WHAT IS CLAIMED IS:

1. A system, comprising:

a plurality of processor cores;

a plurality of graphics processing units;

a plurality of peripheral devices distinct from the processor cores and graphics processing units;

one or more memory controller circuits configured to interface with a system memory; and

an interconnect fabric configured to provide communication between the one or more memory controller circuits and the processor cores, graphics processing units, and peripheral devices;

wherein the processor cores, graphics processing units, peripheral devices, and memory controller circuits are configured to communicate via a unified memory architecture; and

wherein the processor cores, graphics processing units, peripheral devices, one or more memory controller circuits, and interconnect fabric are included in a system on a chip (SOC) integrated onto one or more co-packaged semiconductor dies.

2. The system of claim 1, wherein the processor cores, graphics processing units, and peripheral devices are configured to access any address within a unified address space defined by the unified memory architecture.

3. The system of claim 2, wherein the unified address space is a virtual address space distinct from a physical address space provided by the system memory.

4. The system of any of claims 1-3 wherein the unified memory architecture provides a common set of semantics for memory access by the processor cores, graphics processing units, and peripheral devices.

5. The system of claim 4 wherein the semantics include memory ordering properties.

6. The system of claims 4 or 5 wherein the semantics include quality of service attributes.

7. The system of any of claims 4-6 wherein the semantics include memory coherency.

8. The system of any preceding claim, wherein the one or more memory controller circuits include respective interfaces to one or more memory devices that are mappable of random access memory.

9. The system of claim 8 wherein one or more memory devices comprise dynamic random access memory (DRAM).

10. The system of any preceding claim, further comprising one or more levels of cache between the processor cores, graphics processing units, peripheral devices, and the system memory.

11. The system as recited in claim 10 wherein the one or more memory controller circuits include respective memory caches interposed between the interconnect fabric and the system memory, wherein the respective memory caches are one of the one or more levels of cache.

12. The system of any preceding claim, wherein the interconnect fabric comprises at least two networks having heterogeneous interconnect topologies.

13. The system of any preceding claim, wherein the interconnect fabric comprises at least two networks having heterogeneous operational characteristics.

14. The system of claim 12 or 13, wherein the at least two networks include a coherent network interconnecting the processor cores and the one or more memory controller circuits.

15. The system of any of claims 12-14, wherein the at least two networks include a relaxed-ordered network coupled to the graphics processing units and the one or more memory controller circuits.

16. The system of claim 15, wherein the peripheral devices include a subset of devices, wherein the subset includes one or more of a machine learning accelerator circuit or a relaxed-order bulk media device, and wherein the relaxed-ordered network is further coupled to the subset of devices to the one or more memory controller circuits.

17. The system of any of claims 12-16, wherein the at least two networks include an input-output network coupled to interconnect the peripheral devices and the one or more memory controller circuits.

18. The system of claim 17, wherein the peripheral devices include one or more real-time devices.

19. The system of any of claims 12-18 wherein the at least two networks comprise a first network that comprises one or more characteristics to reduce latency compared to a second network of the at least two networks.

20. The system of claim 19 wherein the one or more characteristics comprise a shorter route than the second network.

21. The system of claim 19 or 20 wherein the one or more characteristics comprise wiring in metal layers nearer a surface of a substrate on which the system is implemented than the wiring for the second network.

22. The system of any of claims 12-21 wherein the at least two networks comprise a first network that comprises one or more characteristics to increase bandwidth compared to a second network of the at least two networks.

23. The system of claim 22 wherein the one or more characteristics comprise wider interconnect compared to the second network.

24. The system of claim 22 or 23 wherein the one or more characteristics comprise wiring in metal layers farther from a surface of a substrate on which the system is implemented than the wiring for the second network.

25. The system of any of claims 12-24, wherein interconnect topologies employed by the at least two networks include at least one of a star topology, a mesh topology, a ring topology, a tree topology, a fat tree topology, a hypercube topology, or a combination of one or of the interconnect topologies.

26. The system of any of claims 12-25, wherein the operational characteristics employed by the at least two networks include at least one of strongly-ordered memory coherence or relaxed-ordered memory coherence.

27. The system of any of claims 12-26 wherein the at least two networks are physically and logically independent.

28. The system of any of claims 12-27 wherein the at least two networks are physically separate in a first mode of operation, and wherein a first network of the at least two networks and a second network of the at least two networks are virtual and share a single physical network in a second mode of operation.

29. The system of any preceding claim, wherein the processor cores, graphics processing units, peripheral devices, and interconnect fabric are distributed across two or more integrated circuit dies.

30. The system of claim 29, wherein a unified address space defined by the unified memory architecture extends across the two or more integrated circuit dies in a manner transparent to software executing on the processor cores, graphics processing units, or peripheral devices.

31. The system of any of claims 29-30 wherein the interconnect fabric is configured to route wherein the interconnect fabric extends across the two integrated circuit dies and wherein a communication is routed between a source and a destination transparent to a location of the source and the destination on the integrated circuit dies.

32. The system of any of claims 29-31 wherein the interconnect fabric extends across the two integrated circuit dies using hardware circuits to automatically route a communication between a source and a destination independent of whether or not the source and destination are on the same integrated circuit die.

33. The system of any of claims 29-32, further comprising at least one interposer device configured to couple buses of the interconnect fabric across the two or integrated circuit dies.

34. The system of any of claims 1-33 wherein a given integrated circuit die includes a local interrupt distribution circuit to distribute interrupts among processor cores in the given integrated circuit die.

