CLAIMS:

1. A method comprising:

collecting data related to usage of a plurality of resources by a plurality of workloads over one or more time periods, the workloads including priority workloads associated with one or more guaranteed performance levels and best effort workloads not associated with guaranteed performance levels;

analyzing the data to identify resource reallocations from one or more of the priority workloads to one or more of the best effort workloads in one or more subsequent time periods while still maintaining the guaranteed performance levels, wherein analyzing comprises:

performing reinforcement learning using the data to identify the resource reallocations while still maintaining the guaranteed performance levels,

wherein performing reinforcement learning further comprises:

generating a first one or more reward values associated with resource allocations to the best effort workloads;

generating a second one or more reward values and/or one or more penalty values responsive to detecting specified performance metrics associated with the priority workloads;

adding the reward values and penalty values to generate a final reward value; and

reallocating resources to attempt to maximize the final reward value;

reallocating the resources from the priority workloads to the best effort workloads for the subsequent time periods;

monitoring execution of the priority workloads with respect to the guaranteed performance level during the subsequent time periods; and

preemptively reallocating resources from the best effort workloads to the priority workloads during the subsequent time periods responsive to detecting that the guaranteed performance level is in danger of being breached.

2. The method of claim 1 wherein the guaranteed performance level comprises a guaranteed latency and/or guaranteed throughput.

3. The method of claim 2 wherein the guaranteed performance level is specified as a key performance indicator (KPI) of a service level agreement (SLA).

4. The method of any of claims 1 to 3 wherein the resource allocations include a cache allocation to the best effort workloads, wherein an increase in the cache allocation is to generate an increased reward value.

5. The method of claim 4 wherein the second one or more reward values comprises a performance reward value for maintaining compliance with one or more of the guaranteed performance levels.

6. The method of any of claims 1 to 5 wherein a first set of the resource allocations are to be performed by resource management circuitry of a processor.

7. The method of claim 6 wherein the first set of resource allocations include a cache occupancy level and cache or memory bandwidth.

8. An apparatus comprising:

a telemetry data collector to collect data related to usage of a plurality of resources by a plurality of workloads over one or more time periods, the workloads including priority workloads associated with one or more guaranteed performance levels and best effort workloads not associated with guaranteed performance levels;

a resource allocation controller to analyze the data to identify resources to be reallocated from one or more of the priority workloads to one or more of the best effort workloads in one or more subsequent time periods while still maintaining the guaranteed performance levels, the resource allocation controller to reallocate the resources from the priority workloads to the best effort workloads for the subsequent time periods,

wherein the resource allocation controller comprises:

a machine learning engine to perform reinforcement learning using the data to identify the resource reallocations while still maintaining the guaranteed performance levels,

wherein performing reinforcement learning further comprises:

generating a first one or more reward values associated with resource allocations to the best effort workloads;

generating a second one or more reward values and/or one or more penalty values responsive to detecting specified performance metrics associated with the priority workloads;

adding the reward values and penalty values to generate a final reward value; and

reallocating resources to attempt to maximize the final reward value;

the telemetry data collector to monitor execution of the priority workloads with respect to the guaranteed performance level during the subsequent time periods; and

the resource allocation controller to preemptively reallocate resources from the best effort workloads to the priority workloads during the subsequent time periods responsive to detecting that the guaranteed performance level is in danger of being breached.

9. The apparatus of claim 8 wherein the guaranteed performance level comprises a guaranteed latency and/or guaranteed throughput.

10. The apparatus of claim 9 wherein the guaranteed performance level is specified as a key performance indicator (KPI) of a service level agreement (SLA).

11. The apparatus of any of claims 8 to 10 wherein the resource allocations include a cache allocation to the best effort workloads, wherein an increase in the cache allocation is to generate an increased reward value.

12. The apparatus of claim 11 wherein the second one or more reward values comprises a performance reward value for maintaining compliance with one or more of the guaranteed performance levels.

13. The apparatus of any of claims 8 to 12 wherein at least a portion of the resource allocation controller comprises resource management circuitry within a processor, the resource management circuitry to perform a first set of the resource allocations.

14. The apparatus of claim 13 wherein the first set of resource allocations include a cache occupancy level and cache or memory bandwidth.

