AMD Genoa and Intel Sapphire Rapids review

Martin Cuma, Scientific Consultant, CHPC

Abstract

We compare the performance of fourth generation AMD EPYC CPUs (code named Genoa), and Intel fourth generation Xeon Scalable processor (code named Sapphire Rapids). In particular, a 2x32 Intel 6430, which is a mid-range offering, and three AMD CPUs, 2x32 9334, 1x64 9554 and 1x96 9654, which are on the medium-higher end. We also compare their performance to older AMD and Intel CPUs where possible. Both AMD and Intel new CPUs provide significant, 20-50%, speedup over the previous generation per node. Comparing the AMD CPUs to the Intel's, the Intel CPU is about 20% slower per node than the AMD counterparts. The 96 core 1P AMD node is a better alternative than a 64 core node, as long as the price of the 96 core node is not 20% or more higher.

Introduction

In this article we look at the performance of fourth generation AMD EPYC CPUs (code named Genoa), and at Intel fourth generation Xeon Scalable processor (code named Sapphire Rapids). The AMD CPUs were released at the end of 2022, and the Intel CPUs in early 2023.

Both the AMD and Intel CPUs are built from multi-CPU chiplets, that are connected with high speed bus in the full CPU package. This approach marks a departure from previous Intel monolithic CPU die approach. AMD has used the chiplets since the first EPYC generation in 2017.

Both AMD and Intel release a wide range of CPU models (SKUs) with different clock speed, core counts and other features, which results in a varying performance of these CPU models. We therefore focus on mid range CPU models, which are of most interest for HPC applications with regards to the price/performance ratio. There appears to be limited number of reviews of these CPU models, most reviews published online focus on the highest performing, and most expensive, CPU models. This article thus should shed some light on the performance comparisons of the CPU models that HPC centers are likely to buy.

For a good overview of the CPU architecture and models, I suggest to read https://hothardware.com/reviews/amd-genoa-data-center-cpu-launch for the AMD and https://www.servethehome.com/4th-gen-intel-xeon-scalable-sapphire-rapids-leaps-forward/ for Intel.

A server vendor gave us an access to one Intel and three AMD CPU models in their lab. The new and previous generation CPUs we look at are summarized in Table 1. Note that the 9554 and 9654 are dual socket models, but, in our tests we had one socket disabled so they effectively used a single CPU socket. Single socket (1P) AMD CPU models are more cost effective and we have been mostly buying those in the previous AMD CPU generations. Table 1 lists the price of the P (single socket) models. The dual socket models are slightly more expensive. Other than having disabled the second CPU channel, the P CPUs are identical to the non-P, which makes the performance we present below for the single dual socket CPU valid for the P model. We also include previous generation AMD and Intel CPU models that we have been purchasing for comparison.

Notice that the Intel CPU is lower price range than the AMDs, at $4.2k/node as compared to AMD's $6k-$10.5k/node, but on the same level of the previous generation Ice Lake 6330 CPU that we have been buying.

CPU	Core count	Base (boost) Frequency	Base TDP	List price (AMD P version)
AMD 9334	2x32	2.7 (3.9) GHz	210 W * 2	$2,990 * 2
AMD 9554	64	3.1 (3.75) GHz	360 W	$7,104
AMD 9654	96	2.4 (3.7) GHz	360 W	$10,625
Intel 6430	2x32	2.1 (3.4) GHz	270 W * 2	$2128 * 2
AMD 7713P	64	2.0 (3.675) GHz	225 W	$5010
Intel 6330	2x28	2.0 (3.1) GHz	205 W * 2	$1894 * 2

Table 1. New and previous generation CPUs used in this review

Benchmarks

There is not a large amount of published benchmarks of these new CPUs, apart from the two documents listed above, the high end models of both AMD and Intel are benchmarked at https://www.phoronix.com/review/amd-epyc-avx512 and https://www.phoronix.com/review/intel-sapphirerapids-avx512. These reviews don't provide sufficient information on the mid range, more cost effective, models, which we focus on in this report.

The test machines have been running Red Hat Enterprise Linux (RHEL) 8.6, CHPC machines have been running Rocky Linux 8.5. Rocky Linux is a clone of RHEL, therefore there is only a minor difference in the OS version. On both sites, we have been using the Spack package manager to build the applications that we have benchmarked. On the new test machines, we have used Spack architecture *linux-rhel8-icelake*, and on the AMD we used *linux-rhel8-x86_64_v4*, since Spack is not completely aware of the new CPUs. On CHPC machines, we used *linux-rhel8-icelake* for the Intel and *linux-rhel8-zen3* for the AMD CPUs. Using Spack may not allow for the best performance, but, it mimics the way how we build applications on our systems and as such gives appropriate comparisons to performance on existing CHPC machines and expectations for the new CPUs performance on our systems.

High Performance Linpack (HPL)

HPL solves a dense system of linear equations in double precision and is a base for the TOP500 list of the fastest computers in the world. It gives a good estimate of the raw double precision CPU performance. It uses the Basic Linear Algebra Subprograms (BLAS), which we have provided either with Intel's Math Kernel Library (MKL), or with OpenBLAS. Note that the MKL uses a CPU check at runtime, which decreases its performance at the AMD CPUs, we have evaluated this and see about 20% performance reduction as compared to when this CPU check is disabled on the AMD 9334 CPU. OpenBLAS performance is about 5% worse than MKL on the same CPU, which is why we in the Table 2 below report only MKL results.

