Endeavour3 and Endeavour4 are two HPE Superdome Flex servers. They replace Endeavour1 and Endeavour2 as the NAS resources for supporting applications that need access to large cache-coherent, global shared-memory capabilities in a single system image (SSI).

Each of these two systems has 896 cores, six terabytes (TB) of memory, and a theoretical double-precision floating point peak performance of 77.414 teraflops. Each system is configured with a total of 32 sockets in this hierarchy:

- 28 cores and 192 gigabytes (GB) of memory in one socket.
- Four sockets in one chassis.
- Eight chassis in one system.

Processors and Memory Configuration in a Socket

At the processor level, Endeavour3 and 4 are based on the Intel Xeon Platinum 8280 processor, which is similar to Aitken'sIntel Xeon Gold 6248 Cascade Lake processors.

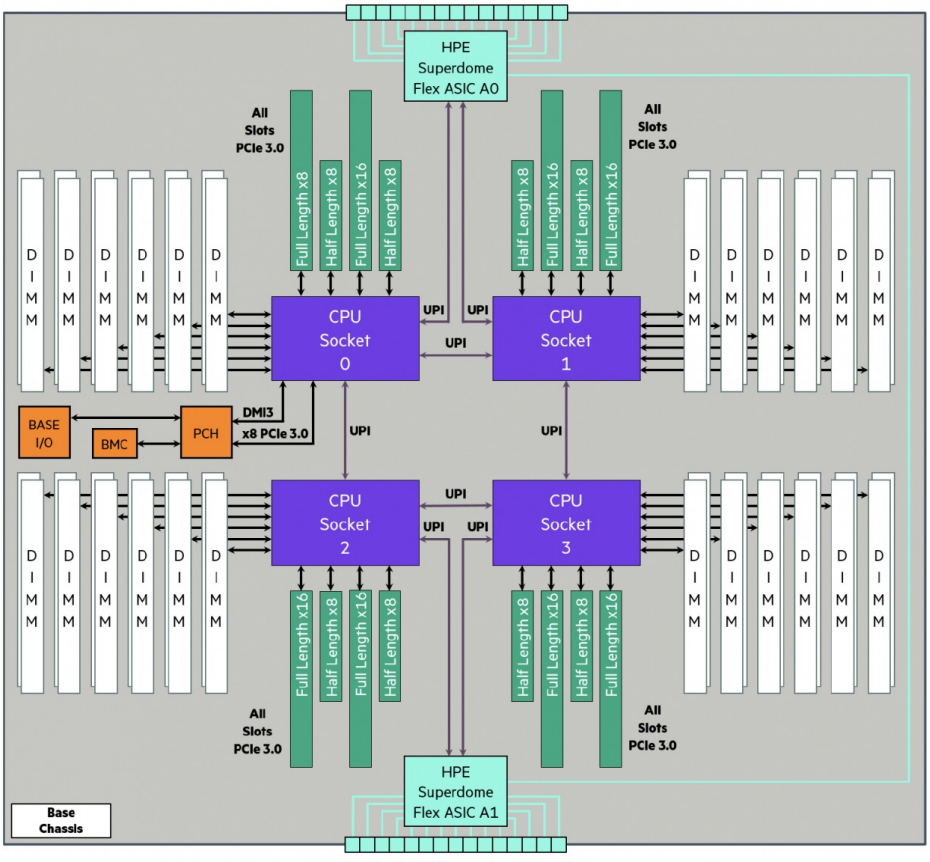
The main improvements provided by the Platinum 8280 processor are:

- The Endeavour systems have more cores per socket: 28 cores vs. 20 cores.
- The CPU clock speed is higher: 2.70 GHz vs. 2.50 GHz.
- The L3 cache size per socket is greater: 38.5 megabytes (MB) vs. 27.5 MB.
- The memory size per socket is increased: 192 GB vs. 96 GB.
- There is more DDR4 memory DIMM per memory channel: 32 GB vs. 16 GB.

For instruction sets, cache hierarchy, and memory subsystem information, see Preparing to Run on Endeavour.

Chassis Configuration

As shown in the following base chassis diagram, published by HPE, there are four sockets connected in a ring fashion via 10.4 GT/s Intel Ultra Path Interconnect (UPI) links. Each socket is also connected to one of two custom-designed HPE Superdome Flex ASICs, via an Intel UPI link, to send remote targeted cache-coherent data traffic to the external chassis.

