

SUMMARY

[07] Example implementations described herein are directed to systems and methods to allocate resources for an intermittent load. The example implementations may involve a per user budget-based resource allocation that can obviate the need to provide deadline or load presence information on a per job basis.

[08] The subject matter includes a computer-implemented method for managing a pool of resources, which can involve receiving a job request for execution on the pool of resources; determining an amount of time to be utilized for executing the job request by: determining available units of resource time based on available resources from the pool of resources and estimated resource commitments over one or more units of time based on historical resource usage of the pool of resources; determining available resources for the job request from the pool of resources based on the available units of resource time; and determining a period of time that is capable of being allocated for the job request based on a budget of resources allocated for the job request and the available units of resource time; determining a resource allocation from the pool of resources, wherein the resource allocation spreads the job request over the amount of time; determining that the job request is capable of being executed for the amount of time; and executing the job request over the amount of time, according to the resource allocation, wherein the step of determining the available units of resource time comprises: determining a first available units of resource time based on available resources from the pool of resources and the estimated

resource commitments over one or more units of time based on the historical resource usage of the pool of resources; and determining a second available units of resource time based on available units of resource time of a user associated with the job request, the available units of resource time of the user associated with the job request determined based on a resource allocation budget assigned to the user associated with the job request and resource commitments of the user associated with the job request.

[09] The subject matter further includes a non-transitory computer readable medium, storing instructions for managing a pool of resources, the instructions which can involve receiving a job request for execution on the pool of resources; determining an amount of time to be utilized for executing the job request based on available resources from the pool of resources and historical resource usage of the pool of resources; determining a resource allocation from the pool of resources, wherein the resource allocation spreads the job request over the amount of time; determining that the job request is capable of being executed for the amount of time; and executing the job request over the amount of time, according to the resource allocation.

[10] The subject matter further includes a system configured to manage a pool of resources, the system which can involve a processor, configured to receive a job request from the pool of resources; determine an amount of time to be utilized for executing the job request, wherein the processor is configured to: determine available units of resource time based on available resources from the pool of resources and estimated resource commitments over one or more units of time based on historical resource usage of the pool of resources; determine available resources for the job request from the pool of resources based on the available units of resource time; and determine a period of time that is capable of being allocated for the job request based on a budget of resources allocated for the job request and the available units of resource time; determine a resource allocation from the pool of resources, wherein the resource allocation spreads the job request over the amount of time; determine that the job request is capable of being executed for the amount of time; and execute the job request over the amount of time, according to the resource allocation, wherein to determine the available units of resource time, the processor is configured to: determine a first available units of resource time based on available resources from the pool of resources and the estimated resource commitments over one or more units of time based on the historical resource usage of the pool of resources; and determine a second available units of resource time based on available units of resource time of a user associated with the job request, the available units of resource time of the user associated with the job request determined based on a resource allocation budget assigned to the user associated with the job request and resource commitments of the user associated with the job request.

[11] The subject matter also includes a system for managing a pool of resources, the system which can involve means for receiving a job request for execution on the pool of resources; means for determining an amount of time to be utilized for executing the job request based on available resources from the pool of resources and historical resource usage of the pool of resources; means for determining a resource allocation from the pool of resources, wherein the resource allocation spreads the job request over the amount of time; means for determining that the job request is capable of being executed for the amount of time; and means for executing the job request over the amount of time, according to the resource allocation.

[12] The methods are implemented using one or more computing devices and/or systems. The methods may be stored in computer-readable media such as a non-transitory computer readable medium.

WHAT IS CLAIMED IS:

1. A computer-implemented method for managing a pool of resources, the method comprising:

receiving a job request for execution on the pool of resources;

determining an amount of time to be utilized for executing the job request by:

determining available units of resource time based on available resources from the pool of resources and estimated resource commitments over one or more units of time based on historical resource usage of the pool of resources;

determining available resources for the job request from the pool of resources based on the available units of resource time; and

determining a period of time that is capable of being allocated for the job request based on a budget of resources allocated for the job request and the available units of resource time;

determining a resource allocation from the pool of resources, wherein the resource allocation spreads the job request over the amount of time;

determining that the job request is capable of being executed for the amount of time; and

executing the job request over the amount of time, according to the resource allocation,

wherein the step of determining the available units of resource time comprises:

determining a first available units of resource time based on available resources from the pool of resources and the estimated resource commitments over one or more units of time based on the historical resource usage of the pool of resources; and

determining a second available units of resource time based on available units of resource time of a user associated with the job request, the available units of resource time of the user associated with the job request determined based on a resource allocation budget assigned to the user associated with the job request and resource commitments of the user associated with the job request.