15. A machine-readable medium having program code stored thereon which, when executed by a machine, causes the machine to perform the operations of:

collecting data related to usage of a plurality of resources by a plurality of workloads over one or more time periods, the workloads including priority workloads associated with one or more guaranteed performance levels and best effort workloads not associated with guaranteed performance levels;

analyzing the data to identify resource reallocations from one or more of the priority workloads to one or more of the best effort workloads in one or more subsequent time periods while still maintaining the guaranteed performance levels,

wherein analyzing comprises:

performing reinforcement learning using the data to identify the resource reallocations while still maintaining the guaranteed performance levels.

wherein performing reinforcement learning further comprises:

generating a first one or more reward values associated with resource allocations to the best effort workloads;

generating a second one or more reward values and/or one or more penalty values responsive to detecting specified performance metrics associated with the priority workloads;

adding the reward values and penalty values to generate a final reward value; and

reallocating resources to attempt to maximize the final reward value;

reallocating the resources from the priority workloads to the best effort workloads for the subsequent time periods;

monitoring execution of the priority workloads with respect to the guaranteed performance level during the subsequent time periods; and

preemptively reallocating resources from the best effort workloads to the priority workloads during the subsequent time periods responsive to detecting that the guaranteed performance level is in danger of being breached.

16. The machine-readable medium of claim 15 wherein the guaranteed performance level comprises a guaranteed latency and/or guaranteed throughput.

17. The machine-readable medium of claim 16 wherein the guaranteed performance level is specified as a key performance indicator (KPI) of a service level agreement (SLA).

18. The machine-readable medium of any of claims 15 to 17 wherein the resource allocations include a cache allocation to the best effort workloads, wherein an increase in the cache allocation is to generate an increased reward value.

19. The machine-readable medium of claim 18 wherein the second one or more reward values comprises a performance reward value for maintaining compliance with one or more of the guaranteed performance levels.

20. The machine-readable medium of any of claims 15 to 19 wherein a first set of the resource allocations are to be performed by resource management circuitry of a processor.

21. The machine-readable medium of claim 20 wherein the first set of resource allocations include a cache occupancy level and cache or memory bandwidth.

22. An apparatus comprising:

means for collecting data related to usage of a plurality of resources by a plurality of workloads over one or more time periods, the workloads including priority workloads associated with one or more guaranteed performance levels and best effort workloads not associated with guaranteed performance levels;

means for analyzing the data to identify resource reallocations from one or more of the priority workloads to one or more of the best effort workloads in one or more subsequent time periods while still maintaining the guaranteed performance levels,

wherein means for analyzing comprises:

means for performing reinforcement learning using the data to identify the resource reallocations while still maintaining the guaranteed performance levels,

wherein means for performing reinforcement learning further comprises:

means for generating a first one or more reward values associated with resource allocations to the best effort workloads;

means for generating a second one or more reward values and/or one or more penalty values responsive to detecting specified performance metrics associated with the priority workloads;

means for adding the reward values and penalty values to generate a final reward value; and

means for reallocating resources to attempt to maximize the final reward value;

means for reallocating the resources from the priority workloads to the best effort workloads for the subsequent time periods;

means for monitoring execution of the priority workloads with respect to the guaranteed performance level during the subsequent time periods; and

means for preemptively reallocating resources from the best effort workloads to the priority workloads during the subsequent time periods responsive to detecting that the guaranteed performance level is in danger of being breached.

23. The apparatus of claim 22 wherein the guaranteed performance level comprises a guaranteed latency and/or guaranteed throughput.

24. The apparatus of claim 23 wherein the guaranteed performance level is specified as a key performance indicator (KPI) of a service level agreement (SLA).

25. The apparatus of any of claims 22 to 24 wherein the resource allocations include a cache allocation to the best effort workloads, wherein an increase in the cache allocation is to generate an increased reward value.

26. The apparatus of claim 25 wherein the second one or more reward values comprises a performance reward value for maintaining compliance with one or more of the guaranteed performance levels.

27. The apparatus of any of claims 22 to 26 wherein a first set of the resource allocations are to be performed by resource management circuitry of a processor.

28. The apparatus of claim 27 wherein the first set of resource allocations include a cache occupancy level and cache or memory bandwidth.