HPL was built with the OS supplied gcc/8.5.0 ,since it relies on the BLAS implementation for performance, as

```
spack install hpl@2.3%gcc@8.5.0 arch=linux-rhel8-icelake
spack install hpl@2.3%gcc@8.5.0 arch=linux-rhel8-x86_64_v4
```

	Intel Saph. Rap.	AMD Genoa	AMD Genoa	AMD Genoa	Intel Ice Lake	AMD Milan
Model	6430	9334	9554	9654	6330	7713P
Clock speed	2.10	2.70	3.10	2.40	2.00	2.00
Core count	2x32	64	64	96	2x28	64
HPL score	**3.436**	**3.527**	**3.319**	**3.811**	**2.346**	**1.749**
Theoretical	4.301	2.765	3.174	3.686	3.584	2.048
% Theoretical	79.89%	127.57%	104.56%	103.38%	65.46%	85.40%

Table 2. HPL results

The most impressive is the raw performance increase from the previous to current generation of both AMD and Intel CPUs, almost 90% for the AMD and 50% for the Intel. I suppose the extra power draw of these CPUs, 60% for the AMD and 30% for the Intel, has to go somewhere. The 96 core AMD predictably gives the best performance, but the 2x32 core Intel CPU is not too far behind the 2x32 AMD counterpart. We don't have the node pricing at this point, and know that the single socket nodes tend to be cheaper, but looking at the pure CPU list pricing the dual socket CPUs seem to be more cost effective, since the 2xAMD 9334 list price is ~$6000 while the single 9354P is ~$7000, and the dual socket 2x32 core setup gives better performance than the 1x64 core setup.

With respect to the theoretical performance, the Intel AVX-512 CPUs tend to have 2 Fused Multiply-add (FMA) vector units of 512 bits, which puts the theoretical double precision FLOPS to 32 per cycle. AMD has two AVX-2 256 bit units, which can do one coupled AVX-512 FMA instruction, which puts the FLOPS per cycle to 16. Now, multiplying that by the frequency and core count, assuming the base frequency, we arrive to the theoretical peak double precision throughput shown in the Theoretical row of Table 2, and the percentage of the theoretical peak we achieve with the HPL. From this can be observed several trends, that correlate with known information. First, the previous Intel chips were known to lower the clock speed, when using AVX-512. This is reflected in the 65% of theoretical peak HPL score for the Ice Lake CPU. It has also been discussed, that the new Sapphire Rapids CPU is doing much better – that is, the clock speed is not lowered that drastically, or not at all, which can be seen in achieving 80% of the theoretical peak, as compared to 65% on the Ice Lake. The new AMD CPUs all achieve above 100% theoretical peak. There could be several explanations for that. There are additional two 256 bit vector add units in the Zen4 in addition to the two 256 bit FMA units, which may in certain situations improve the throughput (providing 24 FLOPS/cycle). Also, it looks like the AMD CPUs are capable of spending more time in the turbo mode, that is, running at the higher than base frequency. This may be the reason for the 2x32 Zen4 achieving 128% of theoretical peak, as compared to the 1x64 or 1x96 ~105% theoretical peak.

Looking at the effect of various BLAS implementations, focusing on the 2x32 AMD 9334, we see about 5% lower performance with OpenBLAS, as compared to unrestricted MKL, and, 20% performance improvement, when disabling MKL's CPU ID check on the AMD CPU, which is achieved with the known LD_PRELOAD workaround.

In sum, both the new AMD and Intel CPUs have achieved an impressive 50-90% increased floating point throughput as compared to the previous generations, and the Intel CPU is almost keeping pace with the AMDs. This is a different situation from the previous CPU generation, where the Intel was significantly better, thanks to 2x the theoretical vectorized floating point throughput.

High Performance Computing Challenge (HPCC) benchmark

HPCC benchmark is a synthetic benchmark suite geared at assessing HPC performance from different angles. It consists of seven main benchmarks, that stress various computer subsystems, such as raw performance, memory access and communication. For detailed description of the benchmark see http://icl.cs.utk.edu/hpcc/. We use version 1.5.0. On the test machines, we installed it with Spack with

```
spack install hpcc target=icelake
```

which resulted in using OpenMPI and OpenBLAS, as that's what Spack uses by default. On the CHPC machines, we also use OpenMPI, but we default to MKL for the BLAS. Thus parts of the benchmark that use BLAS (HPL, DGEMM) may be better if we used MKL on the new AMD CPUs, but the important trends are preserved.