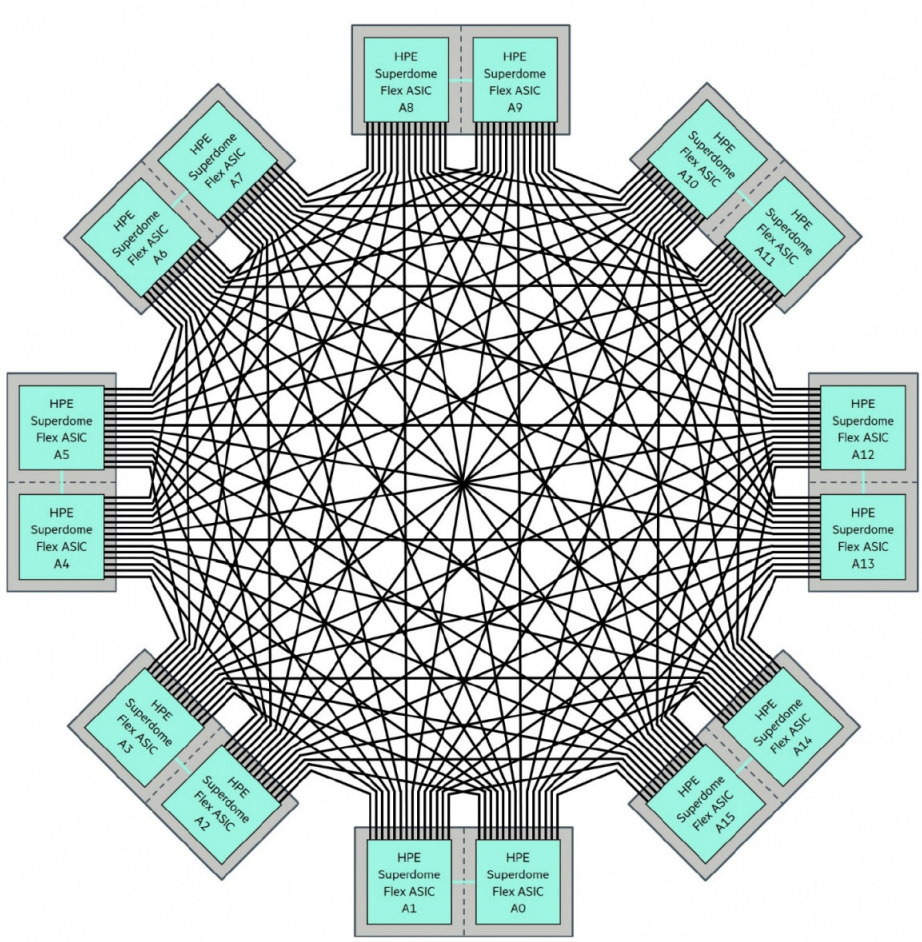


Memory latency within each socket is approximately 100 nanoseconds (ns). However, due to sub-NUMA design within a socket (see the Cascade Lake configuration diagram), it will be faster if the CPU that a process runs on is in the same sub-domain as the memory where the data is located. Memory latency between two sockets through a UPI link is approximately 130 ns.

All memory channels in each chassis are fully independent, and they can run simultaneously at DRAM data transfer rates up to 2,933 MegaTransfers per second (MT/s). This provides a theoretical total memory bandwidth of 564 GB/s; (memory bandwidth per socket is 141 GB/s). HPE's published STREAM TRIAD memory bandwidth is more than 360 GB/s per chassis.

System Interconnect

The HPE Superdome Flex server has a modular architecture based on a four-socket chassis. As shown in the following diagram published by HPE, the eight chassis of a 32-socket server are connected through a cabled crossbar interconnection fabric called *HPE Superdome Flex Grid*.



Each of the 16 application-specific integrated circuits (ASICs) in the grid has a very large directory cache to track the cache line state and maintain coherency across all attached processor sockets. A Superdome Flex ASIC (also known as NUMAlink 8) provides 16 *flex grid* ports, each capable of 13.3 GB/s data rates for maximum flex grid bandwidth. The total bi-sectional crossbar grid bandwidth for a 32-socket Superdome Flex server is more than 850 GB/s. With this crossbar interconnection, memory latency is minimized (< 400 ns), since there is only one hop between any two ASICs.

