Electra, NASA's first prototype modular supercomputing system, is housed in a module a short distance from the primary NASA Advanced Supercomputing (NAS) facility. Electra is configured as follows:

- 16 Broadwell racks and 8 Skylake E-cells
- 3,384 nodes
- 121,536 cores
- 576 terabytes total memory
- 8.10 petaflops theoretical peak performance

Hostnames

There are 4 individual rack units (IRUs) in each rack. For every two racks, there is a rack leader controlling them. The naming convention for the nodes residing in every two racks uses only odd rack numbers.

Broadwell
: 18 nodes per IRU, with 144 nodes residing in every two racks (2 racks x 4 IRUs x 18 nodes).
The hostnames for the Broadwell nodes are r[*x*]i[0-7]n[0-17], where *x* = odd numbers between 1 and 15.

Skylake
: 36 nodes per IRU, with 288 nodes residing in every two racks (2 racks x 4 IRUs x 36 nodes).
The hostnames for the Skylake nodes are r[*x*]i[0-7]n[0-35], where *x* = odd numbers between 133 and 147.

Processor, Memory and Network Subsystems Statistics

Electra's system architecture is ICE X. The following table provides detailed configuration statistics for the processor, memory, and network subsystems:

	Processor	
	Broadwell	**Skylake**
CPU	14-Core Xeon E5-2680v4	20-Core Xeon Gold 6148
Newest Instruction Set	AVX2	AVX-512
Hyperthreading	ON	ON
TurboBoost	ON	ON
CPU-Clock	2.4 GHz	2.4 GHz
Maximum Double Precision Floating Point Operations per Cycle per Core	16	32
# of Cores/node	28	40
Total # of Nodes	1,152	2,232
Total # of Cores	32,256	89,280
Total Double Precision TFlops	1,239	6,857
	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
L2 Cache	256 KB per core; Associativity: 8; Cache line size: 64B	1 MB per core; Associativity: 16; Cache line size: 64B
L3 Cache	35 MB shared inclusive by the 14 cores; Associativity: 20; Cache line size: 64B	27.5 MB shared non-inclusive by the 20 cores; Associativity: full; Cache line size: 64B
TLB	Local to each core	
Default Page Size	4 KB	4 KB
Memory/Core	4.6 GB; DDR4	4.8 GB; DDR4
Total Memory/node	128 GB;	192 GB
Memory Speed and Bandwidth	2400 MHz; 4 channels; 76.8 GB/sec read/write	2666 MHz; 6 channels; 128 GB/sec read/write
Intersocket Interconnect	QuickPath Interconnect 4.8 GHz, 9.6 GT/s, or 38.4 GB/sec	Ultra Path Interconnect 5.2 GHz, 10.4 GT/s, or 41.6 GB/sec
	Inter-node Network	
IB Device on node	Dual single-port 4x FDR IB Mezzanine card (2 single-port HCAs); 56 Gbits/s	Dual single-port 4x EDR IB Mezzanine card (2 single-port HCAs); 100 Gbits/s
IB Switches between nodes	4x FDR; 56 Gbits/s	4x EDR; 100 Gbits/s

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

Overview of Electra Skylake Nodes

Electra includes 2,196 Skylake nodes, which are partitioned into sixteen physical racks. Each node contains two 20-core Xeon Gold 6148 sockets (2.4 GHz) and 192 GB of memory. The nodes are connected to the Electra InfiniBand network (ib0 and ib1) via four-lane Enhanced Data Rate (4X EDR) devices and switches for internode communication. The ib1 fabric is used mainly for I/O and is connected to the NAS Lustre filesystems. In addition, Electra has access to the NAS compute resources home filesystems, front-end systems (PFEs), and PBS server (pbspl1). You can access Electra only through PBS jobs submitted from the PFEs.

Compiling Your Code For Skylake Nodes

The Skylake 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, which may include better optimizations for AVX-512.

For Skylake-specific optimizations, use the compiler option -xCORE-AVX512.

If you want a single executable that will run on any of the 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 Electra Skylake Nodes

To request Electra Skylake nodes, use :model=sky_ele in your PBS script, as shown in the example below.

