TBD:

Reschedule: (BS-NACK, air/train - cancellation from service provider) – Not applicable as it can be treated as a new user request.

Dynamic physical resource (Addition/removal of different resource at different time, addition of car in train, changing of channel/carrier in BS) – Too early to decide

Dynamic time resource?

Cancellation/Yield request

Blocked physical/time resources – under maintenance

# Definitions

**Generic scheduler** – airline, railway, base station, os, travel and hotel, competitive exam.. etc.

**Use Case**: When we have limited resources and more requests for usage, in that case we would like to have effective utilization of resources, based on criteria like priority etc.

**What will this scheduler do?**

Input to scheduler will be N number of user requests and it will generate output of X selected user request and how much resources to allocate for each selected request.

**Physical resource** – airline/railway -> seats, BS -> BW, OS -> processor cores

1. Max physical resource - N
2. Physical resources list (1-N)
3. Number of divisions
4. Division list
   1. Start resource id
   2. End resource id
   3. Vulnerable: Any other division can utilize this resource
   4. Capable: Can get resource from other division
   5. Available list - List of available physical resources.
   6. Blocked list – List of blocked physical resources (due to allocation/maintenance/other reasons)

**Time resource**

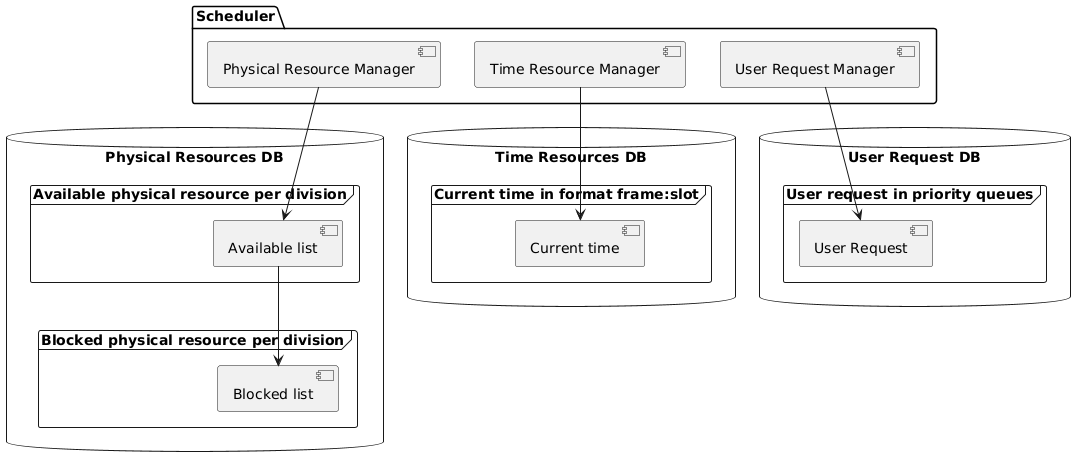
1. Time Slot – Smallest unit of time resource allocation possible
2. Frame – Group of time slots that repeat periodically
   1. Number of Time slots per frame
   2. Division list
      1. Start slot id
      2. End slot id
      3. Vulnerable
      4. Capable
3. Current time (Frame : Time slot)

**User Requests**

1. User request identifier
2. Request priority
3. Priority type (fix/flexi)
4. Priority modification function (mandatory in case of priority type flexi)
5. Physical resource Division Id
6. Number of physical resources requested
7. Start physical resource id?
8. Is physical division flexible
9. Time resource Division Id
10. Is time resource division flexible
11. Number of time slot requested
12. Start time slot?
13. Preemption vulnerability
14. Preemption capability
15. Delta time
16. Partial allocation of physical resource allowed
17. Partial allocation of time resource allowed
18. User request information.
19. Current priority (internal not from user)