The HPE Superdome Flex Grid provides adaptive routing features designed specifically to enhance performance by routing traffic through the optimal latency path available, and provides superior uptime by automatically routing traffic around failed components.

Processor, Memory, and Network Subsystems Configuration

Detailed configuration statistics for Endeavour 3 and Endeavour 4 are listed below:

Processor, Memory, and Network Subsystems Statistics

Architecture	HPE Superdome Flex
CPU	Cascade Lake 28-Core Xeon 8280
CPU Base Clock	2.70 GHz
Hyperthreading	OFF
Turbo Boost	On (set to performance instead of powersave)
Maximum Double-Precision Flops per Cycle per Core	32
# of Cores/Socket	28
# of Sockets/Chassis	4
# of Chassis/System	8
Total # of cores/System	896
Total Double Precision TFlops/System	77.414
Memory/Core	6.85 GB

Memory/Socket	192 GB
Memory/System	6 TB
Memory Speed	2,933 MHz
System Interconnect	Superdome Flex ASICs (NUMAlink 8)
Topology	Crossbar

References

HPE Superdome Flex Server Architecture and RAS (PDF)

The Endeavour supercomputer comprises two nodes, Endeavour3 and Endeavour4. Each of the two systems includes 896 cores and six terabytes (TB) of global shared memory. See Endeavour Configuration Details for more hardware information.

Endeavour 3 & 4 are accessible through the PFEs and are served by PBS server pbspl4. They have access to the NAS compute resources home filesystem and NAS Lustre nobackup filesystems.

Update to Endeavour Support Model: Due to the age of Endeavour's hardware, we are lowering the priority of those nodes in our monthly patching work. This means that Endeavour may not always be returned to service at the same time as newer systems, and the patch windows may need to be extended to account for additional time needed to bring the nodes online.

Operating System

Endeavour 3 & 4 are running the Red Hat Enterprise Linux-based Tri-Lab Operating System Stack (TOSS4).

Connecting to Endeavour

To access the systems, submit PBS jobs from the PFEs with the server name pbspl4 included in the qsub command line or PBS script. You cannot log directly into Endeavour3 and 4 from the PFEs unless you have a PBS job running there.

Accessing Your Data

Use your NAS home filesystems (/u/*username*) and your Lustre filesystem (/nobackup) for Endeavour.

For more information on using Lustre filesystems, see Lustre Basics and Lustre Best Practices.

Compiling and Running Your Code

There are no default modules for compilers, Message Passing Interface (MPI) components, or math libraries. To use these modules, you must first load them.

OpenMP Applications

Load an Intel compiler module and use the Intel compiler -qopenmp option to build the executable. Since the Cascade Lake processors used in Endeavour3 and 4 support AVX512, you can experiment with different compiler versions and flags for better performance. For example:

```
% module load comp-intel/2020.0.166
% ifort -O2 -qopenmp program.f
```

or

```
% module load comp-intel/2020.4.304
% ifort -Ofast -ipo -axCORE-AVX512,CORE-AVX2,AVX -xSSE4.2 -qopenmp program.f
```

Set a specific number of OpenMP threads in your PBS job, and use a thread-pinning method to ensure that the threads do not get in the way of each other on the same core or migrate from one core to another. For example:

```
#PBS ...

module load comp-intel/2020.4.304
setenv OMP_NUM_THREADS 28

/u/scicon/tools/bin/mbind.x -cs -t28 -v ./a.out
```

or

```
setenv KMP_AFFINITY compact # or: setenv KMP_AFFINITY scatter
./a.out > output
```

or

```
setenv KMP_AFFINITY disabled
dplace -x2 ./a.out > output
```

See Process/Thread Pinning Overview for more information about different pinning methods.

Note: The default OMP_STACKSIZE is set to 200 megabytes (MB). If your application experiences segmentation fault (segfault), experiment with increasing the $OMP_STACKSIZE value to see if it helps.

Applications that Require Scientific and Math Libraries

The Intel Math Kernel Library (MKL) is included in Intel compiler modules (versions 11.1 and later). SeeMKL to learn how to link to

MKL libraries. In many cases, the following commands may be sufficient:

```
% module load comp-intel/2020.4.304
% ifort -O2 program.f -mkl
```

Note: By itself, the -mkl option implies -mkl=parallel, which will link to the threaded MKL library. If your application can benefit from using multiple OpenMP threads when the MKL routines are called, you should also set the environment variable OMP_NUM_THREADS, as shown in the OpenMP example in the previous section.

