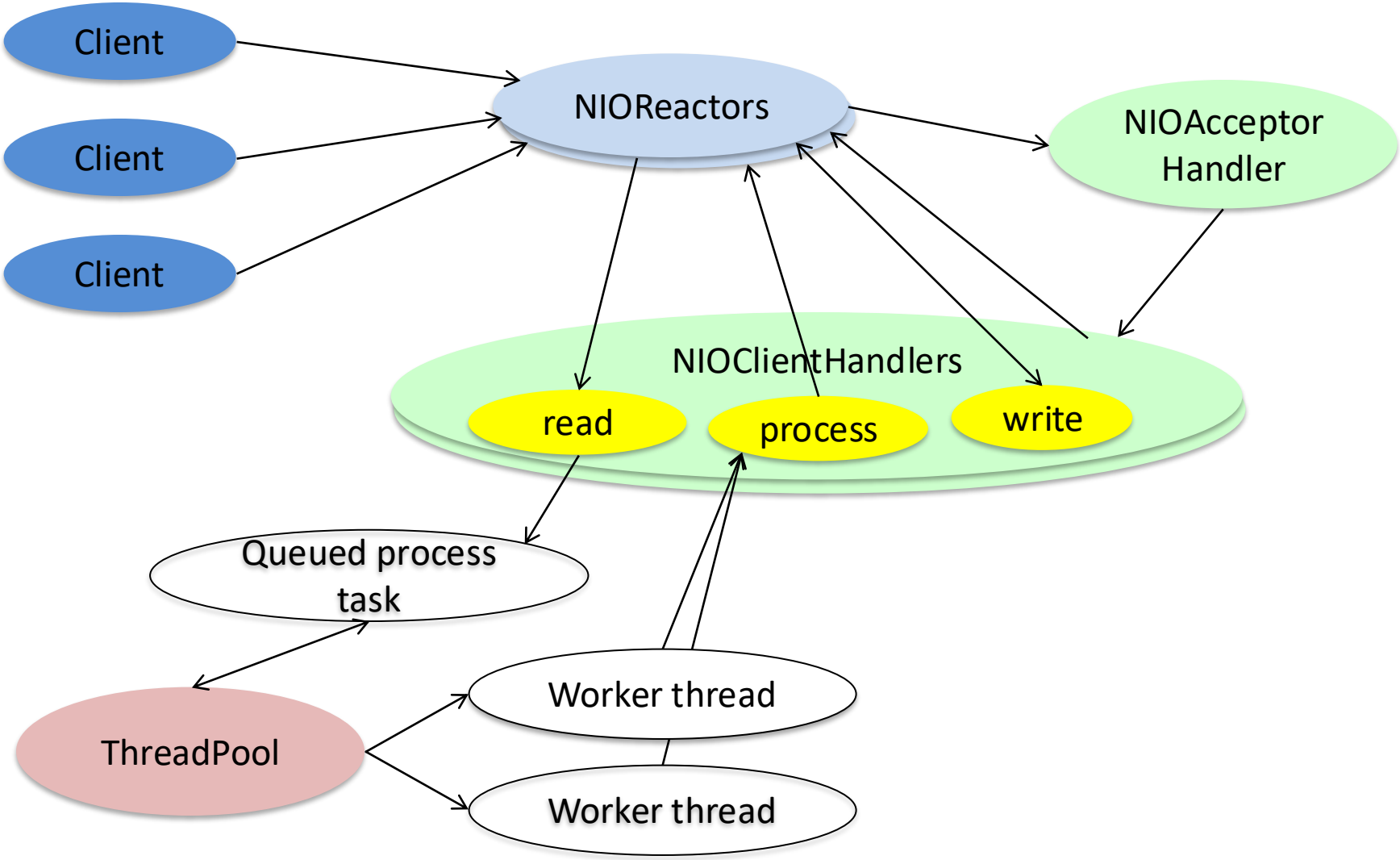
» Selector with registered one or more channels
- int **select**() – blocks until at least one channel is ready
- int select(long timeout) – with timeout milliseconds
- int selectNow() – doesn't block at all, returns immediately

returns the number of channels that are ready since the last call

Set<SelectionKey> selector.selectedKeys();

