Aitken (pronounced AY-ken) is a modular computing system housed in NASA's Modular Supercomputing Facility, located less than a mile from the main NAS building. Aitken is configured as follows:

Intel Cascade Lake-Based Compute Resources

- 4 E-Cells (8 E-racks)
- 1,116 Cascade Lake nodes
- 44,640 cores
- 214 terabytes (TB) total memory

AMD Rome-Based Compute Resources

- 16 racks
- 2,028 Rome nodes
- 259,584 cores
- 1.038 petabyte (PB) total memory

AMD Milan-Based Compute Resources

- 4 racks
- 512 Milan nodes
- 65,536 cores
- 262 terabytes (TB) total memory

Aitken's total theoretical peak performance: 15.49 petaflops (PF)

Hostnames

For the Cascade Lake-based resources, there are four enclosures (individual rack units, or IRUs) in each E-rack, with 36 compute nodes per enclosure. For every two racks, there is a rack leader controlling them. The naming convention for the 288 nodes residing in every two racks (2 racks x 4 enclosures x 36 nodes) uses only odd rack numbers. Therefore, the hostnames of the Aitken compute nodes are r[x]i[0-7]n[0-35], where x are odd numbers between 901 and 907.

For the Rome-based resources, there are four enclosures in each rack, eight compute trays in each enclosure, and four nodes in each compute tray. The hostnames of the Aitken Rome nodes are r[x]c[1-4]t[1-8]n[1-4], where x are numbers between 201 - 216.

For the Milan-based resources, there are four enclosures in each rack, eight compute trays in each enclosure, and four nodes in each compute tray. The hostnames of the Aitken Milan nodes are r[x]c[1-4]t[1-8]n[1-4], where x are numbers between 705 and 708.

Processor, Memory, and Network Subsystems Statistics

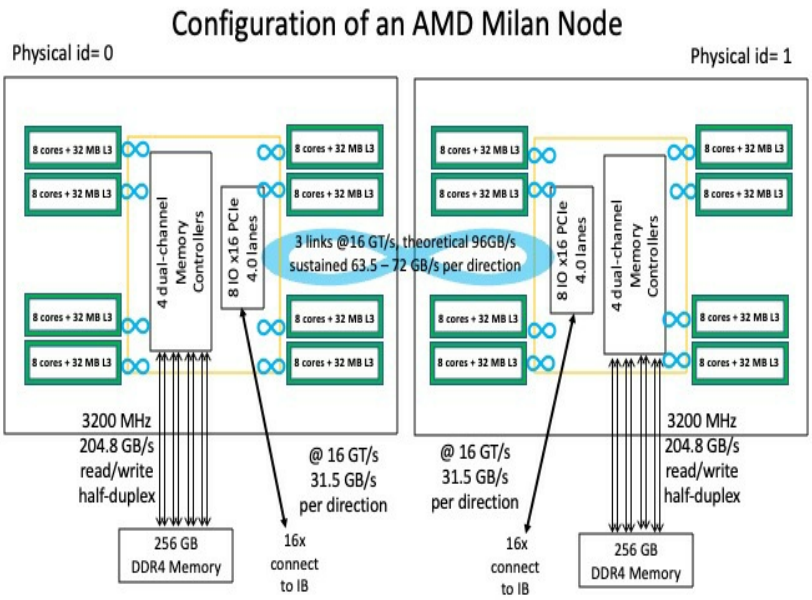
The following table provides detailed configuration statistics for the processor, memory, and network subsystems for the Aitken compute nodes:

Aitken Processor, Memory, and Network Subsystems Statistics

Architecture	HPE SGI 8600-XA730i Gen10	HPE Apollo 9000	HPE Apollo 9000
		Processor	
	Cascade Lake	**Rome**	**Milan**
CPU	20-Core Xeon Gold 6248	64-Core EPYC 7742	64-Core EPYC 7763
Newest Instruction Set	AVX-512	AVX2	AVX2
Hyper-Threading	ON	ON	ON
TurboBoost	ON	ON	ON
non-Turbo base CPU-Clock	2.5 GHz	2.25 GHz	2.45 GHz
Maximum Double Precision Floating Point Operations per Cycle per Core	32	16	16
# of Cores/Node	40	128	128
Total # of Nodes	1,116	2,028	512
Total # of Cores	44,640	259,584	65,536
Total Double Precision TFlops	3,686	9,437	2,569
		Memory	

L1 Cache	Local to each core; Instruction cache: 32K Data cache: 32K; Associativity: 8; Cache line size: 64B	Local to each core; Instruction cache: 32K Data cache: 32K; Associativity: 8; Cache line size: 64B	Local to each core; Instruction cache: 32K Data cache: 32K; Associativity: 8; Cache line size: 64B
L2 Cache	1 MB per core; Associativity: 16; Cache line size: 64B	512 KB per core; Associativity: 8; Cache line size: 64B	512 KB per core; Associativity: 8; Cache line size: 64B
L3 Cache	27.5 MB shared non-inclusive by the 20 cores; Associativity: full; Cache line size: 64B	16 MB shared among 4 cores in a core complex; 256 MB per socket; Associativity: 16; Cache line size: 64B	32 MB shared among 8 cores in a core complex; 256 MB per socket; Associativity: 16; Cache line size: 64B
TLB	Local to each core	Local to each core	Local to each core
Default Page Size	4 KB	4 KB	4 KB
Memory/Core	4.8 GB; DDR4	4.0 GB; DDR4	4.0 GB; DDR4
Total Memory/node	192 GB	512 GB	512 GB
Memory Speed and Bandwidth	2,933 MHz; 6 channels; 141 GB/sec read/write	3,200 MHz; 8 channels; 204.8 GB/sec read/write	3,200 MHz; 8 channels; 204.8 GB/sec read/write
Inter-socket Interconnect	Ultra-Path Interconnect; 5.2 GHz, 10.4 GT/s, or 62.4 GB/sec	Global Memory Interconnect; 8.0 GHz, 16.0 GT/s, or 96.0 GB/sec	Global Memory Interconnect; 8.0 GHz, 16.0 GT/s, or 96.0 GB/sec
Inter-node Network			
IB Device on Node	Dual single-port 4x EDR IB Mezzanine card (2 single-port HCAs); 100 Gbits/s	Dual single-port 4x HDR IB Mezzanine card (2 single-port HCAs); 200 Gbits/s	Dual single-port 4x HDR IB Mezzanine card (2 single-port HCAs); 200 Gbits/s
IB Switches Between Nodes	4x HDR; 200 Gbits/s	4x HDR; 200 Gbits/s	4x HDR; 200 Gbits/s

The Milan nodes are currently installed on the computer floor of the main NAS building.



This is a simplified configuration of an EPYC Milan node with two sockets. Each socket contains 64 cores in eight Core Complex Dies (CCDs, each enclosed in a green box) and one I/O die (IOD, enclosed in a yellow box). The ♾ sign represents the Infinity Fabric.

The Milan processor to be incorporated into the Aitken cluster is EPYC 7763 with a base clock of 2.45 GHz and a maximum boost clock of 3.5 GHz.