# Components

**Component Diagram:**



1. Physical Resource DB
   1. APIs for data modification and retrieval
2. Physical Resource Manager
   1. Initialization/Configuration of Physical Resource
   2. Physical resource modification
3. Time Resource DB – TBD
   1. APIs for data modification and retrieval
4. Time Resource Manager - TBD
   1. Initialization/Configuration of Time Resource
   2. Time resource modification – TBD
   3. Maintaining current time resource
5. User Request DB
   1. APIs for data modification and retrieval
6. User Request Manager
   1. Selection of User Request
   2. Updating user request priorities

# Procedures

1. Initialization/Configuration of Physical Resources
2. Initialization/Configuration of Time Resource
3. Physical resource modification
   1. Addition of resource
   2. Deletion of resource
   3. Blocking of resource
      1. Due to resource allocation
      2. Due to maintenance
   4. Unblocking of resource
      1. The allocated time slot(s) for scheduled resource is over.
      2. Due to completion of maintenance.
      3. Due to cancellation
   5. Modification of Division in physical resource.
4. Check physical resource availability
5. Time resource modification (TBD)
6. Selection of User Request
7. Updating user request priorities
8. Deletion of user request queue.
9. Deletion from resource allocation queue

# Basic Design

**Request Queue** - The user request pending to be processed (OS – waiting for a resource, Train – waiting list)

**Invocation** – Hybrid (Periodic/event)

**Priority handling** – Pick from user request queue and put into resource allocation queue

Invocation Periodic: Check priority queues and schedule

Invocation Event: Check if resource available then schedule, Check for lowest priority

Preemptable user request, if found preempt it and schedule, else put it in wait

**Resource Allocation Queue** - The user requests selected for resource allocation.

**Resource allocation**

1. Start allocation of resources priority wise
2. If for a given request the resource cannot be allocated, check if partial allocation is possible.
3. If partial allocation not possible then
4. Approach1: drop the request, no resource allocated remains vacant.
5. Approach2 (Optimization): drop the request take another request (same priority or lower priority) for which the resource can be accommodated. Delete from queue if request resources are allocated.
6. Approach3 (Optimization): Check if the existing request(s) support partial resource allocation then reduce the resources to accommodate. Delete from queue if request resources are allocated.
7. Approach4 (Optimization) – Use Approach1 till resource allocation deadline time has not reached. After deadline time if the current physical resource division can utilize resources of other division and there is a division which is vulnerable and has free resources then update then otherwise. If the previous step fails, then try first try Approach 2 and if that too fails try Approach3.
8. Approach5 (Optimization) – Use Approach1 till resource allocation deadline time has not reached. After deadline time if the current physical resource division can utilize resources of other division and there is a division which is vulnerable and has free resources then update then otherwise. If the previous step fails, then try first try Approach 3 and if that too fails try Approach2.
9. After resource allocation inform user at least delta time in advance.
10. Delete from resource allocation queue.
11. Update priorities of unscheduled requests

**Flow Chart:**

**Basic Design for Resource allocation via Approach1:**

A diagram of a project

Description automatically generated

KPIs - (Per time slot/Till current time)

1. Number of Request per queue
   1. Number of User Request at a time in request queue (min/max/avg)
   2. Number of Resource allocation requests in request queue. (min/max/avg)
   3. Number of Request per physical resource division (min/max/avg/percentage)
2. Priority
   1. Highest priority
   2. Lower priority
   3. Avg priority
3. Latency
   1. Priority handling latency
   2. Resource allocation latency
   3. Delta time overrun
   4. Slot time overrun
   5. Deletion of request after scheduling
   6. Updating priorities for unscheduled requests
   7. Time spent in request queue
4. Resource utilization: - At a given timeslot maximum possible physical resource utilization. It should not be the case that user request is present, but resource is not scheduled.
   1. Number of unused resources total
   2. Number of unused resources per division
5. Number of Preemption
   1. Number of preemptions of higher priority request by lower priority.
   2. Number of preemptions of lower priority request by higher priority.

# Log Design

# Profiling

# Coverage

# Test Framework