If you want to use a non-threaded MKL library, change the -mkl option to -mkl=sequential.

MPI Applications

HPE's latest Message Passing Toolkit (MPT) library is recommended. Load the library as shown in the following example:

```
% module load comp-intel/2020.4.304
% module load mpi-hpe/mpt
% ifort -O2 program.f -lmpi
```

During runtime, load the Intel compiler and the MPT modules and use mpiexec to launch the executable:

```
#PBS ...
module load comp-intel/2020.4.304
module load mpi-hpe/mpt

mpiexec -np xx ./a.out > output
```

Running PBS Jobs

Each of the Endeavour nodes has 32 sockets, and each socket includes 28 cores and 192 gigabytes (GB) of physical memory. In each socket, PBS has access to 28 cores and 185 GB of memory, which is known as a minimum allocatable unit (MAU). In addition, one of the sockets is set aside for system operations, leaving 31 sockets available for user jobs.

PBS does not allow jobs to share resources in the same socket. However, the amount of resources in a socket that a job will receive is restricted to the amount requested. This restriction is enforced by the Linux kernel cgroups feature. For example, if you request one core and 100 GB of memory, one socket will be assigned exclusively for your job, but 27 cores and 85 GB of memory of the socket will not be accessible by your job.

When job accounting is enabled, your job will be charged by the number of MAUs assigned exclusively for your job.

The maximum amount of resources a job can request is:

- **Endeavour3:** 868 cores with 5,750 GB are the maximum resources available for PBS jobs.
- **Endeavour4:** 868 cores with 5,750 GB are the maximum resources available for PBS jobs.

Endeavour PBS Queues

The PBS server pbspl4 manages jobs for Endeavour3 and 4 (:model=cas_end) and the NAS A100 GPU nodes (:model=[mil_a100]). Among the queues listed by the command qstat -Q @pbspl4, you can run jobs on the two systems using the e_normal, e_long, e_vlong, and e_debug queues. You can also find detailed settings of each queue, such as the maximum walltime and the default or minimum ncpus and memory, using the following command line (the e_normal queue is used in the example):

```
pfe% qstat -fQ e_normal@pbspl4
```

PBS Commands

Run PBS commands (such as qsub, qstat, and qdel, etc.) from the PFEs and specify the PBS server, pbspl4. For example:

```
pfe21% qsub -q queue_name@pbspl4 job_script
pfe21% qstat -nu username @pbspl4
pfe21% qstat job_id.pbspl4
pfe21% qdel job_id.pbspl4
```

If you do not normally run jobs on Aitken and Electra (which also use the PFEs to submit jobs), and your primary workload is on Endeavour3 and Endeavour4, you may want to consider setting the PBS_DEFAULT environment variable to pbspl4. This allows you to simplify the PBS commands, as follows:

```
pfe21% setenv PBS_DEFAULT pbspl4
pfe21% qsub job_script
pfe21% qstat -nu username
pfe21% qstat job_id
pfe21% qdel job_id
```

Job Submission Examples

The following examples demonstrate how PBS allocates resources for your job.

Note: If you do not specify a resource attribute (such as ncpus or mem), a default value for that attribute will be assigned to your job.

- To request using all the resources in two MAUs:

  pfe*21*% qsub -I -lselect=2:ncpus=28:mem=185G:model=cas_end -q *queue_name*@pbspl4

- To request 56 cores:

  pfe*21*% qsub -I -lselect=1:ncpus=56:model=cas_end -q *queue_name*@pbspl4

  Your job will be given access to 56 cores across in two sockets and only the default amount of memory (32G).

- To request 1,850 GB of memory:

  pfe*21*% qsub -I -lselect=1:mem=1850G:model=cas_end -q *queue_name*@pbspl4

  Your job will be given access to 1,850 GB of memory spread in ten sockets and only the default number of cores (one).