The Milan (Zen 3) microarchitecture bears many similarities with the Rome (Zen 2) microarchitecture. For a description on the processor, cache, memory subsystems, intra-socket and inter-socket interconnect for Milan, refer to the information provided for the Rome processor.

Main Differences Between Milan and Rome

Key differences between the two microarchitectures are described in this section.

Core Complex Dies (CCDs)

A Rome CCD contains two separate Core CompleXes (CCXs), each with 4 cores and 16 MB of L3 cache, while a Milan CCD contains a unified CCX with 8 cores and 32 MB of L3 cache which results in:

- Direct communication and therefore lower core-to-core latency among the 8 cores in a CCD.
- Doubling the size of L3 cache: all 32 MB can be accessed by a single core in the CCD.
- Average L3 cache latency: 39 cycles for Rome, 46 cycles for Milan.
- 19% higher instructions per cycle (IPC) on average.

Clock Coupling

With Rome, you can choose whether to couple the Infinity Fabric clock (FCLK) and the DRAM memory clock (MEMCLK) via a BIOS setting. With Milan, these two clocks are synchronized, resulting in a reduced memory latency.

Note: The maximum Infinity Fabric clock speed is 1,467 MHz for Rome and 1,600 MHz for Milan.

NUMA Domain

In Milan, when the BIOS setting BIOS.ProcSettings.CcxAsNumaDomain is enabled, each CCD is exposed as a NUMA node and a dual-socket Milan node will have a total of 16 NUMA nodes instead of 8. Use the following command to find out the number of NUMA nodes configured in a Milan node: `numactl -H`

Note: Each NAS Rome node is configured with 4 NUMA Nodes (NPS4) per socket, for a total of 8 NUMA Nodes per Rome node.

CPU Core

Milan is a 64-bit x86 server microprocessor. Like the Rome processor, Milan supports AVX2 (256 bit) but not the AVX-512 instructions. For a complete list of instructions supported by Milan, run the following command on a Milan node: `cat /proc/cpuinfo`

Each Milan core has two 256-bit Fused Multiply-Add (FMA) units and can deliver up to 16 double-precision floating point operations (flops) per cycle. Therefore, the peak double-precision flops per node is: 128 cores x 2.45 GHz x 16 = 5.02 teraflops (TF). With 512 Milan nodes in Aitken, the peak is 2.57 petaflops (PF).

The simultaneous multi-threading (SMT) capability allows 2 execution threads to execute simultaneously per core, is enabled/disabled in BIOS.

Note: When the Milan nodes are in their early production period, it is likely that some nodes will have SMT enabled while others will have SMT disabled. To check the SMT setting on a Milan node, use the following command:

```
Milan_node% lscpu | grep "Thread(s) per core"
```

If the output is 2, SMT is enabled. If it is 1, SMT is disabled.

Note: Linux support for the Zen microarchitecture started with Linux kernel 4.10. The current Red Hat-based operating system—Tri-Lab Operating System Stack Version 4 (TOSS 4)—uses Linux kernel 4.18.0, which supports the Rome and Milan nodes.

You can use the AMD Optimizing C/C++ (AOCC), GNU, and Intel compilers to compile codes to run on the Milan nodes. To generate optimized x64 code for the Zen3 microarchitecture, consider these compiler flags:

- AOCC compiler: `-march=znver3`

  Note: AOCC consists of a C/C++ compiler (clang) and a Fortran compiler (flang). AOCL (AMD Open CPU Libraries = AMS CPU math library suite) consists of optimized math libraries. AOCC 3.0 and AOCL 3.0 became available when the 3rd Gen EPYC processors launched. For information about additional optimization for the 4th generation EPYC processors, see AOCC and AOCL.

- GNU compiler: `-march=znver3`

  Note: `-march=znver3` is available in gcc 11.x and above but not earlier versions. With gcc 8.x, use `-march=znver1` instead.

- Intel compiler: `-march=core-avx2` (preferred) or `-axCORE-AVX2`

For more information, see AMD's Compiler Options Quick Reference Guide for the AMD EPYC 7xx3-series Processors.

Inter-Node Network

There are 512 Milan nodes in the Aitken cluster, which uses the HPE Apollo 9000 system architecture. They are partitioned as follows:

- 4 racks in the cluster (512 nodes total)
- 4 enclosures per rack (128 nodes/rack)
- 8 trays per enclosure (32 nodes/enclosure)
- 4 nodes per compute tray

The hostname of each node is r[x1]c[x2]t[x3]n[x4] where x1 = 705 - 708; x2 = 1 to 4; x3 = 1 to 8; and x4 = 1 to 4.

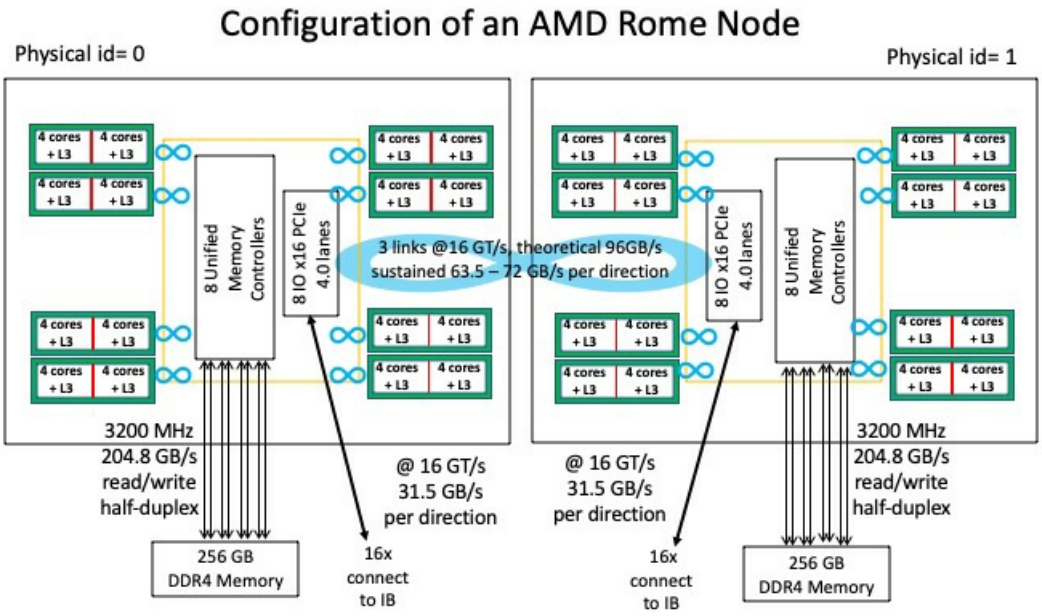
Like the Rome nodes, the inter-node connection of the Milan nodes in Aitken relies on the use of PCIe 4.0 lanes and one dual single-port IB HDR 200 Gbit/s Mezzanine card on each node and two IB HDR 200 Gbit/s premium switches per enclosure. One switch per enclosure facilitates the ib0 fabric while the other facilitates the ib1 fabric. For more information, see the Inter-Node Network section of the Rome Processors.

Connection between Milan and Rome

When the Milan nodes are relocated to the new compute module, they will be connected to the Rome nodes with additional HDR cables on the 8th dimension of the Aitken cluster.