Year	2023	2023	2023	2023	2021	2021
CPU generation	Saph R.	Genoa 2x32	Genoa 96	Genoa 64	Milan 64	Ice Lake
Model	6430	9334	9654	9554	7713P	6330
Core count	2x32	2x32	96	64	64	2x28
Frequency_GHz	2.1	2.7	2.4	3.1	2.0	2.0
HPL_Tflops	3.23	3.11	**3.26**	2.94	1.38	1.87
SingleDGEMM_ Gflops	***71.14***	*59.87*	*50.66*	*58.47*	31.83	60.29
PTRANS_GBs	18.20	**29.85**	29.13	24.62	15.34	22.04
MPIRandomAccess_GUPs	0.337	0.445	**0.579**	0.472	0.475	0.309
SingleRandomAccess_GUPs	*0.107*	*0.148*	*0.144*	***0.151***	0.100	0.049
SingleSTREAM_ Triad	*13.28*	***43.20***	*33.88*	*41.08*	24.67	14.28
MPIFFT_Gflops	57.73	**89.58**	68.11	65.44	22.69	24.01

Table 3. HPCC results that scale or remain the same with increased number of CPU cores used, shown for the full node (using all CPU cores). In **bold** the highest value, in *italic* values that don't change much with increased CPU core count (that is, single CPU/serial performance).

The HPL result is about 20% less, than with the specific HPL runs above, that's likely due to the parameters of the run (put into the HPL.dat file). Nevertheless, the trends of impressive gains with the new CPUs are preserved. The *SingleDGEMM* result is in Intel's favor, highlighting the two 512 bit FMA units per Intel CPU core. The rest of the benchmarks are more memory bound and favor the AMD CPUs. Based on https://www.nextplatform.com/2022/11/10/amd-genoa-epyc-server-cpus-take-the-heavyweight-title/ and https://infohub.delltechnologies.com/p/memory-bandwidth-for-next-gen-poweredge-servers-significantly-improved-with-sapphire-rapids-architecture, the AMD has all core memory bandwidth up to ~350 GB/s, while Intel has ~250 GB/s using DDR5-4800 memory, thus the AMD has an advantage. The 6430 Intel CPU also only supports up to DDR5-4400, not 4800, which will make the Intel peak even lower.

NAS Parallel Benchmarks

NAS Parallel Benchmarks (NPB) are a set of programs derived from computational fluid dynamics (CFD) applications. Some basic information about the benchmarks is here: https://en.wikipedia.org/wiki/NAS_Parallel_Benchmarks. Each of these benchmarks can be run with different problem sizes. Class A is a small problem, Class B is medium size, Class C is a large problem, and Class D is a very large problem (needing about 12 GB of RAM). There are also even larger classes E and F. We have ran Classes A-D and present results for Class C.

The build was done with Spack and the OpenMP variant as `spack install npb%gcc@11.2.0 implementation=openmp`, and run with OpenMP threads pinned to each CPU core, `OMP_PROC_BIND=spread, OMP_PLACES=cores`, which was determined to provide the best performance, giving up to 30% performance boost, although in one benchmark, the IS, no thread pinning was advantageous by 5-10%. For consistency, we show the pinned IS result, though.

In Figures 1 a) and b) we look at the single core performance of the NPB. The first five columns of each results are Intel CPUs, the remainder are AMD CPUs of the last 5-7 years. The results below Ice Lake and Milan have been obtained with previous builds of NPB,that were not necessarily using the same compiler options like the newer CPU results built with Spack, so, the comparison is not exact, but it gives some idea how the single core performance has evolved over the last years.

There are two main takeaways from the single core NPB results. The first is considerable advantage of the AMD CPUs over the Intel, and especially of the new Genoa/Zen4 generation. The other is not a big, if any, improvement of the Sapphire Rapids core over the previous Ice Lake. The reason for this is hard to speculate, perhaps the code does not vectorize well, so the Sapphire Rapids better clock speed at AVX-512 does not transfer here. Or the *icelake* optimizations are not beneficial. It will be useful to re-visit this when Spack and the compilers support the Sapphire Rapids microarchitecture.

Figures 2 a), b) show the whole node NPB performance, that is engaging all the CPU cores via OpenMP threads. Similarly to the single CPU results, the new AMD CPUs are again performing much better than Intel, though the Sapphire Rapids node does provide a significant boost as compared to Cascade Lake in most of the benchmarks. Likely due to better memory bandwidth and more CPU cores. The MG and IS, and to certain degree also FT and LU benchmarks are memory bound, which benefits the dual socket AMD node. Parallel scaling noticeably drops for the single socket AMD CPUs in these benchmarks. For the MG and IS, the best performance, though only ~5% better than when using all the CPU cores, is achieved with only ½ of the cores being used. This demonstrates that for memory intensive workloads the dual socket system is advantageous. For the remainder of the benchmarks, the single socket AMD variants keep up with the dual socket alternative.

In Figure 3 a) we show the parallel scaling of the memory bound MG benchmark on the select latest AMD and Intel CPUs. The Intel CPU seems to scale better to higher core counts, though it still performs significantly less that the equivalent 2x32 AMD alternative. The single socket AMD scales poorly after 16 CPU cores.