- If you do not specify which Endeavour node to run your job, PBS will assign one for you. The following example specifies Endeavour3 and requests 28 cores and 185 GB of memory:

  pfe*21*% qsub -I -lselect=host=endeavour3:ncpus=28:mem=185G:model=cas_end -q *queue_name*@pbspl4

Job Accounting

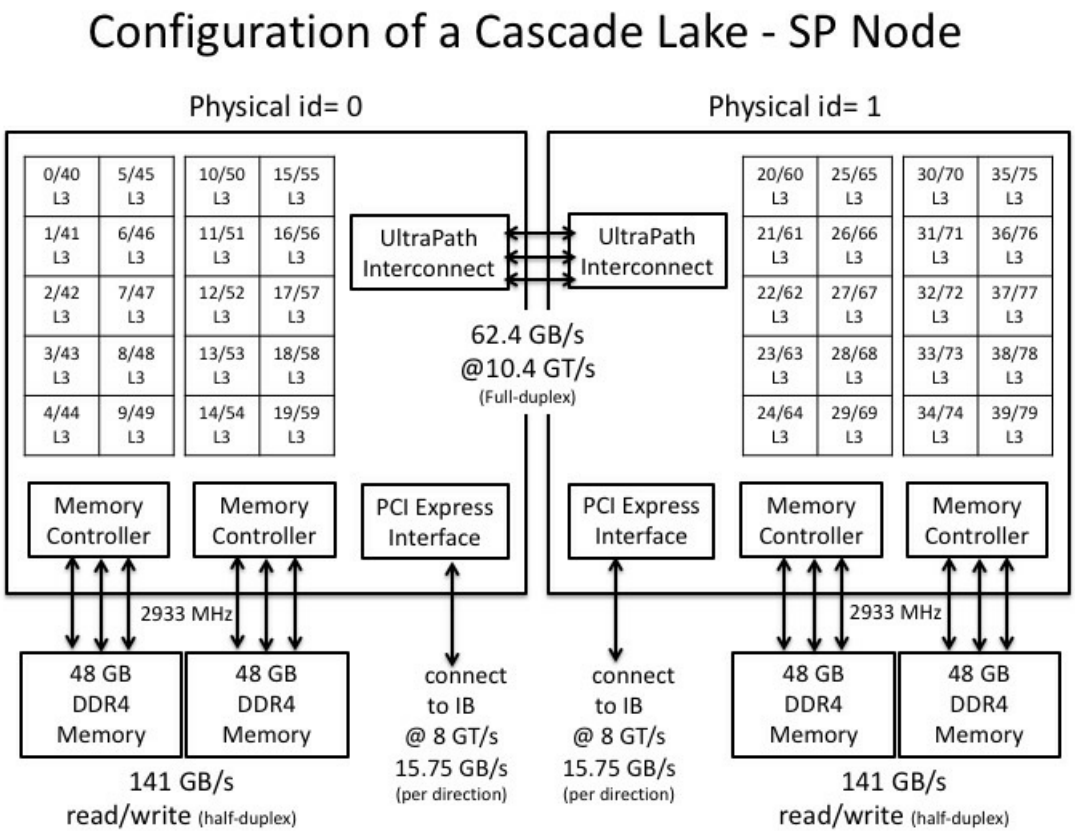
The Standard Billing Unit (SBU) rate for using one MAU per hour is 1.31.

Your group's HECC SBU allocation is charged for usage on any HECC resources, including Electra, Aitken, and Endeavour. The output from the acct_ytd command does not have separate entries for the different systems. To find Endeavour3 and Endeavour4 usage for your group, use the acct_query command instead.

For example, to view the statistics of all of your jobs that ran on Endeavour3 on March 10, 2021, type:

pfe*21*% acct_query -c endeavour3 -d 03/10/21 -u *your_username* -olow

See Job Accounting Utilities to learn more about using the acct_ytd and acct_query tools.



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.

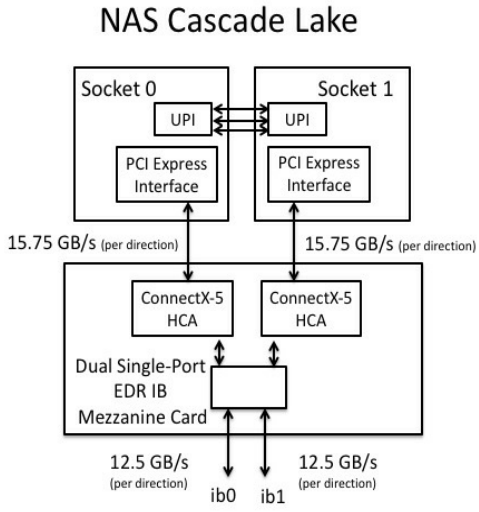
Inter-socket Interconnect

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)