HYPERION RESEARCH

HPC User Forum Update

NAVGEM on the Cloud: Computational Evaluation of Commercial Cloud HPC with a Global Atmospheric Model

Bob Sorensen, Melissa Riddle
February 2020

IN THIS UPDATE

The HPC User Forum was established in 1999 to promote the health of the global HPC industry and address issues of common concern to users. In September 2019, the 73rd HPC User Forum took place at Argonne National Laboratory in Illinois. This update summarizes a presentation from that meeting entitled *NAVGEM on the Cloud: Computational Evaluation of Commercial Cloud HPC with a Global Atmospheric Model*, given by Daniel Arevalo from Devine Consulting.

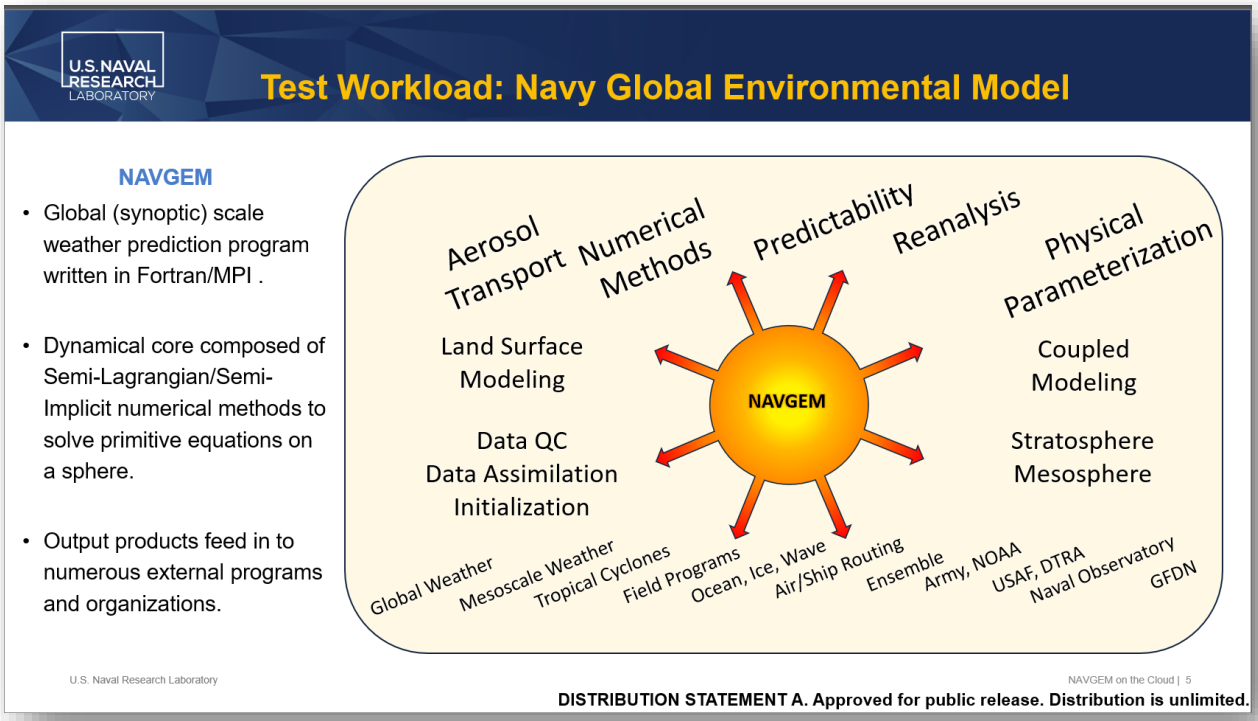
Source: DeVine and Hyperion Research, 2020

COMPUTATIONAL EVALUATION OF COMMERCIAL CLOUD HPC WITH A GLOBAL ATOMSPHERIC MODEL, DANIEL AREVALO FROM DEVINE CONSULTING

Daniel Arevalo from DeVine Consulting presented an evaluation of their work with the U.S. Naval Research Laboratory on the use of HPC and cloud capabilities for atmospheric modelling and forecasting. The Navy DoD Supercomputing Resource Center (Navy DSRC), which is the Navy's arm of the DoD HPC Modernization Program, is located at the Stennis Space Center in Mississippi. It supports various Defense computational areas and has a total computing capability of ~10 petaflops, consisting of two HPE SGI 8600 systems and two Cray XC40s.