Important: Because MPT 2.15 and earlier versions do not support the ConnectX-5 host channel adapters (HCAs), the environment variablesMPI_IB_XRC and MPI_XPMEM_ENABLED have been disabled for jobs running on Skylake. If your MPI applications perform significant MPI collective operations and rely on having these two variables enabled to get good performance, use MPT 2.16 or newer versions. To use the NAS-recommended MPT version, use the command:
module load mpi-hpe/mpt

Sample PBS Script For Electra Skylake Nodes

```
#PBS -l select=10:ncpus=40:mpiprocs=40:model=sky_ele
#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
```

Checking Allocation Usage For Electra Skylake Jobs

To track allocation usage for jobs running on Electra Skylake nodes, run:

```
acct_query -c electra_S
```

Examples:

- To track the total SBUs usage for one of your GIDs (for example, s0001) since 9/30/17:

  ```
  % acct_query -c electra_S -ps0001 -b 9/30/17
  ```

- To track the SBUs usage for each of your jobs run today:

  ```
  % acct_query -c electra_S -olow
  ```

For more information, see Skylake Processors.

To help you prepare for running jobs on Electra's Broadwell compute nodes, this short review includes information on the general configuration, compiling your code, running PBS jobs, and checking allocation usage.

Overview of Electra Broadwell Nodes

Electra is housed in a module outside of the primary NAS facility. The system has 16 Broadwell racks, each containing 72 nodes. Each of the 72 nodes contains two 14-core E5-2680v4 (2.4 GHz) processor chips and 128 GB of memory.

Electra's InfiniBand fabric (ib1) is used mainly for I/O and is connected to the NAS Lustre filesystems. In addition, Electra and all other NAS compute systems share the same home filesystems. It also share the front-end systems (PFE) and PBS server (pbspl1) with some NAS compute nodes such as sky_ele, sky_el, and rom_ait. You can access Electra only through PBS jobs submitted from the PFEs.

Usage of Electra Broadwell nodes is charged to your project's HECC Standard Billing Unit (SBU) allocation.

Compiling Your Code For Broadwell Nodes

Like the Haswell processors, the Broadwell processors support the Advanced Vector Extensions 2 (AVX2) instructions, in addition to AVX (introduced with Sandy Bridge processors), SSE4.2 (introduced with Nehalem processors), and earlier generations of SSE.

AVX2 supports floating point fused multiply-add, integer vector instructions extended to 256-bit, and vectorization gather support, among other features. Your application may be able to take advantage of AVX2, however, not all applications can make effective use of this set of instructions. We recommend that you use the latest Intel compiler, by using the command module load comp-intel and experiment with the following sets of compiler options before making production runs on the Broadwell nodes:

- -O2 -xCORE-AVX2
- -O3 -xCORE-AVX2

These compiler options generate an executable that is optimized for running on Broadwell and Haswell, but cannot run on any of the older processor types. If you prefer to generate a single executable that will run on many Intel processor types (Sandy Bridge, Ivy Bridge, Haswell, Broadwell, Skylake, and Cascade Lake), try using one of the following sets of compiler options:

- -O2 (or -O3)
- -O2 (or -O3) -axCORE-AVX512,CORE-AVX2 -xAVX

The use of -axCORE-AVX512,CORE-AVX2 allows the compiler to generate multiple code paths with suitable optimization to be determined at run time.

If your application does not show performance improvements with -xCORE-AVX2 or -axCORE-AVX512,CORE-AVX2 -xAVX (as compared with just -O2 or -O3) when running on Broadwell nodes, it is more advantageous to use the executable built with just -o2 or -o3 for production runs.

TIP: You can add the compiler options -ip or -ipo to instruct the compiler to look for ways to better optimize and/or vectorize your code. Also, to generate a report on how well your code is vectorized, add the compiler flag -vec-report2.

Notes:

- If you have an MPI code, we strongly recommend that you load the mpi-hpe/mpt module, which always points to the NAS recommended MPT version.
- Ensure your jobs run correctly on Broadwell nodes before you start production work.

Running PBS Jobs on Broadwell Nodes

To request Broadwell nodes, use model=bro_ele in your PBS script.