References

- AMD EPYC 7763
- High Performance Computing (HPC) Tuning Guide for the AMD EPYC 7003 Series Processors
- Workload Tuning Guide for AMD EPYC 7003 Processors
- Compiler Options Quick Reference Guide for the AMD EPYC 7xx3-series Processors



This is a simplified configuration of an EPYC Rome node with two sockets. Each socket contains eight Core Complex Dies (CCDs, each enclosed in a green box) and one I/O die (IOD, enclosed in a yellow box). The infinity sign (♾) represents the Infinity Fabric. Each CCD contains two Core Complexes (CCXs). Each CCX has 4 cores and 16 MB of L3 cache. Thus, there are 64 cores per socket and 128 cores per node.

The AMD EPYC (pronounced "epic") 7742 Rome processor, incorporated into Aitken with the HPE Apollo 9000 system architecture, is in AMD's second generation System-on-Chip (SoC) processor family. Rome has the following high-level features.

- Zen 2 microarchitecture:

  The EPYC 7742 Rome processor has a base CPU clock of 2.25 GHz and a maximum boost clock of 3.4 GHz. There are eight processor dies (CCDs) with a total of 64 cores per socket.

- Hybrid multi-die design:

  Within each socket, the eight processor dies are fabricated on a 7 nanometer (nm) process, while the I/O die is fabricated on a 14 nm process. This design decision was made because the processor dies need the leading edge (and more expensive) 7 nm technology in order to reduce the amount of power and space needed to double the number of cores, and to add more cache, compared to the first-generation EPYC processors. The I/O die retains the less expensive, older 14 nm technology.

- 2nd-generation Infinity Fabric technology:

  Infinity Fabric technology is used for communication among different components throughout the node: within cores, between cores, between CCXs in a CCD, among CCDs in a socket, to the main memory and PCIe, and between the two sockets. The Rome processors are the first x86 systems to support 4th-generation PCIe, which delivers twice the I/O performance (to InfiniBand, storage, NVMe SSD, etc.) over 3rd-generation PCIe.

Processor Hierarchy

The Rome processor hierarchy is as follows:

- Core: A CPU core has private L1I, L1D, and L2 caches, which are shared by two hyperthreads on the core.
- CCX: A core complex includes four cores and a common L3 cache of 16 MB. Different CCXs do not share L3.
- CCD: A core complex die includes two CCXs and an Infinity Link to the I/O die (IOD). The CCDs connect to memory, I/O, and each other through the IOD.
- Socket: A socket includes eight CCDs (total of 64 cores), a common centralized I/O die (includes eight unified memory controllers and eight IO x16 PCIe 4.0 lanes—total of 128 lanes), and a link to the network interface controller (NIC).
- Node: A node includes two sockets and a network interface controller (NIC). In an Aitken Rome node, the NIC is a dual single-port InfiniBand (IB) High Data Rate (HDR) 200 Gbit/s mezzanine card.

CPU Core

Rome is a 64-bit x86 server microprocessor. A partial list of instructions and features supported in Rome includes SSE, SSE2, SSE3, SSSE3, SSE4a, SSE4.1, SSE4.2, AES, FMA, AVX, AVX2 (256 bit), Integrated x87 FPU (FPU), Multi-Precision Add-Carry (ADX), 16-bit Floating Point Conversion (F16C), and No-eXecute (NX). For a complete list, run `cat /proc/cpuinfo` on a Rome node.

Unlike the Intel Skylake and Cascade Lake processors, Rome does not come with the AVX-512 instructions.

Each Rome core:

- Can sustain execution of four x86 instructions per cycle, using features such as the micro-op cache, advanced branch prediction, and prefetching. The prefetcher works on streaming data and on variable strides, allowing it to accelerate many different data structures.
- Has two 256-bit Fused Multiply-Add (FMA) units and can deliver up to 16 double-precision floating point operations (flops) per cycle. Thus, the peak double-precision flops per node is: 128 cores x 2.25 GHz x 16 = 4.6 teraflops. With 2,048 Rome nodes in Aitken, the peak is 9.44 petaflops.
- Can support Simultaneous Multi-threading (SMT), allowing two threads to execute simultaneously per core. SMT can be enabled/disabled in the Basic Input/Output System (BIOS) settings.

  Note: The final setting is yet to be decided. If you plan to take advantage of SMT, use `cat proc/cpuinfo` to determine whether it has been enabled.

Note: Linux support for the Zen microarchitecture started with Linux kernel 4.10. The SUSE Linux Enterprise Server Version 15 (SLES 15) operating system, which was previously used on the Aitken Rome nodes, uses Linux kernel 5.3. Although the current Red Hat-based operating system, Tri-Lab Operating System Stack Version 3 (TOSS 3), uses Linux kernel 3.10, it includes the needed patches—backported from the newer kernels—that are required to support the Rome nodes.

The AMD Optimizing C/C++ (AOCC), GNU, Intel, and PGI compilers can be used to compile codes for running on the Rome nodes. To generate optimized x64 code for the Zen 2 microarchitecture, consider using these compiler flags:

- AOCC compiler: `-march=znver2`
- Intel compiler: `-march=core-avx2` (preferred) or `-axCORE-AVX2`
- PGI compiler: `-tp=zen`
- GNU compiler (GCC): `-march=znver2`

  Note: `-march=znver2` is available in the `gcc 9.x` compiler and later versions. With `gcc 8.x`, use `-march=znver1` instead.

For more information, see Compiler Options Quick Reference Guide for AMD EPYC 7xx2 Series Processors.

Cache Hierarchy

The Rome cache hierarchy is as follows:

- op cache (OC): 4K ops, private to each core; 64 sets; 64 bytes/line; 8-way. OC holds instructions that have already been decoded into micro-operations (micro-ops). This is useful when the CPU repeatedly executes a loop of code. Using OC improves:
  - Pipeline latency: because the op cache pipeline is shorter than the traditional fetch and decode pipeline.
  - Bandwidth: because the maximum throughput from the op cache is eight instructions per cycle, whereas the maximum throughput from the traditional fetch and decode pipeline is four instructions per cycle.
  - Power: because there is no need to re-decode instructions.
- L1 instruction cache: 32 KB, private to each core; 64 bytes/line; 8-way. The processor fetches instructions from the instruction cache in 32-byte naturally aligned blocks.
- L1 data cache: 32 KB, private to each core; 64 bytes/line; 8-way; latency: 7-8 cycles for floating point and 4-5 cycles for integer; 2 x 256 bits/cycle load bandwidth to registers; 1 x 256 bits/cycle store bandwidth from registers; write-back policy.

  Note: With the write-back policy, data is updated in the current level cache first. The update in the next level storage is done later when the cache line is ready to be replaced.
- L2 cache: 512 KB, private to each core; unified; inclusive of L1 cache; 64 bytes/line; 8-way; latency: >= 12 cycles; 1 x 256 bits/cycle load bandwidth to L1 cache; 1 x 256 bits/cycle store bandwidth from L1 cache; write-back policy.
- L3 cache: 16 MB shared among four cores in a core complex (CCX); different CCXs do not share L3; total of 256 MB per socket. Within each CCX: 64 bytes/line; 16-way; latency: 39 cycles on average.