Following a mandate issued by the U.S. Congress, the Navy introduced a "Cloud First" policy stipulating that for any upgrade or change the Navy would look toward cloud capability first. Although many of their public facing websites and back-end admin are now using cloud technology, HPC is still done in-house.

FIGURE 1

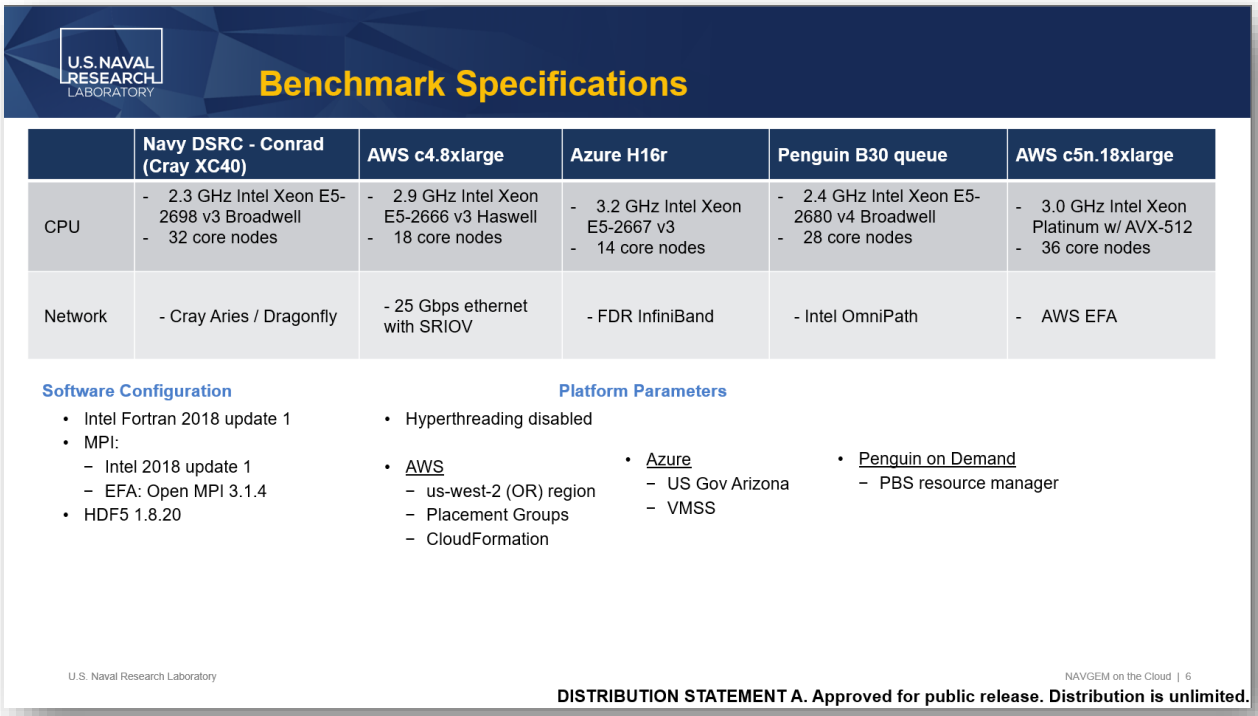


Source: DeVine and Hyperion Research, 2020

The Navy's workhorse, the Navy Global Environmental Model (NAVGEM), is the keystone of the Navy's current global atmospheric weather prediction capabilities. It was developed at Naval Research Laboratory (NRL) in Monterey, California, and is used operationally by the fleet numerical and meteorological command. The model is run every six hours, and the Navy runs a 180-hour forecast in

addition to a twice-daily sixteen-day guidance forecast. The dynamical core of the forecast is a semi-Lagrangian/semi-implicit numerical method solving the primitive equations on a 2D decomposition of the earth. The model is not overly parallel; there is a lot of transposing into and out of Fourier space as well as halo exchanges using certain nonblocking and MRA MPI calls. Researchers there believed that this code would be a good test of the viability of the cloud's capabilities for HPC work.