- **wakeUp**() – different thread can "wake up" thread blocked in select()
- **close**() – invalidates selector, channels are not closed

Java NIO Server – Using Multiple Reactors



Threads

- **processes** vs. **threads**
  - both support concurrent execution
  - a process has one or more threads
  - threads share the same address space (data and code)
  - local variables, exception handling, debugging and profiling
  - context switching between threads is usually less expensive
  - thread inter-communication is relatively efficient using shared memory
- **JVM**
  - a thread executes sequence of code with own stack with frames

    t.getStackTrace()
  - own local variables
  - own method parameters
- thread creation by
  - subclass of **java.lang.Thread**
  - implementation of **java.lang.Runnable**

Java Thread Pool - ExecutorService

- concept of **thread pooling**
- suitable for execution of large number of asynchronous tasks
  - e.g., processing of requests in server
- **reduces the overhead of thread creation for each task, context switching**
- interface - java.util.concurrent.**ExecutorService**
  - shutdown(), shutdownNow(), awaitTermination
  - **execute**(Runnable r)
  - Future<?> **submit**(Runnable r), Future<T> **submit**(Callable<T> c)
- java.util.concurrent.**Future**<T>
  - boolean cancel(boolean mayInterruptIfRunning)
  - isCancelled(), isDone()
  - V get(), V get(long timeout, TimeUnit unit)
- java.util.concurrent.Executors (optionally with ThreadFactory)
  - **newSingleThreadExecutor**()
  - **newFixedThreadPool**(nThreads)
  - **newCachedThreadPool**() – default 60 seconds keep-alive

Java Virtual Threads - Introduction

» **lightweight implementation of Java thread**

» available since Java 21

- preview feature since Java 19 (attribute --enable-preview)

» **standard thread**

- thin wrapper around OS-managed platform thread
- basic unit of OS scheduling
- creation and removal are expensive and resource-heavy operations
- fixed thread stack size -> StackOverflowException
- doesn't scale

» **alternatives**

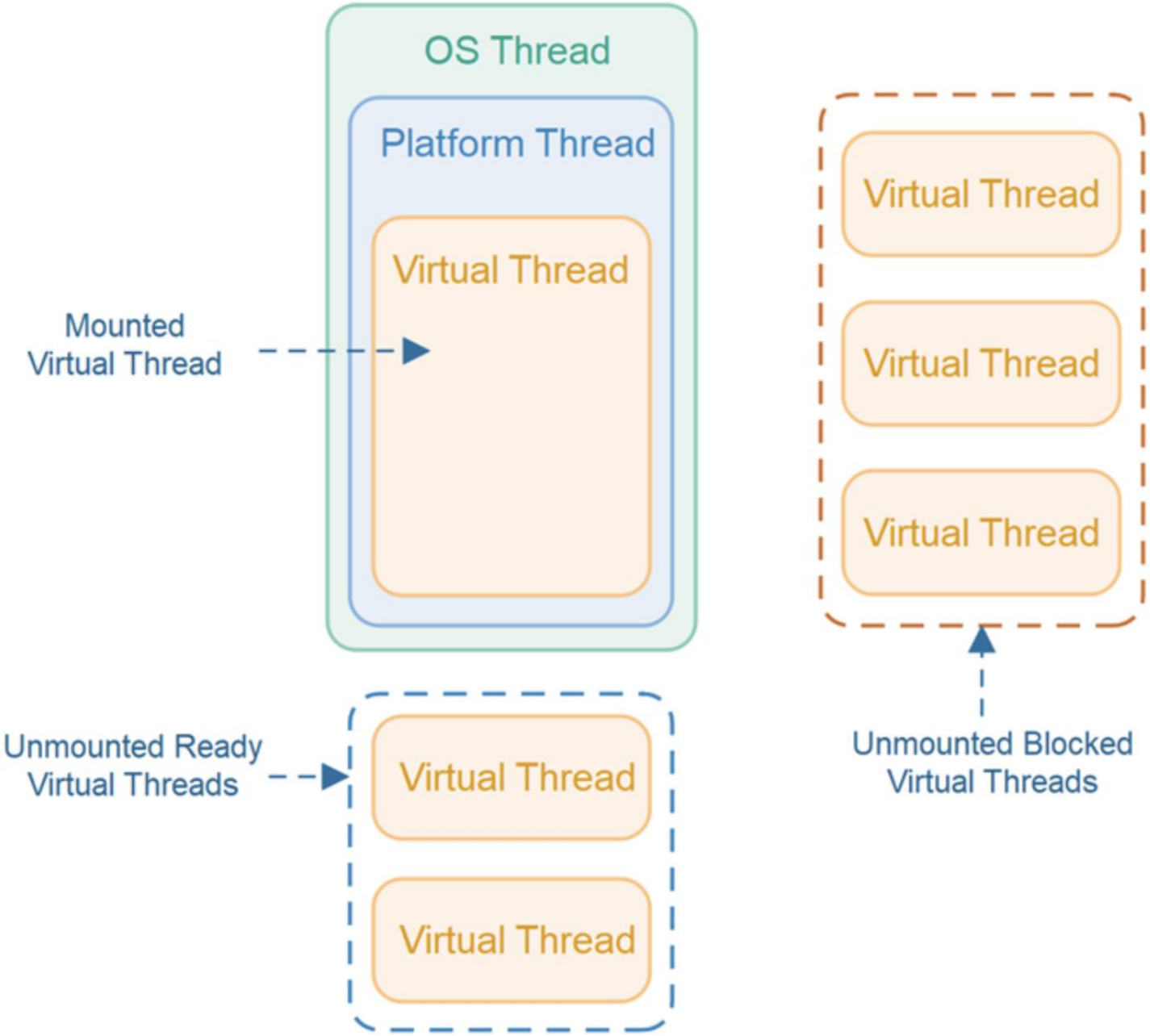
- async/await – reactive-style programming (e.g. Kotlin)
  - asynchronous operations with callback
- issues with readable stack-traces, debugging and observability
- complex workflow for sequential composition, iteration, try-catch blocks