  Note: If a core misses in its local L2 and also in the L3, the shadow tags are consulted. If the shadow tag indicated that the data resides in another L2 within the CCX, a cache-to-cache transfer is initiated.
  1 x 256 bits/cycle load bandwidth to L2 of each core; 1 x 256 bits/cycle store bandwidth from L2 of each core; write-back policy; populated by L2 victims.

Memory Subsystem

Unlike the Zen 1 microarchitecture, in which each CCD has its own dual-channel memory controller, the Zen 2 microarchitecture places eight unified memory controllers in the centralized I/O die so that the memory latency is reduced (roughly at 200+ ns) and is more consistent. The memory channels can be split into one, two, or four Non-Uniform Memory Access (NUMA) Nodes per Socket (NPS1, NPS2, and NPS4). The BIOS default for HPE Apollo 9000 systems is NPS4, which is the highest memory bandwidth configuration geared toward HPC applications.

With eight 3,200-GHz memory channels, an 8-byte read or write operation taking place per cycle per channel results in a maximum total memory bandwidth of 204.8 GB/s per socket.

Each memory channel can be connected with up to two Double Data Rate (DDR) fourth-generation Dual In-line Memory Modules (DIMMs). For the Aitken Rome configuration, each channel is connected to a single 32-GB DDR4 registered DIMM (RDIMM) with error correcting code (ECC) support. In total, the amount of memory is 256 GB per socket and 512 GB per node.

Note: Using the one DIMM per Channel (DPC) configuration enables the system to run the memory DIMMs at the highest possible speed; a two-DPC configuration typically requires slightly reduced memory speed.

The memory frequency can be uncoupled, or it can be coupled with the Infinity Fabric frequency through BIOS settings to benefit either bandwidth-bound or latency-bound workloads. The uncoupled mode is used on the Aitken Rome nodes.

- Uncoupled Mode: For throughput-sensitive applications, to obtain higher read/write throughput, the Maximum Memory Bus Frequency option can be set to the maximum allowed (3,200 MT/s). In this case, the Memory Bus will not be synchronized optimally with the slower Infinity Fabric Clock, causing a slight increase in memory access latency.
- Coupled Mode: For latency sensitive applications, memory access latency can be reduced by setting the Maximum Memory Bus Frequency to 2933 MT/s or 2667 MT/s, in order to synchronize with the Infinity Fabric clock.

Intra-Socket Interconnect

The Infinity Fabric, evolved from AMD's previous generation HyperTransport interconnect, is a software-defined, scalable, coherent, and high-performance fabric. It uses sensors embedded in each die to scale control (Scalable Control Fabric, or SCF) and data flow (Scalable Data Fabric, or SDF).

- The SCF uses sensors to monitor die temperature, speed, and voltage across all cores within the dies and controls power management, security, reset, etc.
- The SDF connects the L3 caches to memory and to the configurable I/O lanes. SDF uses the configurable I/O lanes for memory-coherent communications between compute elements on a single die, between different dies on a socket, and between sockets in a node.
- The die-to-die Infinity Fabric bandwidth is 32 bytes for read and 16 bytes for write per Infinity Fabric clock (which has a maximum speed of 1,467 MHz).

Inter-Socket Interconnect

Two EPYC 7742 SoCs are interconnected via Socket to Socket Global Memory Interconnect (xGMI) links, part of the Infinity Fabric that connects all the components of the SoC together. In each Rome node configured with the HPE Apollo 9000 system architecture, there are 3 xGMI links using a total of 48 PCIe lanes. With the xGMI link speed set at 16 GT/s, the theoretical throughput for each direction is 96 GB/s (3 links x 16 GT/s x 2 bytes/transfer) without factoring in the encoding for xGMI, since there is no publication from AMD available. However, the expected efficiencies are 66-75%, so the sustained bandwidth per direction will be 63.5-72 GB/s.

Note: The xGMI link speed and width can be adjusted via BIOS setting. The xGMI Link Max Speed can be set to 10.667, 13, 16 or 18 GT/s. Setting it to a lower speed can save uncore power that can be used to increase core frequency or reduce overall power. It will also decrease cross-socket bandwidth and increase cross-socket latency. xGMI Dynamic Link Width Management saves power during periods of low socket-to-socket data traffic by reducing the number of active xGMI lanes per link from 16 to 8.

Inter-Node Network

There are 2,008 Rome nodes in the Aitken cluster. They are partitioned as follows:

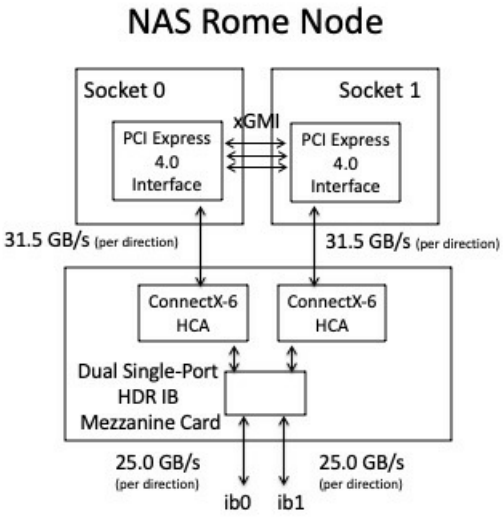
- 16 racks in the cluster (2,048 nodes total)
- 4 enclosures per rack (128 nodes/rack)
- 8 trays (HPE XL925g quad-node tray) per enclosure (32 nodes/enclosure)
- 4 nodes per compute tray

The hostname of each node is r[x1]c[x2]t[x3]n[x4], where x1 = 201 to 216; x2= 1 to 4; x3=1 to 8; and x4 =1 to 4.

The inter-node connection of Aitken's Rome nodes relies on the use of PCIe 4.0 lanes and one dual single-port IB HDR 200 Gbit/s mezzanine card on each node, and two IB HDR 200 Gbit/s premium switches per enclosure. One switch per enclosure facilitates the ib0 fabric, while the other facilitates the ib1 fabric.

As shown in the top diagram of this page, of the 128 PCIe 4.0 lanes per socket, 16 lanes are used to connect a node outwards. With a transfer rate of 16 GT/s, this enables a maximum bandwidth of 31.5 GB/s (when factoring in the 128/130 encoding) for each direction.

The two PCIe 4.0 x16 slots (one from each socket) are connected to the dual single-port IB HDR 200 Gbit/s mezzanine card, as shown below. The mezzanine card contains two separate Mellanox ConnectX-6 Host Channel Adapters (HCA), one for ib0 and the other for ib1, with a bandwidth of 25.0 GB/s per direction and a sub-microsecond latency. Note that the Mellanox ConnectX-6 200-Gbp/s adapters can achieve their maximum bandwidth (25.0 GB/s) in a PCIe 4.0 x16 slot (31.5 GB/s) but not in a PCIe 3.0 x16 slot (15.75 GB/s, as used in NAS Intel Xeon processors).