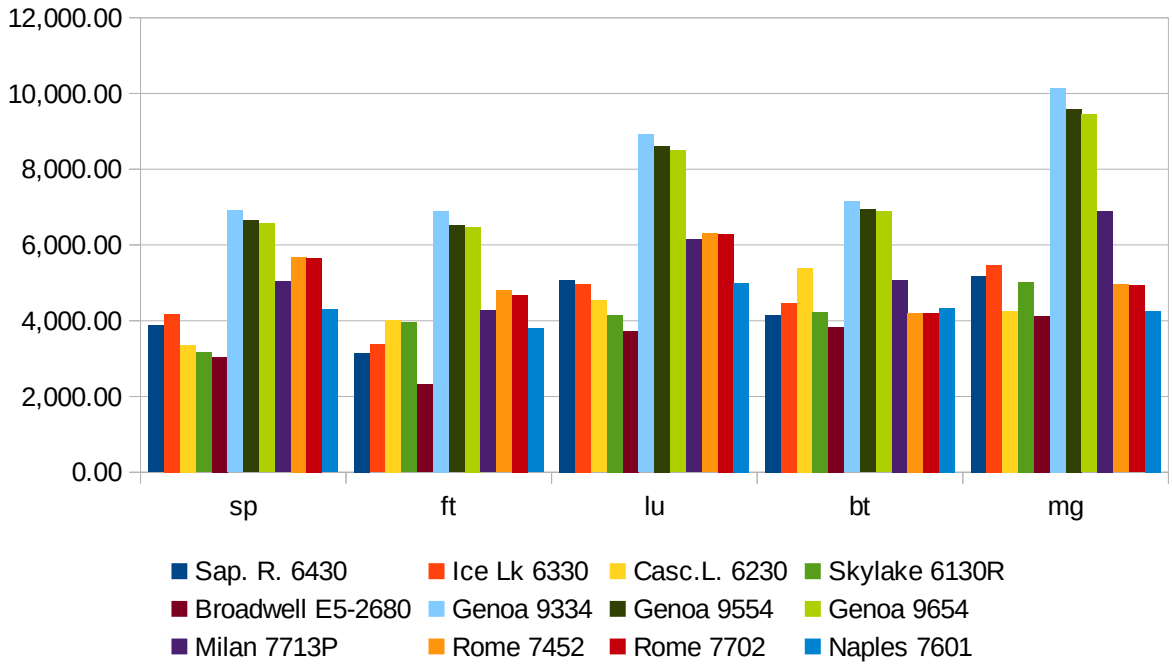

Figure 1 a) Single CPU core results of select NPB benchmarks (in Mop/s), higher is better.

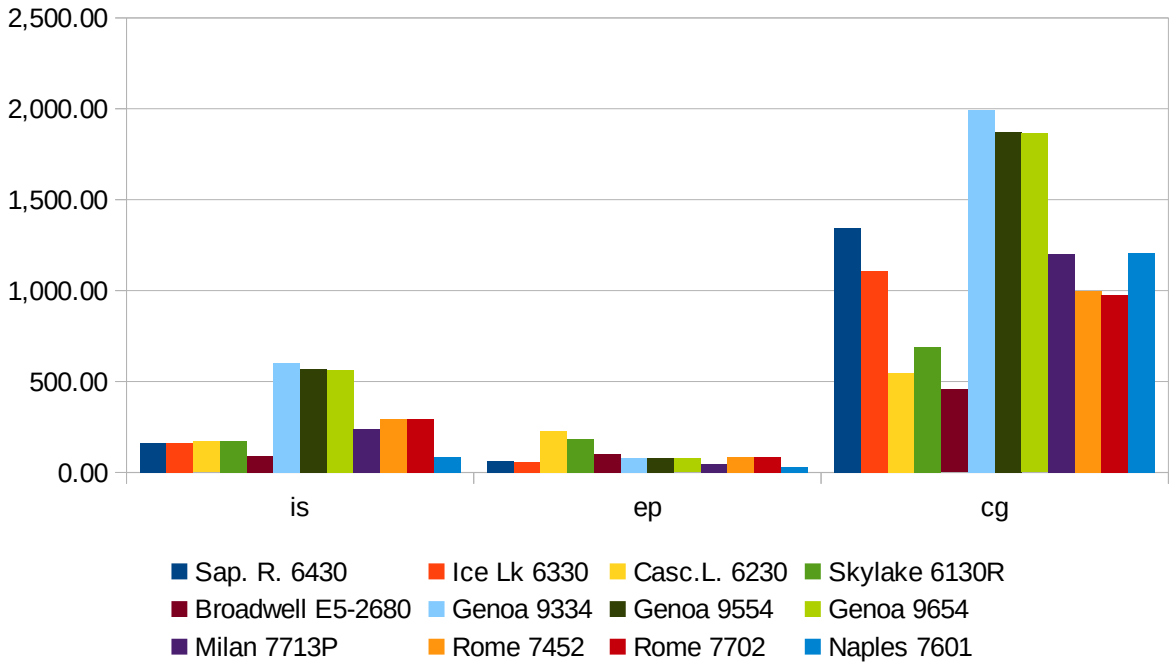

Figure 1 b) Single CPU core results of select NPB benchmarks (in Mop/s), higher is better.