Java Virtual Threads - Details

» **virtual thread**

- reduces the effort required to write high-throughput concurrent applications
- **thread-per-request** approach with almost optimal hardware utilization
- compatible with Thread API
- support debugging and profiling with existing tools
- stack frames in heap
  - stack size dynamically resizes as needed – expand and shrink
- OS still manages only platform threads
- **virtual thread** is **mounted** to **carrier thread** for execution
  - copies stack frames from the heap to the stack of the carrier thread
  - **unmounted** when blocked for I/O, lock or other resource
  - mounting/unmounting is invisible from Java code
  - thread dump, stack trace do not include carrier thread frames
  - carrier threads are from ForkJoinPool operating in FIFO mode
    - using the number of available logical CPU cores

Java Virtual Threads - Details



Java Virtual Threads - Details

- **virtual thread API**
  - Thread::ofVirtual()
  - implementation of the ordinary Thread
    - Thread::currentThread() returns virtual thread, not carrier thread
    - ThreadLocal, interruption, stack walking works the same way
    - always daemon thread, Thread::setDaemon has no effect
    - priority cannot be changed
  - Executors.newVirtualThreadPerTaskExecutor()
    - each task runs in its own virtual thread
- **scalability**
  - fast creation, small memory footprint
  - execution efficiency is the same as for platform threads
  - **scales well for I/O-bound workloads** (even for short-lived tasks)
    - simplified design with thread-per-request
    - suitable for server applications
  - no additional value for CPU-bound workloads

Java Virtual Threads - Details

» example with 100k virtual threads

```
try (var executor = Executors.newVirtualThreadPerTaskExecutor()) {
    IntStream.range(0, 100_000).forEach(i -> {
        executor.submit(() -> {
            Thread.sleep(Duration.ofSeconds(1));
            return i;
        });
    });
}
```

» after warm-up takes about 1.1 seconds

» with Executors.*newFixedThreadPool(1000)* it takes about 1000 seconds

» **drawbacks**

- synchronized pins virtual thread to its carrier -> **use ReentrantLock**
- execution of JNI pins as well
- releases the carrier thread only during blocking operations, **no preemption!**