The 32 nodes in each enclosure are connected to one HPE Apollo 9000 IB HDR 40 Up premium switch (200 Gbit/s) for ib0, and to another for ib1. In each switch, there are two 40-port ConnectX-6 Application-Specific Integrated Circuit (ASIC) chips—80 ports total, of which 32 are used as downlinks to connect to the 32 nodes in the enclosure, 40 are used as uplinks to connect to external targets, and four are used internally to connect between the two ASICs.

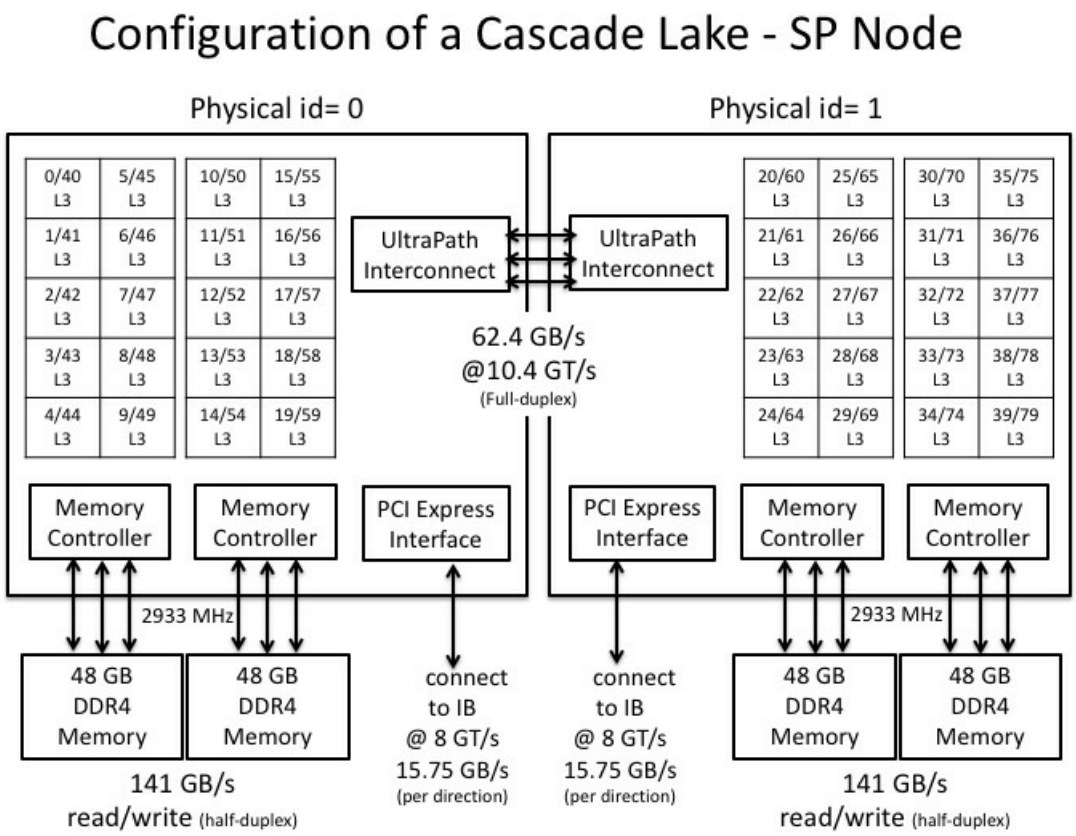
For each IB fabric—with 32 nodes in an enclosure connecting to the same IB switch, forming the first-dimension in the topology—the 2,048 Rome nodes form a 7-dimensional enhanced hypercube.

Connection Between Rome and Cascade Lake Nodes

The Rome and Cascade Lake nodes in Aitken will be connected with additional HDR cables on the 8th dimension.

References

- AMD EPYC 7742
- AMD EPYC 7742 specifications
- Compiler Options Quick Ref Guide for AMD EPYC 7xx2 Series Processors (PDF)
- High Performance Computing: Tuning Guide for AMD EPYC 7002 Series Processors (PDF)
- Workload Tuning Guide for AMD EPYC 7002 Series Processor Based Servers (PDF)
- Software Optimization Guide for AMD Family 17h Models 30h and Greater Processors
- AMD Infinity Architecture



Notes:

- Each small square in the diagram represents a combination of a physical core and a L3 cache slice. For each physical core, there are two logical core labelings obtained from the "processor" entry of the cat /proc/cpuinfo output.
- Although Cascade Lake processors support up to three Intel Ultra Path Interconnect (UPI) links with a bandwidth of 62.4 GB/s, the HPE 8600 Saxon board used for the Cascade Lake nodes at NAS implements only two links, with a bandwidth of 41.6 GB/s.

The Intel Cascade Lake processor incorporated into the Aitken cluster is the 20-core Xeon Gold 6248 model. Its base clock speed is 2.5 GHz for non-AVX, 1.9 GHz for AVX2, and 1.6 GHz for AVX-512. It uses an enhanced 14-nanometer (nm) fabrication process and the Skylake microarchitecture with some optimization.

Each Aitken Cascade Lake node contains two Cascade Lake processors and uses a dual single-port 100 Gbits/s Enhanced Data Rate (EDR) Mezzanine card to connect outward (see Inter-node Network). The use of four 200 Gbits/s High Data Rate (HDR) switches per enclosure forms the MPI and I/O InfiniBand fabrics within the Aitken cluster. Connecting the Aitken cluster to the NAS filesystems relies on additional sets of HDR switches and cables.

Instruction Sets

In addition to the instruction sets SSE, SSE2, SSE3, Supplemental SSE3, SSE4.1, SSE4.2, AVX, AVX2, AVX-512, and AVX512[F,CD,BW,DQ,VL], which are available in its Skylake predecessor, Cascade Lake also includes the new AVX-512 Vector Neural Network Instructions (VNNI), which provide significant, more efficient deep-learning inference acceleration. Cascade Lake also introduces in-hardware mitigations for the Spectre and Meltdown security flaws.

With 512-bit floating-point vector registers and two floating-point functional units, each capable of Fused Multiply-Add (FMA), a Cascade Lake core can deliver 32 double-precision floating-point operations per cycle.

Use the Intel compiler flag -xCORE-AVX512 for Skylake and Cascade Lake-SP specific optimizations. The optimization flag -qopt-zmm-usage=high -xCORE-AVX512 may benefit floating-point heavy applications running on Skylake and Cascade Lake.

Tip: If you want a single executable that will run on any of the Aitken and Electra processor types, with suitable optimization to be determined at run time, you can compile your application using the option:
-O3 -ipo -axCORE-AVX512,CORE-AVX2 -xAVX.

Hyperthreading

Hyperthreading is turned ON.

Turbo Boost

Turbo Boost is turned ON. Maximum Turbo Frequency is 3.90 GHz for non-AVX, and 3.8 GHz for AVX2 and AVX-512.