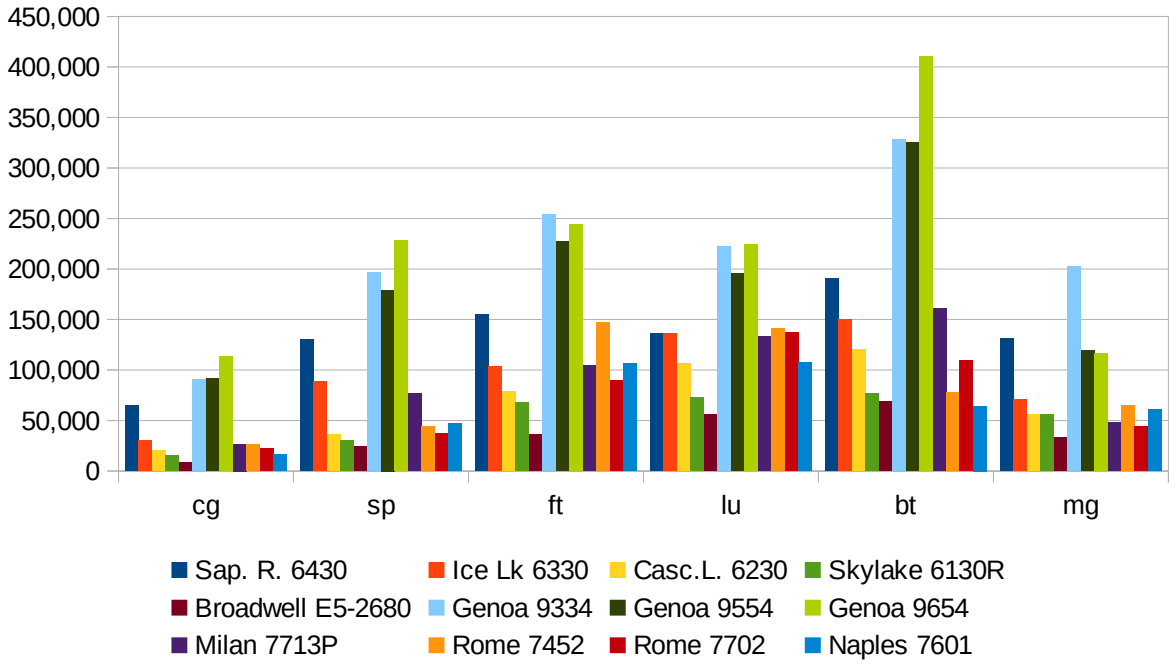

Figure 2 a) Whole node (all CPU cores) results of select NPB benchmarks (in Mop/s), higher is better.

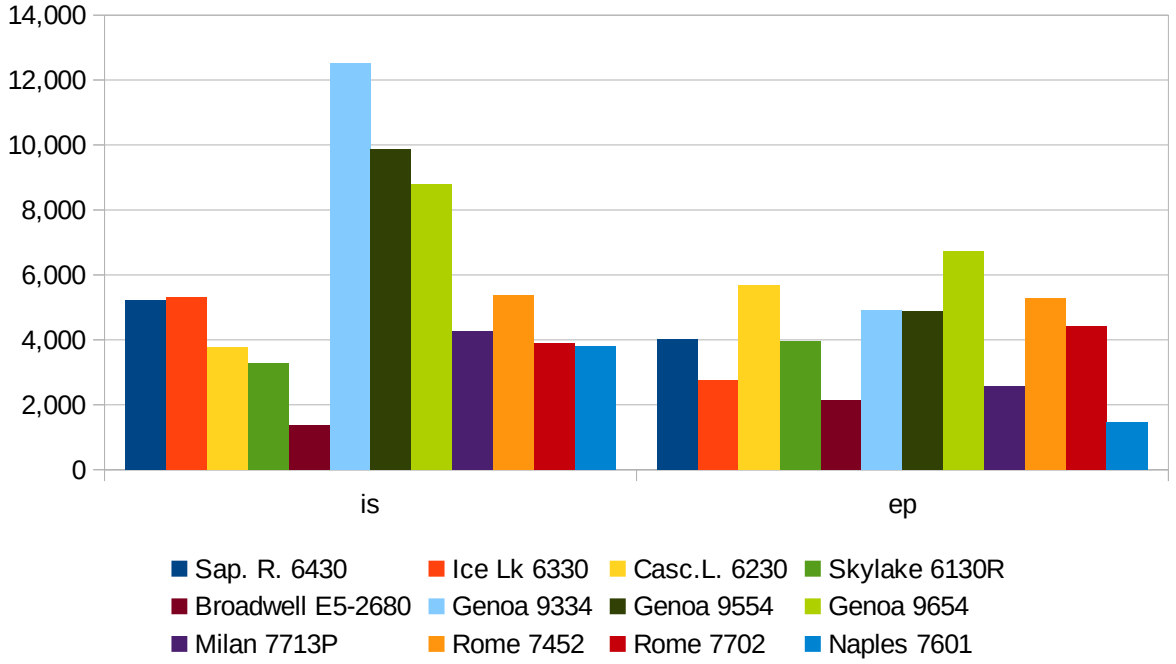

Figure 2 b) Whole node (all CPU cores) results of select NPB benchmarks (in Mop/s), higher is better.