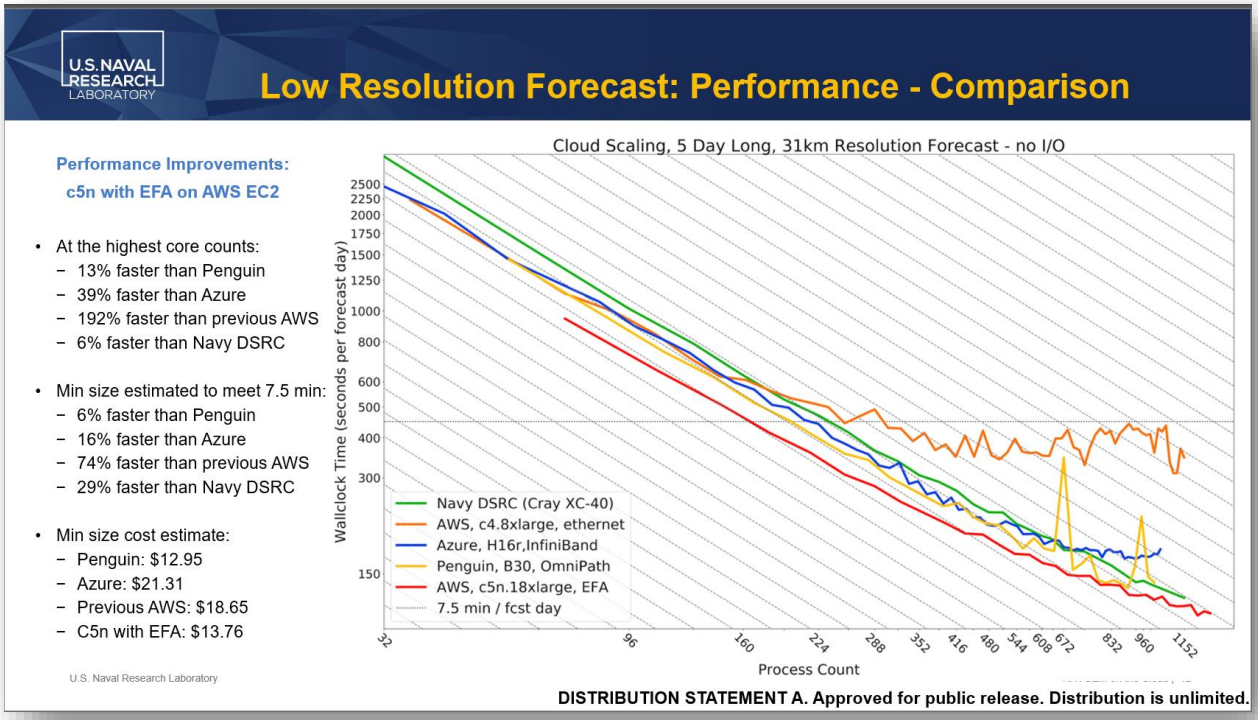
FIGURE 2



Source: DeVine and Hyperion Research, 2020

Arevalo reviewed the different cloud providers they tested, starting with AWS and then Azure. There was a pilot program with HPCMP that also allowed them to test on Penguin on Demand (POD) and, finally, a revisit to AWS. For the sake of consistency, Intel Fortran was used the entire test – mostly Intel MPI with a brief switch to Open MPI for AWS's most updated networking capability. Different Intel CPUs were tested as well as different networks, not just InfiniBand.