Cache Hierarchy

The cache hierarchy of Cascade Lake is as follows:

- L1 instruction cache: 32 KB, private to each core; 64 sets; 64 B/line; 8-way
- L1 data cache: 32 KB, private to each core; 64 sets; 64 B/line; 8-way; fastest latency: 4 cycles; 128 B/cycle load bandwidth; 64 B/cycle store bandwidth; write-back policy
- L2 cache: 1 MB, private to each core; 64 B/line; 16-way; fastest latency: 14 cycles; 64 B/cycle bandwidth to L1 cache; write-back policy
- L3 cache: shared non-inclusive 1.375 MB/core; total of 27.5 MB, shared by 20 cores in each socket; 2,048 sets; 64 B/line; fully associative; fastest latency: 50 - 70 cycles; write-back policy

Memory Subsystem

Like Skylake, there are two sub-NUMA clusters in each Cascade Lake socket, creating two localization domains. There are three memory channels per sub-NUMA cluster. Each channel can be connected with up to two memory DIMMs. For the Aitken Cascade Lake configuration, there is one 16-gigabyte (GB) dual rank DDR4 DIMM with error correcting code (ECC) support per channel. In total, the amount of memory is 48 GB per sub-NUMA cluster, 96 GB per socket, and 192 GB per node.

The speed of each memory channel is increased from 2,666 MHz in Skylake to 2,933 MHz in Cascade Lake. An 8-byte read or write can take place per cycle per channel. With a total of six memory channels, the total half-duplex memory bandwidth is approximately 141 GB/s per socket.

On-chip Interconnect

The on-chip architecture of Cascade Lake SP uses the same mesh layout as that of Skylake—where the cores and L3 caches are organized in rows and columns—instead of the ring architecture used by earlier Xeon processors.

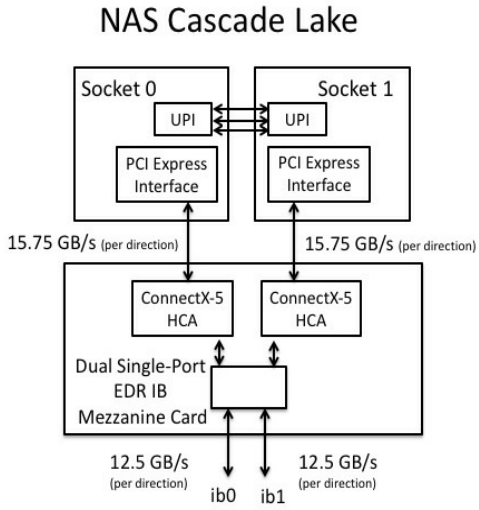
Inter-socket Interconnect

The Cascade Lake processors support up to three UPI links. The configuration at NAS has two links enabled. The UPI runs at a speed of 10.4 gigatransfers per second (GT/s). Each link contains separate lanes for the two directions. The total full-duplex bandwidth is 62.4 GB/s with three links and 41.6 GB/s with two links.

Inter-node Network

Like the network subsystems of the Electra Broadwell nodes and the Electra Skylake nodes, each Cascade Lake node uses two PCI Express interfaces (one from each socket) to connect to the Aitken InfiniBand (IB) fabrics. The use of PCIe 3.0 16-lane (x16) links enables a maximum bandwidth of 15.75 GB/s for each direction.

Like the Electra Skylake nodes, the Aitken Cascade Lake inter-node network makes use of the 100 Gbits/s EDR technology. As shown below, one PCIe is connected to the ib0 fabric via a single-port, four-lane (4X) EDR host channel adapter (HCA), in a dual single-port EDR IB mezzanine card, with an effective bandwidth of 100 Gbits/s (that is, 12.5 GB/s) for each direction. The other PCIe is connected to the ib1 fabric via another single-port 4x EDR HCA in the same mezzanine card.



There are 1,152 Cascade Lake nodes in the Aitken cluster, which uses the liquid-cooled HPE 8600 system architecture. The nodes are partitioned as follows:

- Four E-cells (288 nodes per E-cell)
- Two compute racks and one cooling rack per E-cell (144 nodes per compute rack)
- Four enclosures (IRUs) per compute rack (36 nodes per enclosure)
- Nine compute trays (HPE XA730i "Saxon" blades) per enclosure (four nodes per tray)

The connection of the compute nodes relies on the use of four standard IB HDR switch blades per enclosure. Two of these switch blades facilitate the ib0 fabric, while the other two facilitate the ib1 fabric. Each switch blade has one 40-port ConnectX-6 ASIC with 18 ports connecting to the compute nodes, and 22 ports available for switch-to-switch links and connections to external targets. The connection of the 1,152 Cascade Lake nodes forms a 6-dimensional enhanced hypercube.

Notes:

- There are eight ASICs per enclosure in Electra Skylake, while there are four in Aitken Cascade Lake. This reduces a 1,152-node topology from seven dimensions in Electra Skylake to six dimensions in Aitken Cascade Lake.
- Only two out of the three UPI links are enabled.

References

- Intel Xeon Gold 6248 Processor
- Second Generation Intel Xeon Scalable Processors
- Intel 64 and IA-32 Architectures Optimization Reference Manual (Apr 2019 edition)
- HPE SGI 8600 QuickSpecs (Apr 15, 2019 edition)

To help you prepare for running jobs on Aitken's Milan nodes, this short user guide includes information on the general configuration of Milan nodes, compiling your code, and running PBS jobs.

Aitken Milan Front-End Systems

Two new Aitken front ends, afe01 and afe02, serve the Aitken Milan nodes. Each AFE includes one or two AMD EPYC 7763 processors with a total of 128 physical cores (sum of two sockets), and 512 gigabytes (GB) of memory.

Note: The Aitken Milan compute nodes are not accessible from the PFEs.

You can access the AFEs using one of the following methods.

Initiating SSH from Your Local Desktop (Recommended)

- Two-step login:

  ```
  local_desktop% ssh sfe6.nas.nasa.gov  (or replace sfe6 with sfe7 or sfe8)
  sfe6% ssh afe01    (or ssh afe02)
  ```

- One-step login (using SSH passthrough):

  Before you begin: Follow the instructions in Setting Up SSH Passthrough, and make sure that the configuration file in the .ssh directory of your local desktop contains the following lines:

  ```
  Host afe01
     HostKeyAlias          afe01.nas.nasa.gov
     ProxyCommand          ssh -ax -oCompression=no sfe ssh-balance %h

  Host afe02
     HostKeyAlias          afe02.nas.nasa.gov
     ProxyCommand          ssh -ax -oCompression=no sfe ssh-balance %h
  ```

  Once you have SSH Passthrough set up, you can log into the AFE in one step:

  ```
  local_desktop% ssh afe01   (or ssh afe02)
  ```

Initiating SSH from a PFE

Note: This method is not recommended but is made available for convenience.

```
pfe% ssh afe01 (or ssh afe02)
```

Aitken Milan Compute Nodes

Aitken includes 512 Milan compute nodes, which are partitioned into four physical racks. Each node contains two 64-core AMD EPYC 7763 Milan sockets (2.45 GHz) and 512 GB of memory.

These nodes are inter-connected with four-lane High Data Rate (4X HDR) devices and four-lane High Data Rate (4X HDR) switches. They are currently housed in the NAS main building and have access to the NAS compute resources home and Lustre filesystems.

Note: For the long term, the Milan nodes will be relocated to be near the existing Aitken resources when a new compute module becomes available. At that time, these nodes will be connected to the Rome nodes with additional HDR cables on the 8th dimension of the Aitken cluster hybercube.