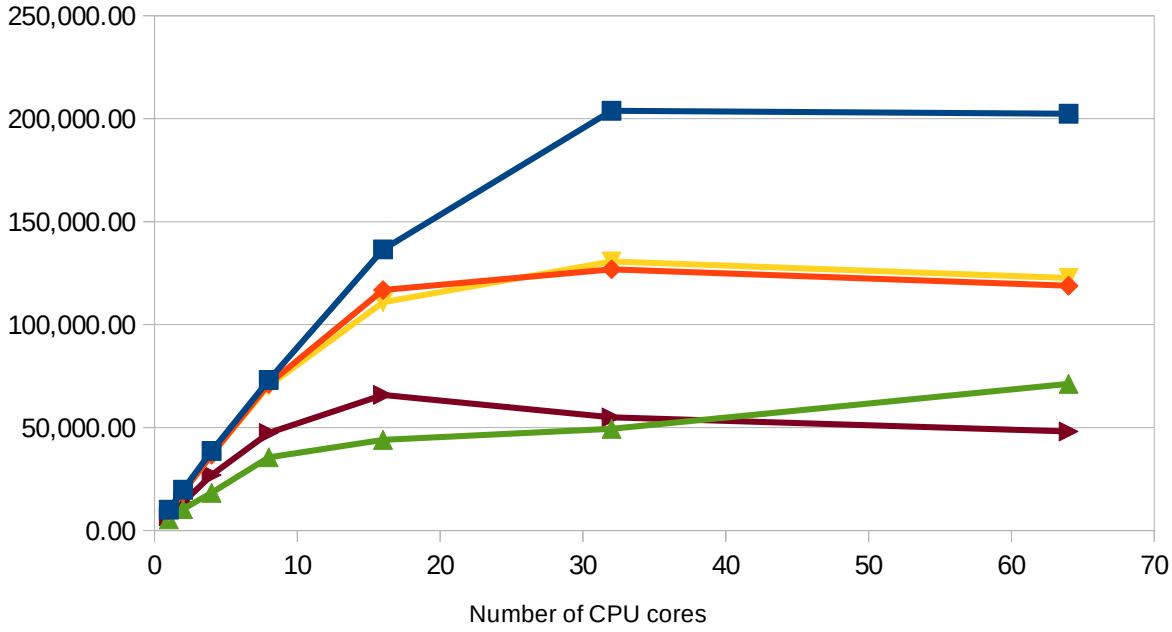

Figure 3 a) Parallel scaling of the NAS MG benchmark (in MOps/s), blue Intel Sapphire Rapids 2x32 cores, red AMD Genoa 9334 2x32 cores, yellow AMD Genoa 9554 1x64 cores, green Intel Ice Lake, magenta AMD Milan

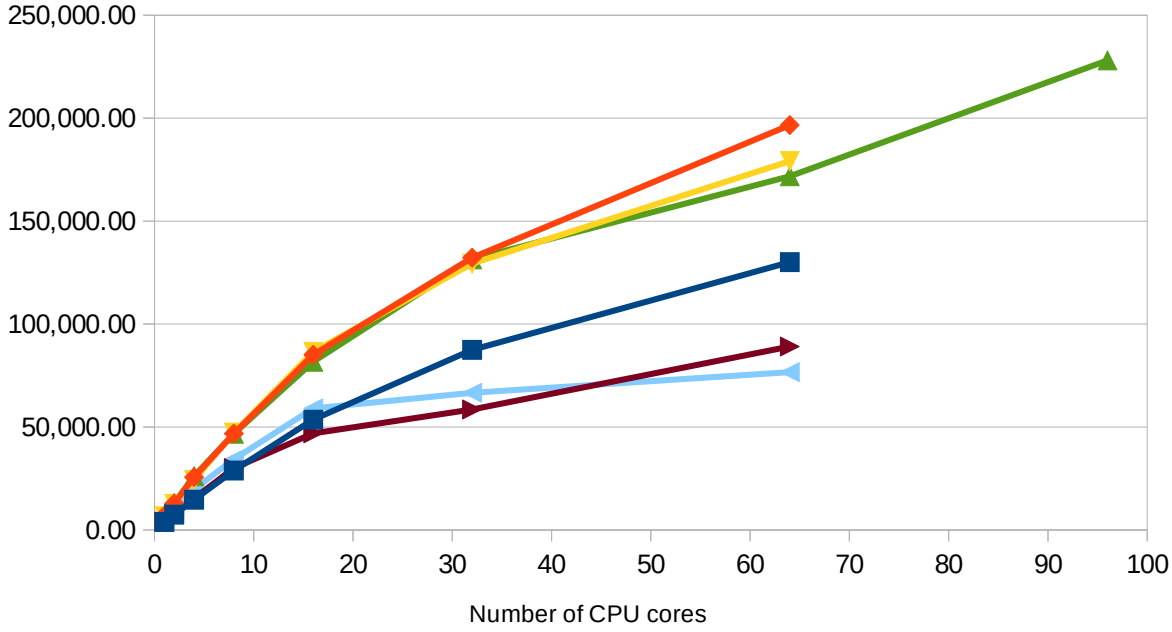

Figure 3 b) Parallel scaling of the NAS SP benchmark (in MOps/s), blue Intel Sapphire Rapids 2x32 cores, red AMD Genoa 9334 2x32 cores, yellow AMD Genoa 9554 1x64 cores, green AMD Genoa 9654 1x96, magenta Intel Ice Lake, light blue AMD Milan

Figure 3 b) demonstrates parallel scaling of a more CPU bound SP benchmark, which scales reasonably to higher CPU core count, and the AMD 9554 one socket CPU is similar, though slightly less at 64 cores than the dual socket 9334 CPU.

LAMMPS

LAMMPS is a popular molecular dynamics simulation program developed at Sandia National Laboratory. It is a good representative for multi-body like simulations, that use internally coded computational kernels, not relying so much on vendor accelerated libraries.