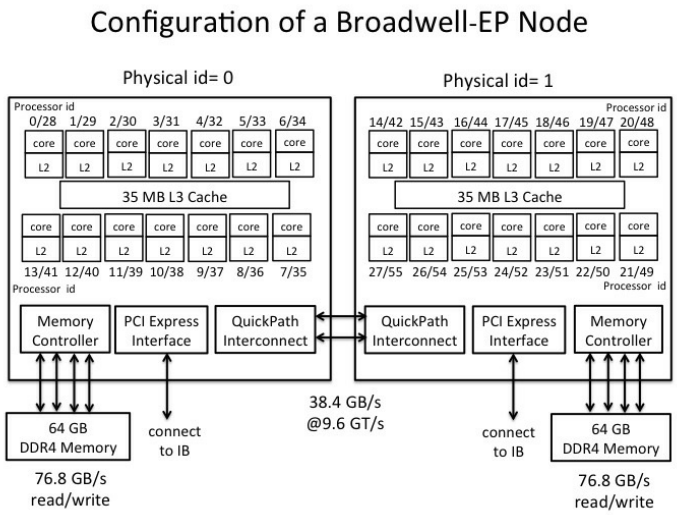
```
#PBS -l select=10:ncpus=28:mpiprocs=28:model=bro_ele
#PBS -l site=static_broadwell
```

Sample PBS Script For Broadwell Nodes

```
#PBS -l select=10:ncpus=28:mpiprocs=28:model=bro_ele
#PBS -l walltime=8:00:00
#PBS -q normal
module load comp-intel/2020.4.304 mpi-hpe/mpt
cd $PBS_O_WORKDIR
mpiexec -np 280 ./a.out
```

For more information about Broadwell nodes, see:

- Electra Configuration Details
- Broadwell Processors



Microarchitecture

The Intel Broadwell processor incorporated into the Electra cluster is the 14-core E5-2680v4 model with a clock speed of 2.4 GHz. As the 14 nanometer (nm) die shrink of the Haswell microarchitecture, the Broadwell processor uses less power and is more efficient than the Haswell processor.

Instruction Sets

Like the Haswell processor, Broadwell supports single instruction, multiple data (SIMD) instruction sets, including several generations of Streaming SIMD Extensions (SSE, SSE2, SSE3, Supplemental SSE3, and SSE4), Advanced Encryption Standard (AES), and Advanced Vector Extensions (AVX and AVX2). Broadwell also supports some new instruction sets, including the Multi-Precision Add-Carry Instruction Extensions (ADX) for arbitrary-precision integer operations.

Main Features of AVX2

Features include:

- Floating-point fused multiply-add (FMA) support, which can double the number of peak floating-point operations compared with those run without FMA. With 256-bit floating-point vector registers and two floating-point functional units, each capable of FMA, a Broadwell core can deliver 16 floating-point operations.
- Integer-vector instructions support has been extended from 128-bit to 256-bit capability.
- Gather vectorization support, enhanced to enable vector elements to be loaded from non-contiguous memory locations.

Compiler Support for AVX2

AVX2 is formally supported by Intel compilers starting with version 12.1. To take advantage of AVX2, use the following compiler and options:

Compiler

We recommend using the latest Intel compiler `comp-intel`.

Options

Use either `-xCORE-AVX2` or `-axCORE-AVX2`.

Both the `-xCORE-AVX2` and `-axCORE-AVX2` options turn on `-fma`, which computes a x b + c in one rounding. This provides higher accuracy than `-no-fma`, which computes a x b + c in two steps (first a x b, then + c) and two roundings. Turn off FMA if you want your results to match legacy ones.

Note: An application that is compiled with AVX2 instructions can run only on Haswell and Broadwell or newer Intel nodes.

For Fortran codes, consider using the option `-align array32byte` to align your vector arrays on 32-byte boundaries for best performance and reliability. This option is available in Intel compiler version 13 and later.

AVX2 is also supported in the GNU Compiler Collection (GCC) starting with version 4.7. Use compiler options such as `-march=core-avx-2` or `-mavx2`, `-mfma` if you want to use AVX2.

Hyperthreading

Hyperthreading is turned ON.

Turbo Boost

Turbo Boost is turned ON.

Memory Subsystems

The memory hierarchy of Broadwell is as follows:

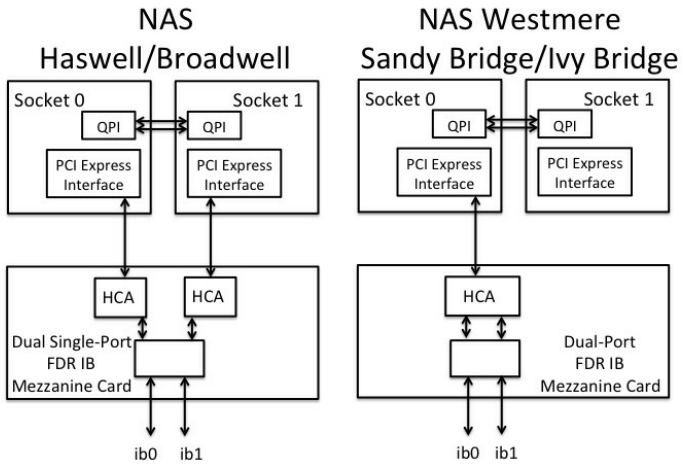
- L1 instruction cache: 32 KB, private to each core
- L1 data cache: 32 KB, private to each core
- L2 cache: 256 KB, private to each core
- L3 cache: 35 MB, shared by 14 cores in each socket
- Memory: 64 GB per socket, total of 128 GB per node

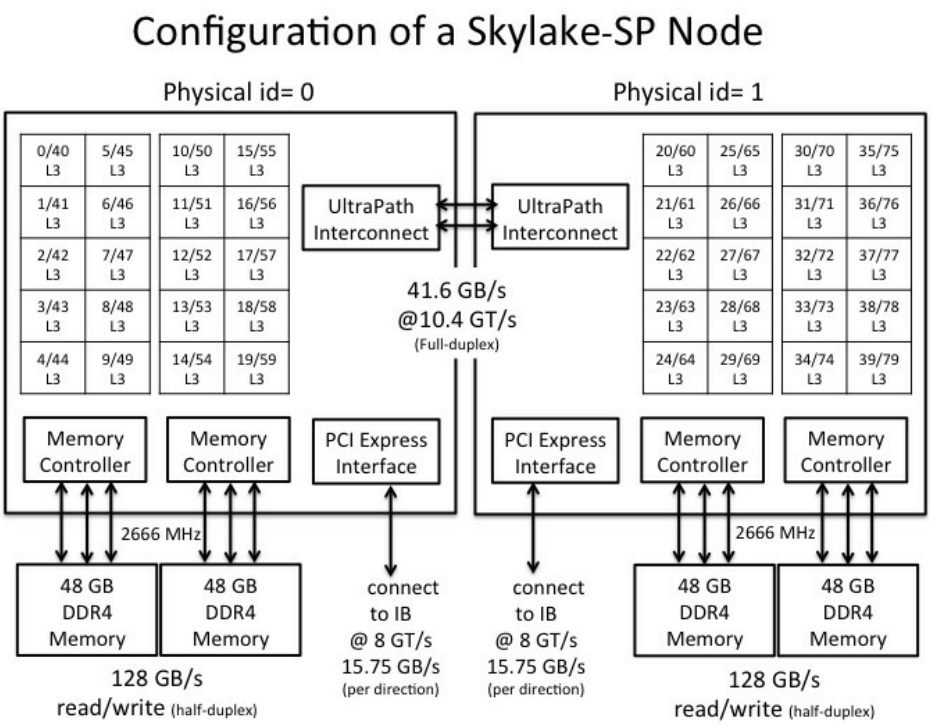
The Broadwell nodes are equipped with 2,400 MHz DDR4 memory to provide higher memory bandwidth. There are four memory channels per socket. Each channel can be connected with a maximum of two memory DIMMs. Of the eight memory DIMM slots for each socket, four are populated with 16-GB error correcting code (ECC)-registered DDR4 memory, for a total of 64 GB per socket. With two sockets in a node, the total memory per node is 128 GB.

Connecting the two sockets are two Intel QPI links running at a speed of 9.6 gigatransfers per second (GT/s). Each link contains separate lanes for the two directions. The total bandwidth (2 links x 2 directions) is 38.4 GB/sec.

Network Subsystem

The network subsystem of the Broadwell nodes is the same as that of the Haswell nodes, as shown in the following diagram. Each Broadwell node is equipped with two PCI Express (PCIe) interfaces (one from each socket). One PCIe interface is connected to the ib0 InfiniBand (IB) fabric via a single-port, four-lane, Fourteen Data Rate (4X FDR) host channel adapter (HCA), in a dual single-port FDR IB mezzanine card. The other PCIe interface is connected to the ib1 fabric via another single-port 4x FDR HCA in the same mezzanine card.





The Intel Skylake processor incorporated into the Electra cluster is the 20-core Xeon Gold 6148 model with a base clock speed of 2.4 GHz. Skylake is a microarchitecture redesign using the same 14-nanometer (nm) manufacturing process technology as Broadwell, with multiple new features and enhancements in the on-chip and intersocket interconnects, memory, cache, and CPU.