Operating System and Software Modules

The operating system on all Milan compute nodes is the Red Hat Enterprise Linux-based Tri-Lab Operating System Stack Version 4 (TOSS 4). Batch jobs use :aoe=toss4 by default, so you do not need to add :aoe=toss4 in your PBS resource request. The default $MODULEPATH includes:

- /usr/share/Modules/modulefiles
- /nasa/modulefiles/toss4
- /nasa/modulefiles/spack/
- /nasa/modulefiles/pkgsrc/toss4/

Compilers For Milan Nodes

There are several compilers you can use to build your application:

Intel Compilers

Some Intel compiler modules are available in the /nasa/modulefiles/toss4 directory, however, several are in the /nasa/modulefiles/testing directory. To see what versions are available in these two directories, run:

```
module use -a /nasa/modulefiles/testing
module avail comp-intel
```

To build your application using the Intel compilers, use `-march=core-avx2`.

Existing executables that are built for the Aitken Rome nodes should run on the Milan nodes.

Note: Although AMD recommends using `-axCORE-AVX2`, we do not recommend using it, as the resulting executable will run but will not actually take advantage of AVX2 instructions. Executables built with `-xCORE-AVX2`, `-xSSE4.2`, and so on will not run on Milan processors.

GNU Compilers

The default version of GNU Compiler Collection (GCC) under TOSS 4 (/usr/bin/gcc) is GCC 8.5.0. However, support for the Rome architecture is only available with GCC 11.xx and later. GCC version 13.2 has been installed and is accessible via:

```
module use -a /nasa/modulefiles/testing
module load gcc/13.2
```

The recommended compiler flag for Milan with gcc is: `-march=znver3`

AMD Optimizing C/C++ Compilers (AOCC)

To see available versions before loading an AOCC compiler, use the `module avail comp-aocc` command. For example:

```
module use -a /nasa/modulefiles/testing
module avail comp-aocc
---------- /nasa/modulefiles/testing ----------
comp-aocc/4.0.0  comp-aocc/4.1.0  cmp-aocc/4.2.0

---------- /nasa/modulefiles/toss4 ------------
comp-aocc/3.2.0
```

The AOCC C, C++ and Fortran compiler drivers are clang, clang++ and flang, respectively.

The recommended compiler flag for Milan with AOCC is: `-march=znver3`

Note: The default optimization level is `-O1` in AOCC 3.2, and `-O2` in AOCC 4.1 and newer versions.

For more compiler options, see AMD's article Compiler Options Quick Reference Guide for AMD EPYC 7xx3 Series Processor.

MPI Libraries For Milan Nodes

HPE MPT

Use the modulefile `mpi-hpe/mpt`, which is available in the /nasa/modulefiles/toss4 directory:

```
module load mpi-hpe/mpt
```

OpenMPI

There are a few OpenMPI modulefiles available in the /nasa/modulefiles/testing directory.

```
module use -a /nasa/modulefiles/testing
module avail openmpi
---- /nasa/modulefiles/testing ----
openmpi/4.1.5-toss4-intel_2022.1  openmpi/4.1.6-toss4-gnu  openmpi/4.1.6-toss4-aocc
```

Note: There are several newer versions that are not listed in the output above.

The compiler used to build each of these modulefiles is indicated in their filenames. For example, `openmpi/4.1.6-toss4-gnu` was built with gcc 8.5.0, and `openmpi/4.1.6-toss4-aocc` was built with aocc 4.2.0.

After loading one of these modulefiles, run the command `ompi_info` to learn more about each installation.

Running OpenMP or MPI/OpenMPI Hybrid Codes

Each Milan node has 128 physical cores. When simultaneous multi-threading (SMT) is enabled in BIOS, there are 256 logical cores for use by OpenMP or MPI/OpenMPI hybrid codes.

To check whether SMT is enabled, run the following command on a Milan compute node:

```
cat /proc/cpuinfo | grep processor | wc -l
```

If the output shows 256, then SMT is enabled.

Note: During Milan's early production, some nodes were configured with the SMT on and some with the SMT off. Currently, all nodes have been set with SMT on.

We recommend using the mbind.x tool for process and thread binding, as shown in the following example with four MPI processes and 32 OpenMP threads per process:

```
mpiexec -n 4 /u/scicon/tools/bin/mbind.x -t 32 a.out
```

If you use HPE MPT, you can also use the omplace tool for pinning MPI/OpenMP hybrid code.

Running PBS Jobs on Aitken Milan Nodes

Aitken's Milan nodes are accessible only through PBS jobs submitted from afe01 and afe02 and served by the PBS server pbspl1.

To request Aitken Milan nodes, use :model=mil_ait in your PBS script, as shown in the example below.

Sample PBS Script For Aitken Milan Nodes

```
#PBS -l select=2:ncpus=128:mpiprocs=128:model=mil_ait
#PBS -l walltime=8:00:00
#PBS -q normal
module use -a /nasa/modulefiles/testing
module load mpi-hpe/mpt
module load comp-intel
cd $PBS_O_WORKDIR
mpiexec -np 256 ./a.out
```

For more information, see AMD Milan Processors.

To help you prepare for running jobs on Aitken's Rome nodes, this short user guide includes information on the general configuration of Rome nodes, compiling your code, and running PBS jobs.

Overview of Aitken Rome Nodes

Aitken includes 2,008 Rome nodes, which are partitioned into sixteen physical racks. Each node contains two 64-core AMD EPYC 7742 Rome sockets (2.25 GHz) and 512 GB of memory. The nodes are connected to the Aitken InfiniBand network (ib0 and ib1) via four-lane High Data Rate (4X HDR) devices and four-lane High Data Rate (4X HDR) switches for internode communication. The ib1 fabric is used mainly for I/O and is connected to the Lustre filesystems. In addition, Aitken Rome nodes share the same home filesystems, PFEs, and PBS server as Electra, and the Aitken Cascade Lake and Milan nodes. You can access Aitken Rome nodes only through PBS jobs submitted from the PFEs.

Note: Aitken Milan nodes use a different set of front ends. See Preparing to Run on Aitken Milan Nodes for more information.

Operating System and Software Modules

The operating system on all Rome compute nodes is the Red Hat Enterprise Linux-based Tri-Lab Operating System Stack Version 4 (TOSS 4). Batch jobs use :aoe=toss4 by default, so you do not need to add :aoe=toss4 in your PBS resource request. The default $MODULEPATH includes:

- /usr/share/Modules/modulefiles
- /nasa/modulefiles/toss4
- /nasa/modulefiles/spack/
- /nasa/modulefiles/pkgsrc/toss4/

Compilers for Rome Nodes

Note: Some executables built under the TOSS 3 environment will work under TOSS 4; however, some may not. If you need to rebuild your executable under TOSS 4, and you experience issues such as missing certain software packages or libraries, send an email to support@nas.nasa.gov

There are several compilers you can use to build your application:

- Intel Compilers:

  Some Intel compiler modules are available in the /nasa/modulefiles/toss4 directory, however, several are in the /nasa/modulefiles/testing directory. To see what versions are available in these two directories, run:

  ```
  module use -a /nasa/modulefiles/testing
  module avail comp-intel
  ```

  Existing executables that are built with support for AVX512 (for running on Skylake and Cascade Lake processors) will not run on the Rome nodes. Others that are built with support for AVX2 may run on the Rome nodes; however, you should always check the result of your runs using existing executables. To rebuild your application using Intel compilers, use -march=core-avx2.