FIGURE 3



Source: DeVine and Hyperion Research, 2020

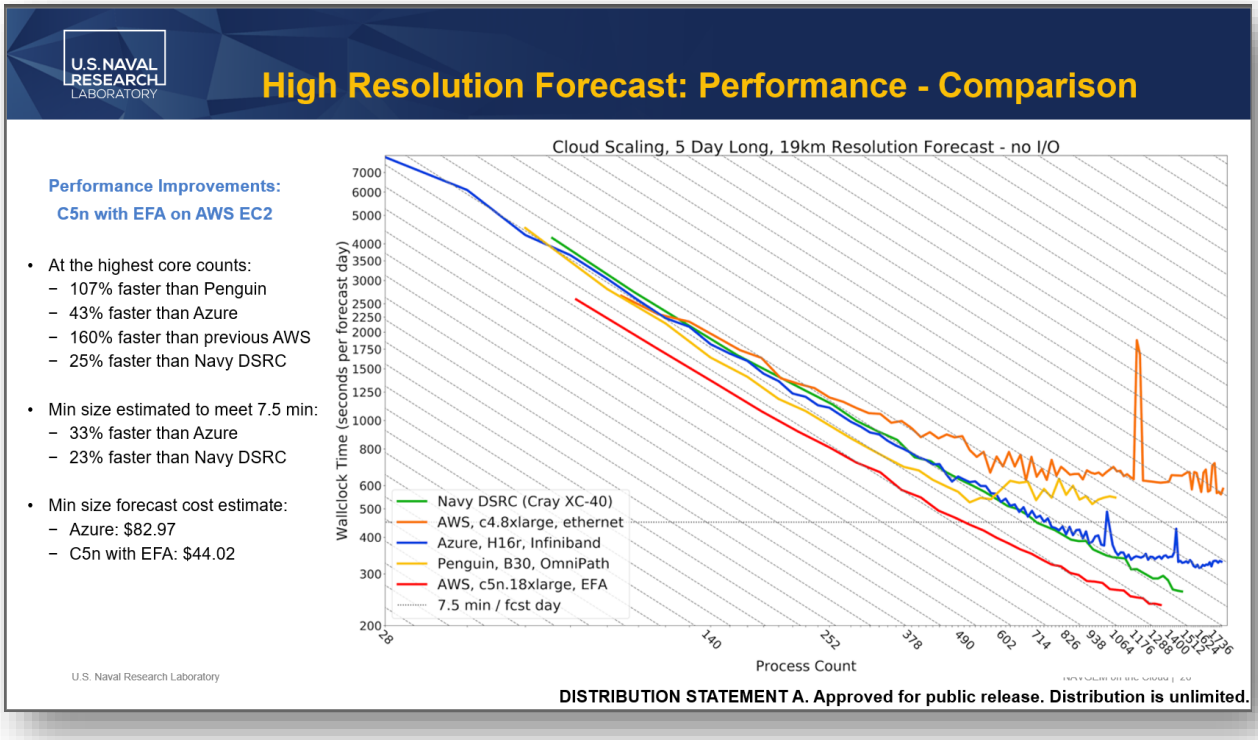
Arevalo next presented test results. In his charts, the forecast day in seconds was represented on the y-axis and the number of cores equal to the MPI process count on the x-axis. As seen in Figure 3, the first line represented the average forecast day performed by the Navy DSRC Cray xC-40. The diagonal lines indicate ideal scaling with the horizontal dashed line representing the Navy's traditional 7.5-minute goal for a forecast day. Arevalo expresses satisfaction with both scalability and surpassing the 7.5-minute goal.

- The next was the first attempt at AWS. At the time, the most updated Amazon compute optimized instance was the c4.8xlarge ethernet. The image offered a 25 Gb/s networking capability for the test scales well for small clusters, up to ~160 cores, at which case the performance flattens out and shows an increase in variability.
- The next example was Azure, with FDR InfiniBand networking. Testers saw an improved performance curve compared to the AWS test. The code scaled well, with variability increasing at higher clusters and the curve flattening out at ~700 cores. Under the advisement of Microsoft, Arevalo and his team underutilized the core count, using 14 of 16 core nodes as a means to address poor initial performance.
- Testers then moved onto Penguin on Demand (POD). The performance was similar or slightly improved from Azure and DSRC until the ~600-700 core count mark, at which time the

variability increased. Arevalo noted that onboarding for POD was much easier than the others, citing a simpler and more user-friendly experience.

- Finally, Arevalo discussed the revisiting of AWS. Amazon had since updated instances, moving from c4 to c5n, which has updated networking capability, along with Elastic Fabric Adapter (EFA), which is their implantation of LibFabrics open source fabric software. There was a notable improvement compared with the last performance from AWS. Testers saw smooth scaling up to ~450 cores and the variability increases beyond that in a stair-step function. They were impressed by the performance increase.

FIGURE 4



Source: DeVine and Hyperion Research, 2020

Arevalo then moved on to discussing their higher-resolution tests. By increasing the resolution, they hoped to push the performance curves "to the right." By increasing the resolution, the compute to communication ratio increases. Again, starting with Navy DSRC, testers noted good scaling, and Arevalo explained that to stay below the 7.5-minute forecast day goal they had to push the core count.

- The first attempt on AWS is not able to meet the 7.5-minute forecast day.
- Azure produced similar performance as before compared to DSRC. Variability was introduced at larger cluster sizes, but it still pushed that performance to the right albeit not until ~500 cores.

- The updated AWS with EFA curve had smooth scaling and the same stair step function variability much later at ~1000 cores.

Arevalo noted a caveat that it is testing only the compute capability of the cloud. They disabled I/O for many of their tests or subtracted it from the performance runtime because of the problems they were having. Because a weather forecast without I/O is not useful, their next steps include incorporating that. The dashed plot lines indicate compute-only time and the solid lines are with I/O included. They are looking to expand the testing of the NAVGEM ensemble, as this is just the core of a larger program.

For more information or to view this and other presentations given at HPC User Forums dating back to 2008, visit www.hpcuserforum.com.

About Hyperion Research, LLC

Hyperion Research provides data driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