35. The system of claim 34 comprising two or more integrated circuit dies that include respective local interrupt distribution circuits and at least one of the two or more integrated circuit dies includes a global interrupt distribution circuit, wherein the local

interrupt distribution circuits and the global interrupt distribution circuit implement a multi-level interrupt distribution scheme.

36. The system of claim 35 wherein the global interrupt distribution circuit is configured to transmit an interrupt request to the local interrupt distribution circuits in a sequence, and wherein the local interrupt distribution circuits are configured to transmit the interrupt request to local interrupt destinations in a sequence before replying to the interrupt request from the global interrupt distribution circuit.

37. The system of any of claims 1-36 wherein a given integrated circuit die comprises a power manager circuit configured to manage a local power state of the given integrated circuit die.

38. The system of claim 37 comprising two or more integrated circuit dies that include respective power manager circuits configured to manage the local power state of the integrated circuit die, and wherein at least one of the two or more integrated circuit die includes another power manager circuit configured to synchronize the power manager circuits.

39. The system of any preceding claim, wherein the peripheral devices include one of more of: an audio processing device, a video processing device, a machine learning accelerator circuit, a matrix arithmetic accelerator circuit, a camera processing circuit, a display pipeline circuit, a nonvolatile memory controller, a peripheral component interconnect controller, a security processor, or a serial bus controller.

40. The system of any preceding claim, wherein the interconnect fabric interconnects coherent agents.

41. The system of claim 40, wherein an individual one of the processor cores corresponds to a coherent agent.

42. The system of claim 40, wherein a cluster of processor cores corresponds to a coherent agent.

43. The system of any preceding claim, wherein a given one of the peripheral devices is a non-coherent agent.

44. The system of claim 43, further comprising an input/output agent interposed between the given peripheral device and the interconnect fabric, wherein the input/output agent is configured to enforce coherency protocols of the interconnect fabric with respect to the given peripheral device.

45. The system of claim 44 wherein the input/output agent ensures the ordering of requests from the given peripheral device using the coherency protocols.

46. The system of claim 44 or 45 wherein the input/output agent is configured to couple a network of two or more peripheral devices to the interconnect fabric.

47. The system of any preceding claim, further comprising hashing circuitry configured to distribute memory request traffic to system memory according to a selectively programmable hashing protocol.

48. The system of claim 47 wherein at least one programming of the programmable hashing protocol evenly distributes a series of memory requests over a plurality of memory controllers in the system for a variety of memory requests in the series.

49. The system of claim 29, wherein at least one programming of the programmable hashing protocol distributes adjacent requests within a memory space, at a specified granularity, to physically distant memory interfaces.

50. The system of any preceding claim further comprising a plurality of directories configured to track a coherency state of subsets of a unified memory address space specified by the unified memory architecture, wherein the plurality of directories are distributed in the system.

51. The system of claim 50 wherein the plurality of directories are distributed to the memory controllers.

52. The system of any preceding claim wherein a given memory controller of the one or more memory controller circuits comprises a directory configured to track a plurality of cache blocks that correspond to data in a portion of the system memory to which the given memory controller interfaces, wherein the directory is configured to track which of a plurality of caches in the system are caching a given cache block of the plurality of cache blocks, wherein the directory is precise with respect to memory requests that have

been ordered and processed at the directory even in the event that the memory requests have not yet completed in the system.

53. The system of claim 52 wherein the given memory controller is configured to issue one or more coherency maintenance commands for the given cache block based on a memory request for the given cache block, wherein the one or more coherency maintenance commands include a cache state for the given cache block in a corresponding cache of the plurality of caches, wherein the corresponding cache is configured to delay processing of a given coherency maintenance command based on the cache state in the corresponding cache not matching the cache state in the a given coherency maintenance command.

54. The system of any of claims 52-53 wherein a first cache is configured to store the given cache block in a primary shared state and a second cache is configured to store the given cache block in a secondary shared state, and wherein the given memory controller is configured to cause the first cache transfer the given cache block to a requestor based on the memory request and the primary shared state in the first cache.

55. The system of any of claims 52-54 wherein the given memory controller is configured to issue one of a first coherency maintenance command and a second coherency maintenance command to a first cache of the plurality of caches based on a type of a first memory request, wherein the first cache is configured to forward a first cache block to a requestor that issued the first memory request based on the first coherency maintenance command, and wherein the first cache is configured to return the first cache block to the given memory controller based on the second coherency maintenance command.

56. An integrated circuit comprising:

a plurality of processor cores;

a plurality of graphics processing units;

a plurality of peripheral devices distinct from the processor cores and graphics processing units;

one or more memory controller circuits configured to interface with a system memory;

an interconnect fabric configured to provide communication between the one or more memory controller circuits and the processor cores, graphics processing units, and peripheral devices; and

an off-chip interconnect coupled to the interconnect fabric and configured to couple the interconnect fabric to a corresponding interconnect fabric on another instance of the integrated circuit, wherein the interconnect fabric and the off-chip interconnect provide an interface that transparently connects the one or more memory controller circuits, the processor cores, graphics processing units, and peripheral devices in either a single instance of the integrated circuit or two or more instances of the integrated circuit.

57. A method comprising:

communicating, via a unified memory architecture, among a plurality of processor cores, a plurality of graphics processing units, a plurality of peripheral devices distinct from the plurality of processor cores and the plurality of graphics processing units, and one or more memory controller circuits over an interconnect fabric in a system, wherein the processor cores, graphics processing units, peripheral devices, one or more memory

controller circuits, and interconnect fabric are included in a system on a chip (SOC) integrated onto one or more co-packaged semiconductor dies.