  Note: Although AMD recommends using -axCORE-AVX2, we do *not* recommend using it, as the resulting executable will run but will not actually take advantage of AVX2 instructions. Executables built with -xCORE-AVX2, -xSSE4.2, etc., will not run on Rome processors.
- GNU Compilers:

  The default version of GNU Compiler Collection (GCC) under TOSS 4 (/usr/bin/gcc) is GCC 8.5.0. However, support for the Rome architecture is only available with GCC 9.x and later. Among the several GCC versions available in the /nasa/pkgsrc/toss4 directory, load the 9.3 or 10.3 version for use with the Rome nodes. For example:

  ```
  module avail gcc
  -- /nasa/modulefiles/pkgsrc/toss4/ --
  gcc/6.5  gcc/7.5  gcc/8.4  gcc/9.3  gcc/10.3  gcc/13.2
  module load gcc/10.3
  ```

  The recommended compiler flag to use with GCC is -march=znver2.
- AMD Optimizing C/C++ and Fortran Compilers (AOCC)

  To see available versions before loading an AOCC compiler, use the module avail comp-aocc command. For example:

  ```
  module avail comp-aocc
  -- /nasa/modulefiles/toss4 --
  comp-aocc/3.2.0

  -- /nasa/modulefiles/testing --
  comp-aocc/4.0.0  comp-aocc/4.1.0  comp-aocc/4.2.0  comp-aocc/5.0.0

  module load comp-aocc/3.2.0
  ```

  The recommended compiler flag to use with AOCC is -march=znver2.

- PGI Compilers:

  Some PGI compiler modules are available in the /nasa/modulefiles/testing directory. There is no support specifically for Rome processors. If you want to compile your code using a PGI compiler, the recommended compiler flag is -tp=zen

For more compiler options, see Compiler Options Quick Reference Guide for AMD EPYC 7xx2 Series Processors.

MPI Library for Rome Nodes

HPE MPT is the only recommended MPI library for use on the AMD Rome nodes under TOSS 4. Use thempi-hpe/mpt.2.30 modulefile in the /nasa/modulefiles/toss4 directory. Earlier versions such as mpi-hpe/mpt.2.25 do not work properly for the Rome nodes.

```
module load mpi-hpe/mpt.2.30
```

Note: Although there is more than one version of MPT library installed on NAS systems, mpi-hpe/mpt.2.30 is currently the only reliable version for running on the Rome nodes. You can access this version by loading mpi-hpe/mpt, which points to the recommended version (currently mpi-hpe/mpt.2.30) and environmental settings:

```
% module load mpi-hpe/mpt
```

Running OpenMP or MPI/OpenMPI Hybrid Codes

Each Rome node has 128 physical cores. When Simultaneous Multi-Threading (SMT) is enabled in BIOS, there are 256 logical cores for use by OpenMP or MPI/OpenMPI hybrid codes.

To check whether SMT is enabled, run the following command on a Rome compute node:

```
cat /proc/cpuinfo | grep processor | wc -l
```

If it shows 256, then SMT is enabled.

We recommend using the mbind.x tool for process and thread binding, as shown in the following example with four MPI processes and 32 OpenMP threads per process:

```
mpiexec -n 4 /u/scicon/tools/bin/mbind.x -t 32 a.out
```

If you use HPE MPT, you can also use the omplace tool for pinning MPI/OpenMP hybrid code.

Running PBS Jobs on Aitken Rome Nodes

To request Aitken Rome nodes, use :model=rom_ait in your PBS script, as shown in the example below.

The devel, normal, long, debug, and low queues are available for use with Rome nodes.

Note: The Standard Billing Unit (SBU) rate for Rome processors is 4.06.

Sample PBS Script For Aitken Rome Nodes

```
#PBS -l select=2:ncpus=128:mpiprocs=128:model=rom_ait
#PBS -l walltime=8:00:00
module load mpi-hpe/mpt
module load comp-intel/2020.4.304

cd $PBS_O_WORKDIR

mpiexec -np 256 ./a.out
```

For more information, see AMD Rome Processors.

To help you prepare for running jobs on Aitken's Cascade Lake compute nodes, this short user guide includes information on the general configuration of Cascade Lake nodes, compiling your code, running PBS jobs, and checking allocation usage.

Overview of Aitken Cascade Lake Nodes

Aitken includes 1,152 Cascade Lake nodes, which are partitioned into eight physical racks. Each node contains two 20-core Xeon Gold 6248 sockets (2.5 GHz) and 192 GB of memory. The nodes are connected to the Aitken InfiniBand network (ib0 and ib1) via four-lane Enhanced Data Rate (4X EDR) devices and four-lane High Data Rate (4X HDR) switches for internode communication. The ib1 fabric is used mainly for I/O and is connected to the NAS Lustre filesystems. In addition, Aitken Cascade Lake nodes use the NAS compute resources home filesystems, front-end systems (PFEs), and PBS server (pbspl1) as Electra and the Aitken Rome nodes. You can access Aitken Cascade Lake nodes only through PBS jobs submitted from the PFEs.

Note: Aitken Milan nodes use a different set of front ends. See Preparing to Run on Aitken Milan Nodes for more information.

Compiling Your Code For Cascade Lake Nodes

The Cascade Lake processors include the Advanced Vector Extensions 512 (AVX-512). Intel AVX-512 optimizations are included in Intel compiler version 16.0 and later versions. We recommend that you test the latest Intel compiler, through the command `module load comp-intel/`, which may include better optimizations for AVX-512.

For Cascade Lake-specific optimizations, use the compiler option `-xCORE-AVX512`.

If you want a single executable that will run on any of the Aitken and Electra processor types, with suitable optimization to be determined at runtime, you can compile your application using the option:
`-O3 -axCORE-AVX512,CORE-AVX2 -xAVX`.

Note: The use of either of these options, `-xCORE-AVX512` or `-axCORE-AVX512`, could either improve or degrade performance of your code. Be sure to check performance with and without these flags before using them for production runs.

Running PBS Jobs on Aitken Cascade Lake Nodes

To request Aitken Cascade Lake nodes, use `:model=cas_ait` in your PBS script, as shown in the example below.

Note: Because MPT 2.15 and earlier versions do not support the ConnectX-5 host channel adapters (HCAs), the environment variables MPI_IB_XRC and MPI_XPMEM_ENABLED have been disabled for jobs running on Cascade Lake. If your MPI applications perform significant MPI collective operations and rely on having these two variables enabled to get good performance, use MPT 2.17 or newer versions. You can use the NAS-recommended MPT version with the command:
`module load mpi-hpe/mpt`

Sample PBS Script For Aitken Cascade Lake Nodes

```
#PBS -l select=10:ncpus=40:mpiprocs=40:model=cas_ait
#PBS -l walltime=8:00:00
#PBS -q normal

module load mpi-hpe/mpt
module load comp-intel

cd $PBS_O_WORKDIR

mpiexec -np 400 ./a.out
```

For more information, see Cascade Lake Processors.