Each Electra Skylake node makes use of the 100 Gbits/s four-lane Enhanced Data Rate (4X EDR) technology to connect to the rest of the Electra InfiniBand network.

Each small square in the diagram above 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.

On-chip and Inter-socket Interconnects

In previous generations of Intel Xeon processors (such as Sandy Bridge, Ivy Bridge, Haswell, and Broadwell), the cores, L3 cache, memory controller, I/O controller and intersocket interconnect ports are connected with a on-chip ring architecture. This has the drawback of increased latency and bandwidth constraints when the number of cores per socket increases. To mitigate this issue, Skylake utilizes a mesh architecture that encompasses an array of vertical and horizontal paths, allowing communication from one core to another through a shortest path. Furthermore, each core and L3 slice has a combined Caching and Home Agent (CHA) that handles address mapping and information routing, and provides scalability of resources across the mesh.

The two sockets are connected with two Intel Ultra Path Interconnect (UPI) links, which deliver increased bandwidth and performance over the Intel Quick Path Interconnect (QPI) that was used in previous generations of Intel Xeon processors. 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 (2 links x 2 directions) is 41.6 gigabytes per second (GB/s).

Note: With full-duplex communication between two components, both ends can transmit and receive information between each other simultaneously. With half-duplex communication, the transmission and reception must happen alternatively.

Memory Subsystem

There are two sub-NUMA clusters in each socket, creating two localization domains. Each domain contains one memory controller and ten L3 cache slices. Processes observe lower L3 and memory latency if they access data mapped to the L3 or memory in the same domain where the processes are running, compared with outside of the domain.

There are three memory channels per sub-NUMA cluster. Each channel can be connected with up to two memory DIMMs. For the Electra Skylake configuration, there is one 16-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 activation or deactivation of the sub-NUMA domains is controlled in the Basic Input/Output System (BIOS) firmware. Currently, they are deactivated for the Electra Skylake nodes.

The speed of each memory channel is 2,666 MHz. One 8-byte read or write can take place per cycle per channel. With a total of 6 memory channels, the total half-duplex memory bandwidth is approximately 128 GB/s per socket.

Cache Hierarchy

The cache hierarchy of Skylake is as follows:

- L1 instruction cache: 32 KB, private to each core; 64 B/line; 8-way
- L1 data cache: 32 KB, private to each core; 64 B/line; 8-way; fastest latency: 4 cycles
- L2 cache: 1 MB, private to each core; ; 64 B/line; 16-way; fastest latency: 12 cycles
- L3 cache: shared non-inclusive 1.375 MB/core; total of 27.5 MB, shared by 20 cores in each socket; fully associative; fastest latency: 44 cycles

In Broadwell, the L2 cache is 256 KB per core and the L3 cache is a shared inclusive cache with 2.5 MB per core. In Skylake, the cache hierarchy has

changed to provide a larger L2 cache of 1 MB per core and a smaller shared non-inclusive 1.375 MB L3 cache per core.

Note: An inclusive L3 cache guarantees that every block that exists in the L2 cache also exists in the L3 cache. A non-inclusive L3 cache does not guarantee this.

A larger L2 cache increases the hit rate into the L2 cache, resulting in lower effective memory latency and lower demand on the mesh interconnect and L3 cache.

If the processor has a miss on all the levels of the cache, it fetches the line from memory and puts it directly into the L2 cache of the requesting core, rather than putting a copy into both the L2 and L3 caches, as is done on Broadwell. When the cache line is evicted from the L2 cache, it is placed into L3 if it is expected to be reused.

Due to the non-inclusive nature of the L3 cache, the absence of a cache line in L3 does not indicate that the line is absent in private caches of any of the cores. Therefore, a snoop filter is used to keep track of the location of cache lines in the L1 or L2 caches of cores when a cache line is not allocated in L3. On the previous-generation processors, the shared L3 itself takes care of this task.

Processor

Like the Broadwell processors, Skylake supports single instruction, multiple data (SIMD) instruction sets, including several generations of Streaming SIMD Extensions (SSE, SSE2, SSE3, Supplemental SSE3, SSE4.1 and SSE4.2), and Advanced Vector Extensions (AVX and AVX2). In addition, it also includes MPX (Memory Protection Extensions), Intel SGX (Software Guard Extensions) and Advanced Vector Extensions 512 (AVX-512).