For the latest Intel and AMD CPUs, we have built LAMMPS with Spack, optimized for the linux-rhel8-icelake or linux-rhel8-x86_64_v4 architecture, with

```
spack install lammps%gcc@11.2.0 +voronoi+class2+extra-fix+reaxff+misc+molecule+manybody+mc+meam+kokkos+colvars+plumed+replica+rigid+kspace+extra-pair
```

For the Intel Ice Lake and AMD Milan, we built and ran at CHPC using the generic `nehalem` Spack target used for production. The MPI used was OpenMPI.

The older results were obtained with the 31Mar17 version using Intel 2019.5 or 2017 compilers, MPI and MKL (using MKL's FFTW wrappers) and with optimization flags `-axCORE-AVX512,CORE-AVX2,AVX,SSE4.2 -O3 -prec-div -fp-model precise`. On the AMD Rome, we used Intel 2019.5 with flags `-march=core-avx2 -ip -prec-div -fp-model precise`. The rest of the flags were taken from the USER-INTEL package makefile. The MPI used was Intel MPI. Thus we again are not doing an exact comparison with these older CPUs, but, rather include them as a reference for how the newer CPUs built in a fairly canned (and possibly not super optimized) fashion compare to likely best optimized older binaries.

We have run three LAMMPS benchmarks from http://lammps.sandia.gov/bench.html:
LJ = atomic fluid, Lennard-Jones potential with 2.5 sigma cutoff (55 neighbors per atom), NVE integration Chain = bead-spring polymer melt of 100-mer chains, FENE bonds and LJ pairwise interactions with a 2^(1/6) sigma cutoff (5 neighbors per atom), NVE integration EAM = metallic solid, Cu EAM potential with 4.95 Angstrom cutoff (45 neighbors per atom), NVE integration. Each problem was scaled 2x in each dimension resulting in 256,000 atoms and was run for 1,000 time steps.

CPU	Saph. R.	Ice Lk	Genoa	Genoa	Genoa	Milan	G9554/SR	G9554/ G9654	G9334/ G9654
Model	6430	6330	9334	9554	9654	7713P			
Core count	2x32	2x28	2x32	64	96	64			
1	123.59	149.92	96.21	99.56	100.80	130.01	0.81	1.01	0.97
2	63.82	76.90	49.41	51.43	52.17	67.61	0.81	1.01	0.96
4	31.63	39.21	25.18	26.18	26.52	33.92	0.83	1.01	0.96
8	16.11	19.91	13.34	13.44	13.61	17.49	0.83	1.01	0.99
16	8.15	10.12	6.64	6.86	6.97	8.94	0.84	1.02	0.97
32	4.18	5.17	3.48	3.52	3.58	4.71	0.84	1.02	0.99
48					2.43				
64 (56)	2.28	3.24	1.99	1.93	1.87	2.99	**0.85**	0.97	**1.03**
96					1.65			**1.17**	

Table 4. LAMMPS LJ benchmark results in seconds, lower is better.

Table 4 shows runtimes in seconds for the LJ benchmark, the other two exhibit similar trends. The last three columns in the table compare the select new CPUs. From this we can see that the Sapphire Rapids node is about 20% slower than the 1P 64 core AMD node, the 2P and 1P 64 CPU core setup performs roughly the same (= LAMMPS is CPU bound), and that the 96 core AMD node gives about 20% better performance than the 64 core node.

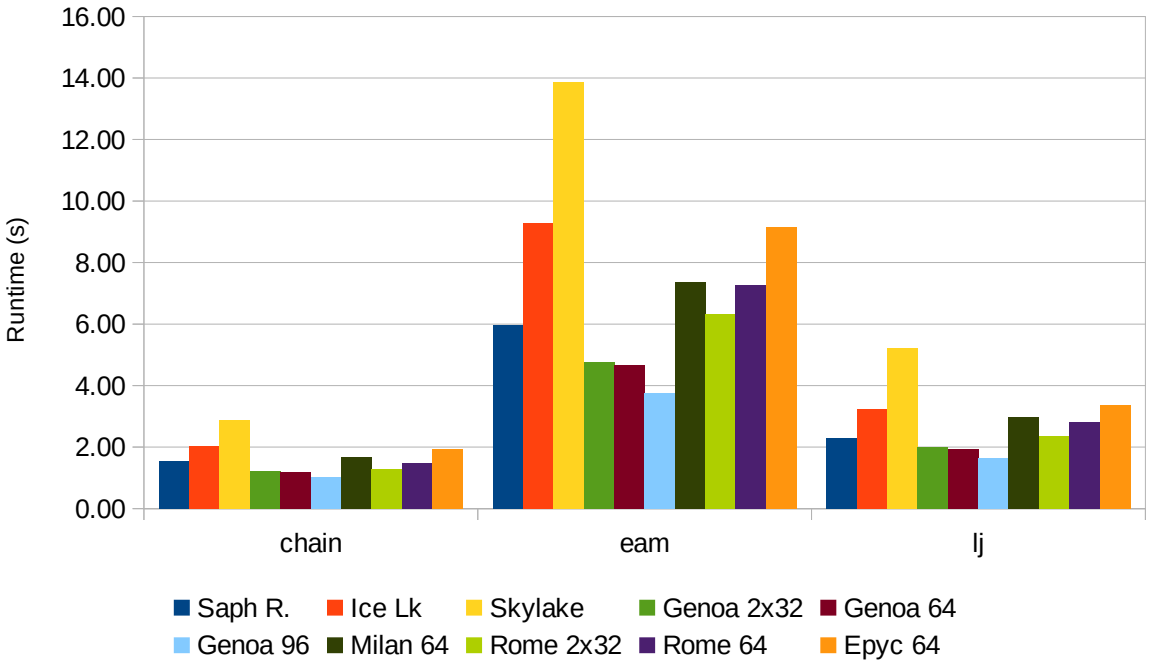

Figure 4. Whole node performance of the three LAMMPS benchmarks, in seconds, lower is better.