2. The computer-implemented method of claim 1, wherein the determining the period of time that is capable of being allocated for the job request based on a budget of resources allocated for the job request and the available units of resource time comprises:

for each unit of time, determining the resource allocation from the pool of resources based on the available resources for the each unit of time and the estimated commitments for the each unit of time until the job request is fully allocated; and

determining the period of time from the resource allocation.

3. The computer-implemented method of claim 2, wherein for each unit of time, the determining the resource allocation from the pool of resources based on the

available resources for the each unit of time and the estimated commitments for the each unit of time until the job request is fully allocated further comprises:

for the each unit of time, determining the resource allocation based on the resource commitments of the user associated with the job request for the each unit of time and a resource allocation budget assigned to the user associated with the job request for the each unit of time.

4. The computer-implemented method of claim 1, wherein the determination that the job request is capable of being executed for the amount of time is based on a determination that the amount of time to be utilized for executing the job request meets a time requirement associated with the job request.

5. The computer-implemented method of claim 1, wherein the determination that the job request is capable of being executed for the amount of time is based on a determination that a downtime associated with the resource allocation does not exceed a maximum downtime associated with the job request.

6. The computer-implemented method of claim 5, wherein the determination that a downtime associated with the resource allocation does not exceed the maximum downtime associated with the job request comprises:

determining the downtime of the resource allocation based on determining one or more units of time in the resource allocation being indicative of non-allocation;

for the downtime not exceeding the maximum downtime, determining that the job request is capable of being executed; and

for the downtime exceeding the maximum downtime, determining that the job request is not capable of being executed.

7. The computer-implemented method of claim 1, wherein the determination that the job request is capable of being executed for the amount of time is based on a determination that the resource allocation can be executed.

8. The computer-implemented method of claim 1, wherein the determining the resource allocation comprises, for each unit of time:

determining the resource allocation for the each unit of time based on a budget associated with the job request and available resources from the pool of resources for the each period of time.

9. The computer-implemented method of claim 1, wherein the executing the job request comprises, for each unit of time in the amount of time, allocating resources from the pool of resources according to the resource allocation corresponding to the each unit of time.

10. The computer-implemented method of claim 1, further comprising:

managing a resource budget for each user of a plurality users of the pool of resources;

determining queuing for the plurality of users based on resource commitments of the plurality of users and the resource budget for the each user of the plurality of users.

11. The computer-implemented method of claim 10, wherein the determining an amount of time to be utilized for executing the job request based on available resources from the pool of resources and historical resource usage of the pool of resources comprises:

estimating resource commitments for each of the plurality of users based on the historical resource usage of the pool, and

estimating the available resources from the pool of resources from the estimated resource commitments.

12. The computer-implemented method of claim 11, wherein the determining the resource allocation from the pool of resources comprises, for each unit of time, determining the resource allocation for the job request based on the estimated available resources for the each unit of time.

13. The computer-implemented method of claim 10, wherein the pool of resources comprises storage and central processing units (CPUs) provided from a cloud pool, wherein the job request comprises a request of at least one of processing and memory.

14. The computer-implemented method of claim 10, wherein the managing the resource budget for each user of a plurality users of the pool of resources comprises managing resource allocation to one or more client devices associated with each user based on the resource budget associated with each user.

15. The computer-implemented method of claim 10, wherein the determining the resource allocation from the pool of resources is based on the queuing of the user from the plurality of the users associated with the job request.

16. The computer-implemented method of claim 15, wherein the determining the resource allocation from the pool of resources is based on the queuing of the user from the plurality of the users associated with the job request comprises:

for each unit of time, estimating the queuing of the user associated with the job request; and

estimating resource commitments of users from the plurality of users ahead of the user associated with the job request in the queue.

17. A system configured to manage a pool of resources, the system comprising:

a processor, configured to:

receive a job request from the pool of resources;

determine an amount of time to be utilized for executing the job request,
wherein the processor is configured to:

determine available units of resource time based on available resources
from the pool of resources and estimated resource commitments over one or
more units of time based on historical resource usage of the pool of resources;

determine available resources for the job request from the pool of
resources based on the available units of resource time; and

determine a period of time that is capable of being allocated for the job
request based on a budget of resources allocated for the job request and the
available units of resource time;

determine a resource allocation from the pool of resources, wherein the
resource allocation spreads the job request over the amount of time;

determine that the job request is capable of being executed for the amount of
time; and

execute the job request over the amount of time, according to the resource allocation,

wherein to determine the available units of resource time, the processor is configured to:

determine a first available units of resource time based on available resources from the pool of resources and the estimated resource commitments over one or more units of time based on the historical resource usage of the pool of resources; and

determine a second available units of resource time based on available units of resource time of a user associated with the job request, the available units of resource time of the user associated with the job request determined based on a resource allocation budget assigned to the user associated with the job request and resource commitments of the user associated with the job request.