AVX-512 was originally introduced with the Intel Xeon Phi processors. Of the multiple AVX-512 instruction groups, Skylake comes with the AVX512F, AVX512CD, AVX512BW, AVX512DQ groups and a new AVX512VL feature.

With AVX-512, programs can pack eight double precision or 16 single precision floating-point numbers, or eight 64-bit integers, or 16 32-bit integers within the 512-bit vectors. With 512-bit floating-point vector registers and two floating-point functional units, each capable of Fused Multiply-Add (FMA), a Skylake core can deliver 32 floating-point operations per cycle—double the number of operations of a Haswell/Broadwell core, or quadruple that of a Sandy Bridge/Ivy Bridge core can deliver.

Unlike SSE and AVX, which cannot be mixed without performance penalties, the mixing of AVX and Intel AVX-512 instructions is supported without penalty. AVX registers YMM0-YMM15 map into the Intel AVX-512 registers ZMM0-ZMM15, very much like SSE registers map into AVX registers. Therefore, in processors with Intel AVX-512 support, AVX and AVX2 instructions operate on the lower 128 or 256 bits of the first 16 ZMM registers.

Intel AVX-512 optimizations are included in Intel compiler version 16.0 and later versions. For Skylake-specific optimizations, use `-xCORE-AVX512`. With Intel compiler 2018 versions, using the optimization flag `-qopt-zmm-usage=high -xCORE-AVX512` may benefit floating-point heavy applications running on Skylake.

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

Hyperthreading

Hyperthreading is turned ON.

Turbo Boost

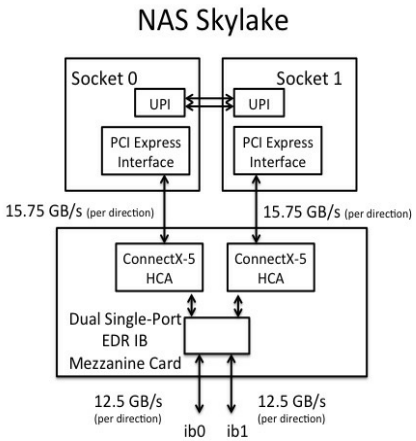
Turbo Boost is turned ON. Maximum Turbo Frequency is 3.70 GHz.

Inter-Node Network

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

One enhancement in the NAS Skylake inter-node network is that it makes use of the 100 Gbits/s Enhanced Data Rate technology instead of the 56 Gbits/s Fourteen Data Rate technology used for the NAS Broadwell nodes. As shown below, one PCIe is connected to the ib0 InfiniBand fabric via a single-port, four-lane, Enhanced Data Rate (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.

Note: For nodes labeled with r[x]i[x]n[9-17,27-35], socket 0 is connected to ib0 and socket 1 is connected to ib1. For nodes labeled with r[x]i[x]n[0-8,18-26], socket 0 is connected to ib1 and socket 1 is connected to ib0.



There are 2,196 Skylake nodes in the Electra cluster. They are partitioned as follows:

- Eight E-cells (288 nodes per E-cell)
- Two compute racks and one cooling rack per E-cell (144 nodes per rack)
- Four individual rack units (IRUs) per compute rack (36 nodes per IRU)
- Nine compute blade (ICE XA SAXON blade) per IRU (4 nodes per blade)

The connection of the compute nodes relies on the use of four premium IB EDR switch blades per IRU. Each switch blade has two 36-port ConnectX-5 ASICs with 9 ports from each ASIC connecting to compute nodes and 18 ports from each ASIC for connecting to external targets. The connection of the Skylake nodes forms an 8-dimensional hypercube. The topology of Electra (including both the Skylake and Broadwell nodes) is a 9-dimensional hypercube.

MPT 2.15 and older versions do not support the ConnectX-5 HCAs. To avoid jobs failing when using MPT 2.15 and older versions, the environment variables MPI_IB_XRC and MPI_XPMEM_ENABLED have been disabled for jobs running on Skylake. If your MPI applications perform significant MPI collective operations and rely on having these two variables enabled to get good performance, use MPT 2.16 or the forthcoming newer versions.

To make sure you use the NAS recommended MPT version, simply load the module `mpi-hpe/mpt`.

References

- Intel Xeon Processor Scalable Family Technical Overview
- Intel 64 and IA-32 Architectures Optimization Reference Manual
- HPE SGI 8600 QuickSpecs