In Figure 4 we compare the runtime on a whole node between the select CPUs. Intel has improved markedly from Skylake to Sapphire Rapids, but, so did the AMD Genoa from the Milan and Rome CPUs, for the Ice Lake and Milan, part of it may be use of the Nehalem optimized binary. Unless the price of the 96 core AMD 1P node is more than 20% higher than of the 64 core 1P node, the 96 core node is a better choice for LAMMPS.

GROMACS

GROMACS is another molecular dynamics program, similar to LAMMPS, but, it does its internal assembly optimizations for various CPU microarchitectures, which makes it important to build it for that particular microarchitecture.

For the new CPUs we built GROMACS 2022.3 with Spack as
`spack install gromacs@2022.5%gcc@11.2.0+lapack arch=linux-rhel8-icelake ^intel-oneapi-mkl`
for the Intel CPU, and similarly with the `arch=linux-rhel8-x86_64_v4` for the AMD CPUs. For MPI, OpenMPI was used.

On CHPC machines, we used the same version built for production for the `icelake` and `zen2` targets, for the Intel and AMD CPUs, respectively.

We run two different benchmarks, benchMEM, 82k atoms, protein in membrane surrounded by water, 2 fs time step, and benchRIB, 2M atoms, ribosome in water, 4 fs time step, obtained from the Max Planck Institute for multidisciplinary sciences.

CPU	Saph. R.	Ice Lk	Genoa	Genoa	Genoa	Milan	G9554/ SR	G9554/ G9654	G9334/ G9554
Model	6430	6330	9334	9554	9654	7713P			
Cores	2x32	2x28	2x32	64	96	64			
benchMEM	134.54	112.75	145.96	150.10	164.88	84.01	1.12	1.10	0.97
benchRIB	11.06	7.85	13.14	12.78	15.36	5.91	1.16	1.20	1.03

Table 5. GROMACS whole node performance in ns/day, higher is better

In Table 5 we present the GROMACS performance on the whole node. The last three columns show some ratios between different CPUs. The AMD 1P 64 core node has 12% to 16% better performance than the Intel node. The 96 core 1P AMD node has 10-20% advantage over the 64 core 1P node. And, the 2P vs 1P 64 core node has similar performance. All in all, GROMACS show similar trends to LAMMPS, with the 96 core 1P AMD node being the best choice, as long as the price is not more than 10-20% higher than that of the 64 core 1P node.

NWCHEM

NWCHEM is a quantum chemistry simulation program, which depends heavily on dense linear algebra provided by BLAS and LAPACK, which performance should indicate that of other quantum chemistry simulations like VASP or Gaussian. The advantage of NWCHEM is that it is open source and buildable by Spack, although with some caveats listed below.

On the new CPUs, we have built NWCHEM 6.8.1 with

```
spack install nwchem@6.8.1%gcc@8.5.0~openmp
```

either with OpenBLAS, or with Intel MKL, and with OpenMPI. On CHPC systems, we have used a prior build of NWCHEM 7.0.2 with Intel MPI and MKL. We did not succeed to build 6.8.1 on CHPC, and 7.0.2 on the new test system, due to dependency issues that were difficult to resolve. Therefore the comparison to the older CPUs is not ideal.

We look at the C240 buckyball benchmark, which is fairly widely used and published.

CPU	Saph. R.	Ice Lk	Genoa	Genoa	Genoa	Milan
Model	6430	6330	9334	9554	9654	7713P
Cores	2x32	2x28	2x32	64	96	64
C240	1,354.23	1,358.50	971.50	968.30	961.1	1,029.00

Table 6. NWCHEM C240 runtime on the whole node in seconds, lower is better.

Table 6 shows the runtime for the C240 benchmark run on the whole node. There is a moderate improvement from Milan to Genoa, and no improvement from Ice Lake to Sapphire Lake. This may be due to the speed improvements done in the version 7.0.2 as compared to 6.8.2 and warrants revisiting once we get nodes with the new CPUs in the CHPC environment.

There’s also not too much difference between the three AMD CPUs. Furthermore, the runtime is about the same for MKL (shown in the table) and OpenBLAS, and, disabling the CPU ID check in MKL only yields about 1% speed up, as compared to about 20% in the HPL, which suggests that this NWCHEM run is not bound by the BLAS/LAPACK.

Conclusions

Both the Intel Sapphire Rapids and AMD Genoa generation CPUs provide a significant performance boost for most applications, anywhere from 20-50%. The AMD CPUs are more performant than those of Intel, and the 96 core 1P AMD CPU is a better choice, if it’s priced no more than 20% more than the 64 core 1P node for CPU bound applications like molecular dynamics, or dense linear algebra.